

Article

Infrared/Visible Light Fire Image Fusion Method Based on Generative Adversarial Network of Wavelet-Guided Pooling Vision Transformer

Haicheng Wei ^{1,*} , Xinping Fu ², Zhuokang Wang ² and Jing Zhao ^{3,*}

¹ School of Medical Technology, North Minzu University, Yinchuan 750021, China

² School of Electrical and Information Engineering, North Minzu University, Yinchuan 750021, China; 20227417@stu.nmu.edu.cn (X.F.); 20227386@stu.nmu.edu.cn (Z.W.)

³ School of Information Engineering, Ningxia University, Yinchuan 750021, China

* Correspondence: wei_hc@nmu.edu.cn (H.W.); zhaojing_nx@nxu.edu.cn (J.Z.)

Abstract: To address issues of detail loss, limited matching datasets, and low fusion accuracy in infrared/visible light fire image fusion, a novel method based on the Generative Adversarial Network of Wavelet-Guided Pooling Vision Transformer (VTW-GAN) is proposed. The algorithm employs a generator and discriminator network architecture, integrating the efficient global representation capability of Transformers with wavelet-guided pooling for extracting finer-grained features and reconstructing higher-quality fusion images. To overcome the shortage of image data, transfer learning is utilized to apply the well-trained model to fire image fusion, thereby improving fusion precision. The experimental results demonstrate that VTW-GAN outperforms the DenseFuse, IFCNN, U2Fusion, SwinFusion, and TGFuse methods in both objective and subjective aspects. Specifically, on the KAIST dataset, the fusion images show significant improvements in Entropy (EN), Mutual Information (MI), and Quality Assessment based on Gradient-based Fusion (Q_{abf}) by 2.78%, 11.89%, and 10.45%, respectively, over the next-best values. On the Corsican Fire dataset, compared to data-limited fusion models, the transfer-learned fusion images enhance the Standard Deviation (SD) and MI by 10.69% and 11.73%, respectively, and compared to other methods, they perform well in Average Gradient (AG), SD, and MI, improving them by 3.43%, 4.84%, and 4.21%, respectively, from the next-best values. Compared with DenseFuse, the operation efficiency is improved by 78.3%. The method achieves favorable subjective image outcomes and is effective for fire-detection applications.

Keywords: fire image fusion; transfer learning; generative adversarial network; vision transformer; wavelet-guided pooling



Citation: Wei, H.; Fu, X.; Wang, Z.; Zhao, J. Infrared/Visible Light Fire Image Fusion Method Based on Generative Adversarial Network of Wavelet-Guided Pooling Vision Transformer. *Forests* **2024**, *15*, 976. <https://doi.org/10.3390/f15060976>

Academic Editor: Viacheslav I. Kharuk

Received: 9 April 2024

Revised: 13 May 2024

Accepted: 31 May 2024

Published: 1 June 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Infrared images contain rich thermal radiation information, which is less affected by complex environments and is widely used in fire detection. However, because infrared images are limited by imaging equipment, the collected images have problems such as low spatial resolution, insufficient texture information, etc., and cannot be popularized in fire recognition [1]. Considering that visible images include rich reflective light details and have high spatial resolution and sufficient detail information, the fusion of visible and infrared images can achieve comprehensive fire images with rich texture details and thermal radiation information and improve the accuracy of fire detection [2].

There are many studies on infrared and visible image fusion in traditional methods. The more common ones are those using Pyramid Transform [3], Wavelet Transform [4], Contourlet Transform [5], etc. To decompose the source image into sub-band images of different scales, image fusion is carried out according to fusion rules, and fusion images are obtained by using the corresponding multi-scale inverse transform [6]. Although these

image fusion methods can complete the image fusion task well, there are some problems, such as noise and artifacts.

With the development of deep learning, image fusion technology based on neural networks has become an important research direction in this field [7]. In terms of Convolutional Neural Network (CNN)-based image fusion methods, reference [8] proposed an encoder- and decoder-based image fusion framework (PG-Fusion), which developed multi-scale gradient residual blocks and a dual-stream pyramid to preserve both visible light details and thermal radiation information. Li et al. [9] constructed a two-stage fusion network through salient object segmentation masks, enhancing fusion effects by identifying and retaining features of salient objects in the images. Reference [10] introduced a robust infrared and visible light image fusion framework using a multi-receptive field attention mechanism and color visual perception, aiming to improve fusion quality and ensure that the fused image aligns better with visual perception in terms of details and colors. Although CNNs possess powerful feature extraction capabilities, their fixed receptive fields limit their ability to provide global information, impacting image fusion results. In the realm of Generative Adversarial Network (GAN)-based image fusion methods, Jin et al. [11] employed a multi-level and multi-classification GAN (MMGAN) to fuse visible light and infrared images, enhancing fusion quality in forest-fire scene imagery. Reference [12] proposed a GAN with instance attention and semantic transition modules (AT-GAN), utilizing a dual discriminator approach with two different attention mechanisms for each image to extract important information, resulting in fused images with richer information. Huang et al. [13] introduced a Multi-Attention GAN (MAGAN) by constructing a multi-attention generator and two multi-attention discriminators for fusing infrared and visible light images. However, GAN-based methods still face several challenges, such as training instability, increased noise in generated images, and the inability to effectively model global information due to using small-size convolutional kernels, leading to the loss of some important scene features. Transformer-based image fusion methods address the lack of long-term dependencies in both CNNs and GANs. In [14], a Transformer fusion network was proposed, which uses the strong feature representation ability of a self-attention mechanism for image fusion. The authors of [15] proposed a multi-scale adaptive Transformer for multimodal medical image fusion, which uses adaptive convolution and an adaptive Transformer for global information extraction but performs poorly in fine granularity, affecting image quality. The authors of [16] proposed TGFuse by combining a Transformer and a GAN. This model uses both channel and space converters to generate the fused image and uses two discriminators to identify the deep information of the fused image and the source image. However, there are still limitations in detail expression. Although the above Transformer-based network has learned the global fusion relationship, compared with the method proposed in this study, it performs poorly in terms of fine granularity, relies too much on the convolution layer when extracting local features, and cannot enhance some meaningful features, affecting image quality.

Based on the mentioned issues, a method for infrared/visible light fire image fusion based on the GAN of Wavelet-Guided Pooling Vision Transformer (VTW-GAN) is proposed. Adversarial learning is conducted through a generator and discriminator, both embedded with Transformer modules. TGFuse is improved by integrating the Transformer with wavelet-guided pooling in the generator. Transfer learning is employed to apply pre-trained models to fire image fusion, thereby enhancing fusion accuracy. The main contributions of this study are as follows:

- (1) The VTW-GAN model uses the GAN as the basic network and embeds the improved Transformer module in the generator, which solves the problem of poor performance in terms of fine granularity of forest-fire fusion images. The improved Transformer module combines the efficient global representation capability of the Transformer with the detail enhancement of wavelet-guided pooling. The fusion model keeps more detailed texture information while learning the global fusion relationship, which lays the foundation for accurate forest-fire detection.

- (2) The VTW-GAN model uses transfer learning to solve the problem of insufficient forest-fire image data. The pre-training model is obtained by training on the KAIST dataset, and the pre-training model is fine-tuned based on the Corsican Fire dataset. In the case of limited data, the performance of the model in forest-fire image fusion is improved.
- (3) This study conducted experiments on the KAIST dataset and Corsican Fire dataset and proved that VTW-GAN has excellent fusion performance in pedestrian street images and forest-fire images, which reflects the better generalization ability of the VTW-GAN model.

2. Materials and Methods

The network structure of the VTW-GAN image fusion method is shown in Figure 1, which is mainly divided into two parts. The first part is the image fusion pre-training part including the generator and discriminators D1 and D2. This part inputs the visible and infrared images of the source domain into the generator to generate the fusion image. Discriminator D1 is used to distinguish the infrared image and fusion image, and discriminator D2 is used to distinguish the visible image and fusion image, providing high-resolution details of infrared thermal information and visible light, respectively. Through the continuous confrontation and iterative update of the generator and discriminator, high-quality fusion images in the source domain are obtained. The second part is the migration learning model part, which fine-tunes the weights and models learned in the first part on the source domain, extracts YUV components from a small amount of fire image data in the target domain, and inputs them into the generator after training to obtain the transfer learning model. Then, it combines the output components I_{FY} , I_{FU} , and I_{FV} and converts them into RGB format to obtain the final fusion result I_F .

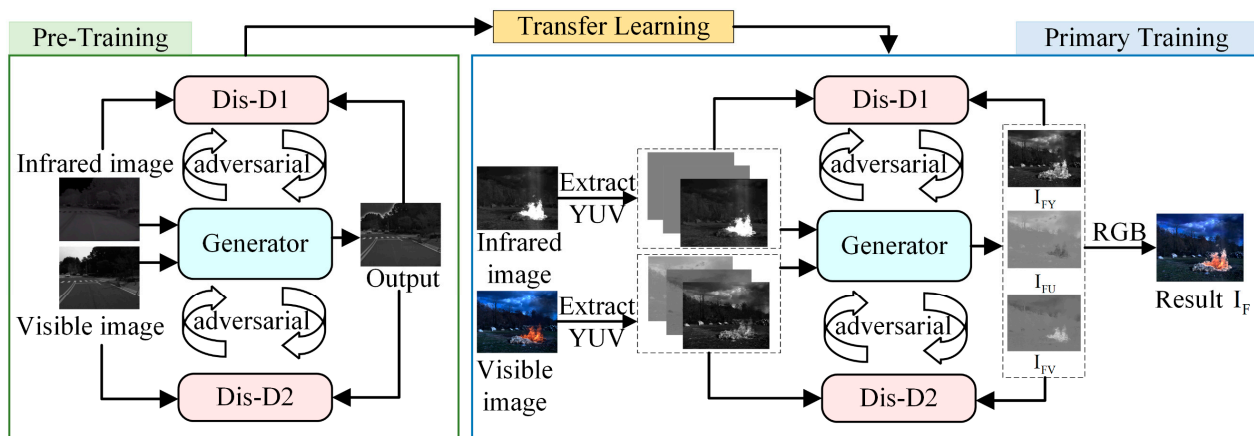


Figure 1. The network structure of VTW-GAN.

2.1. Generator

The network framework of the generator is shown in Figure 2. After merging the infrared and visible images in the channel dimension, the initial feature extraction is carried out through the convolution layer with a convolution kernel of 3×3 . The Res Block layer contains two convolution layers and a residual block. In order to obtain the best model performance with limited computing resources and balance computing resource consumption and feature representation capabilities, three downsampling operators are added to form a four-layer Res Block layer for extracting the deep-layer features of different scales. In order to make up for the lack of global dependency in the CNN fusion method, the Vision Transformer (ViT) Fusion Module is added, and the mixed CNN features are input into the ViT Fusion Module to learn the global fusion relationship. In the ViT Fusion Module, the learned Relationship Map is used to enhance the feature representation. The

fusion features of different scales are sampled to the original image size, and they are superimposed to obtain the single-channel fusion results.

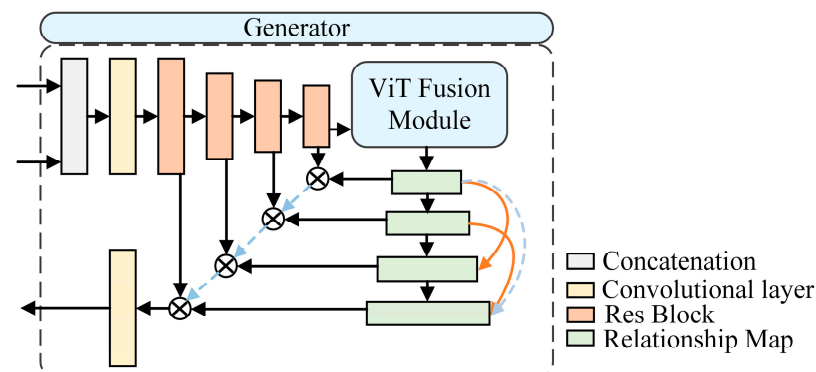


Figure 2. The network framework of the generator.

In order to extract global context information, the ViT Fusion Module in the generator consists of a Channel Transformer and Spatial Transformer, which refines the fusion relationship within the spatial scope and across channels. The network structure is shown in Figure 3. In the Spatial Transformer, the position embedding is canceled, and the image is divided into blocks and stretched into vector groups. After entering the Transformer model to learn the global spatial relationship between image blocks, the vector group is restored to the image, where “ p ” represents the block size, “ w ” and “ h ” represent the number of image blocks in the width and height dimensions, respectively, and “ E ” represents the reduced dimension. The Channel Transformer is similar to the Spatial Transformer. It changes the number of tokens input into the Transformer encoder to the number of channels to learn the information correlation of cross-channel dimensions. The Channel and Spatial Transformer are combined into a ViT Fusion Module in turn to learn the relationship mapping suitable for infrared and visible image fusion.

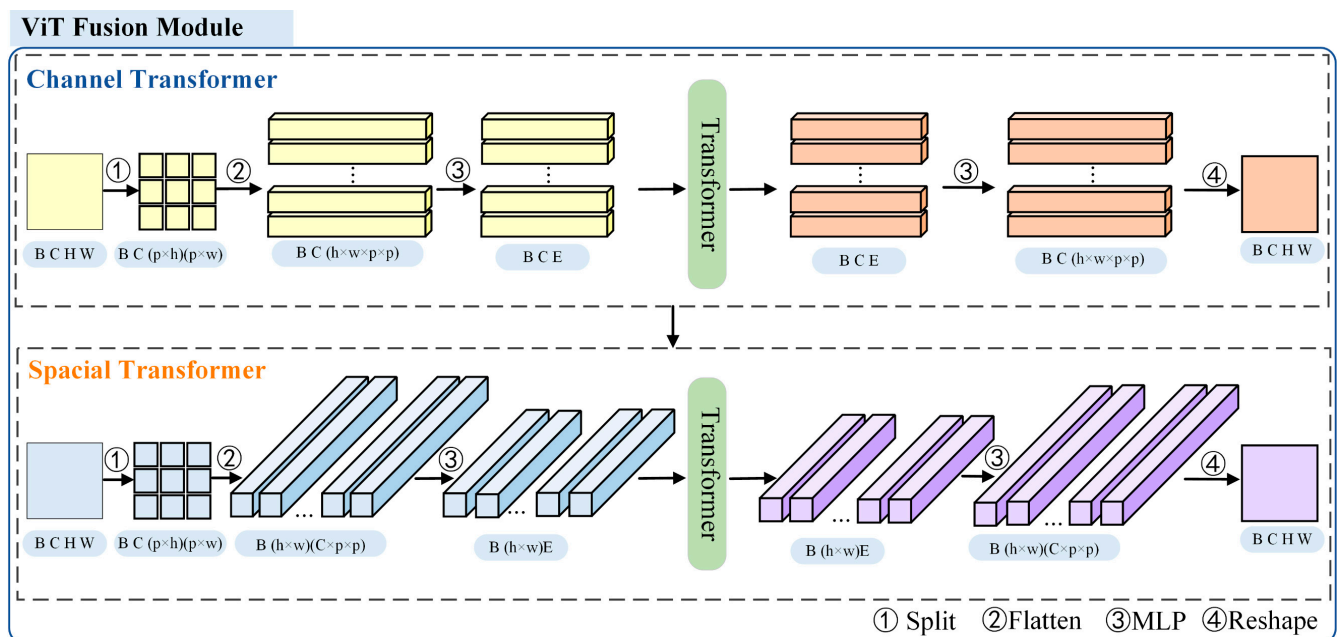


Figure 3. ViT Fusion Module.

However, the ordinary Transformer encoder cannot obtain the details of the fire fusion image. In order to obtain more fine-grained features [17], Haar Wavelet-Guided Pooling

was embedded into the Spatial Transformer encoder to add multi-scale features, and wavelet de-pooling was added after the encoder to accurately reconstruct the image. Haar wavelet boot pooling has four cores:

$$\begin{aligned} LL &= \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} & VD &= \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix} \\ HD &= \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix} & DD &= \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \end{aligned} \quad (1)$$

Compared with ordinary pooling, Haar Wavelet-Guided Pooling decomposes the input into four component channels to extract multi-scale information and texture details. Low-frequency components acquire global basic information, and the high-frequency components VD , HD , and DD extract vertical, horizontal, and diagonal edge details. In order to maintain the continuity and smoothness of the edge of the feature map, the LL is filled with reflection filling, while the high-frequency components are deconvoluted to enhance the clarity and detail of the edge, and then the four components are summed to completely restore the image. The network structure of the improved Transformer encoder is shown in Figure 4. After the global feature F is extracted through the attention mechanism, the global multi-scale features are obtained through the multi-scale kernel LL , VD , HD , and DD of the Haar Wavelet-Guided Pooling layer to capture the vertical, horizontal, and diagonal texture details. The definition is as follows:

$$\{F_{LL}, F_{VD}, F_{HD}, F_{DD}\} = \text{Pooling}(F) \quad (2)$$

Pooling ($\cdot$) refers to the wavelet-guided pooling operation, F_{LL} refers to the acquired global basic information, and F_{VD} , F_{HD} , and F_{DD} refer to the acquired details. The F_{LL} is transferred to the Patch Merging layer to increase the number of output channels, and the detailed features are given in the wavelet solution pool for fire image reconstruction. The convolution block contains multiple convolution layers using the activation function, which is used to obtain local semantic information and map it to the higher-order feature space.

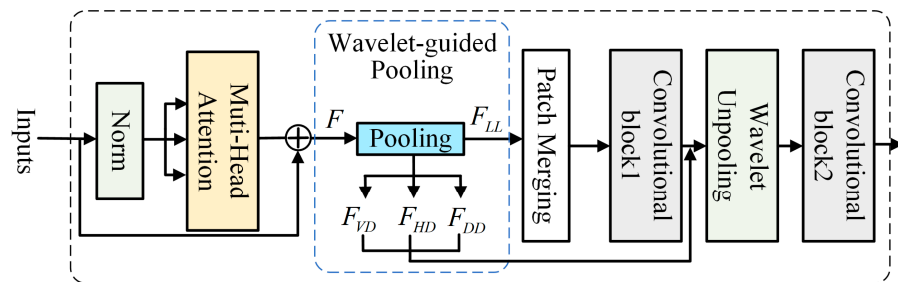


Figure 4. Improved Network Structure of Spatial Transformer Encoder.

2.2. Discriminator

Discriminator D1 and D2 both use the pre-trained VGG-16 network as the network structure. VGG-16 is divided into four layers, and different layers have different feature depths and feature shapes, as shown in Figure 5. Taking discriminator D1 as an example, the infrared image and the fused fire image are input into VGG-16 network for feature extraction, respectively, and the fused image is approximated to the infrared image by calculating L1 loss. TGFuse is referred to to distinguish infrared and visible light features of fire images by using different depth features extracted by VGG-16: the D1 discriminator focuses on retaining more significant information in the fourth-layer features of the network, while the D2 discriminator uses the first-layer features to retain more detailed information. Through the interaction between the discriminator and generator in confrontation training, high-quality fusion images are finally generated.

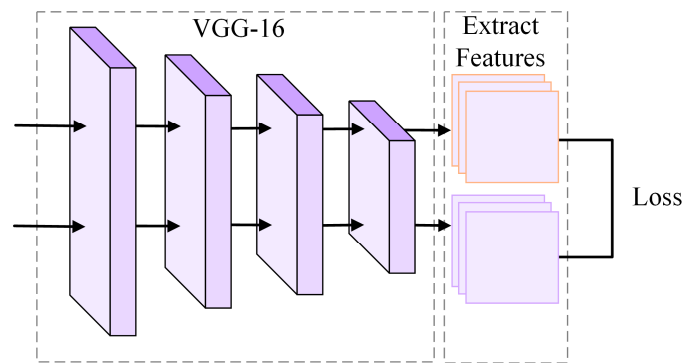


Figure 5. The network framework of discriminators D1 and D2.

2.3. Loss Function

To prevent conflicts between loss functions when multiple loss functions optimize fused images from different perspectives during training, we refer to TGFuse’s introduction of the SSIM [18] loss function to measure the structural similarity between images as the loss of generators. SSIM is defined as follows:

$$SSIM(X, Y) = \frac{(2\mu_X\mu_Y + C_1)(2\sigma_{XY} + C_2)}{(\mu_X^2 + \mu_Y^2 + C_1)(\sigma_X^2 + \sigma_Y^2 + C_2)} \quad (3)$$

where X and Y represent infrared and visible images, respectively, μ and σ represent mean and Standard Deviation, respectively, σ_{XY} represents the covariance between X and Y , C_1 and C_2 represent the stability coefficients, and σ^2 is the variance, which reflects the contrast of the image. The higher the contrast, the more helpful the human visual system is to capture information. The variance is defined as follows:

$$\sigma^2(X) = \frac{\sum_{i=0}^{M-1} \sum_{j=0}^{N-1} [X(i, j) - \mu]^2}{MN} \quad (4)$$

M , N represents the image size in the horizontal and vertical directions, respectively. In order to make the fused image coordinate the consistency between different image blocks, the generator loss function is defined as follows:

$$Patches\_SSIM(I_X, I_Y, I_F|W) = \begin{cases} SSIM(I_X, I_F), & \text{if } \sigma^2(X) > \sigma^2(Y) \\ SSIM(I_Y, I_F), & \text{if } \sigma^2(Y) \geq \sigma^2(X) \end{cases} \quad (5)$$

$$L_{Patches\_SSIM} = 1 - \frac{1}{N} \sum_{W=1}^N Patches\_SSIM(I_X, I_Y, I_F|W)$$

$Patches\_SSIM(I_X, I_Y, I_F|W)$ calculates the structural similarity of image blocks, I_X and I_Y represent infrared images and visible images, respectively, I_F represents fusion images, and W represents the number of image blocks. By calculating the structural similarity between the fused image and the reference image, the fused image is continuously optimized to make it close to the reference image. Discriminators D1 and D2 use feature level loss to enable the fused image to obtain features consistent with infrared and visible images at different levels. The loss function is defined as follows:

$$L_{D_i} = \|\varphi_j(I) - \varphi_j(I_F)\|_1 \quad (6)$$

I represents infrared images or visible images, and φ_j denotes the features extracted from the j th layer of the pre-trained VGG-16 network. For the discriminators D1 and D2, b is 1 and 4, respectively.

2.4. Transfer Learning

In order to achieve the best results, image fusion models often need a large number of infrared and visible images to participate in model training. However, in practical applications, fire images are difficult to obtain. Through migration learning, the information and features in the source domain can be transferred to the fire domain to make up for the lack of data in the fire domain and improve the generalization ability and performance of the model. Therefore, it is necessary to use prior knowledge to build the model and then use migration learning to fine-tune the model and migrate it to the dataset to be fused. After pre-training based on the KAIST dataset [19] to form a basic network and learn the characteristics of the source domain, in order to retain the feature extraction ability learned in the pre-training model, all generators trained in the source domain are copied to the target domain generator, and only the discriminators D1 and D2 of the target domain are trained. By transferring the feature mapping between the street image in the source domain and the corresponding fusion image to the network of the fire image in the target domain, the fusion image is trained on the Corsican Fire dataset [20] and KAIST dataset, and the pre-training model is fine-tuned.

3. Experiment and Discussion

The experimental method is based on the following hardware platform: the CPU is Inter (R) Core (TM) i9-9900k CPU@3.60GHz (Intel, Santa Clara, CA, USA), the graphics card is NVIDIA GeForce RTX 2080Ti GPU (NVIDIA, Santa Clara, CA, USA), and the memory is 32 G. The algorithm model is implemented by building the python 3.8+ Python framework on the Ubuntu 18.04 system.

The dataset for model training is randomly divided into a training set and verification set, with a ratio of 4:1. The training process is divided into two stages: the first stage is the pre-training stage: 40,000 pairs of visible and infrared images are randomly selected on the KAIST dataset as the training set. The training batch is set to 16, the optimizer is Adam, the training phase contains 20 epochs, and the learning rate is set to 0.0001. The second stage is the migration learning stage: 512 images from the KAIST dataset and Corsican Fire dataset are used to carry out migration training for visible and infrared fire images. After data enhancement (rotation, overturning, blurring, and noise increase) is performed on the fire images, 4096 pairs of visible and infrared fire images are obtained. The training batch is set to 16, the optimizer is Adam, the training phase contains 10 epochs, and the learning rate is set to 0.0001.

3.1. Evaluation Metrics

The subjective evaluation of the image fusion results is mainly performed to evaluate the color, lightness, fidelity, and other visual effects of the fused image with the naked eye. The objective evaluation is performed to quantitatively analyze the fused image by calculating the relevant index information of the image, including Entropy (EN), Spatial Frequency (SF), Mutual Information (MI), Average Gradient (AG), Standard Deviation (SD), and Quality Assessment based on Gradient-based Fusion (Q_{abf}). They are defined as follows:

$$EN = - \sum_{i=0}^n X_i \log_2 X_i \quad (7)$$

where n is the gray level and is the proportion of gray value pixel i in the total number of pixels. EN is used to measure the amount of information contained in an image. The more uniform the image histogram, the larger the EN .

$$\begin{aligned}
SF &= \sqrt{RF^2 + CF^2} \\
RF &= \sqrt{\frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N (I(i, j) - I(i, j-1))^2} \\
CF &= \sqrt{\frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N (I(i, j) - I(i-1, j))^2}
\end{aligned} \tag{8}$$

where RF is the row frequency and CF is the column frequency. M, N is the width and height of the image, and I is the pixel value of the image at (i, j) . SF reflects the change rate of the image grayscale. The larger the SF is, the higher the quality of the fused image is.

$$MI(A, B) = H(A) + H(B) - H(A, B) \tag{9}$$

MI is calculated according to the information Entropy $H(A)$ and joint information Entropy $H(A, B)$ of the image. The larger the MI , the more information of the source image is retained in the fused image.

$$AG = \frac{1}{(M-1)(N-1)} \sum_i^{M-1} \sum_j^{N-1} \sqrt{\frac{(I(i+1, j) - I(i, j))^2 + (I(i, j+1) - I(i, j))^2}{2}} \tag{10}$$

AG is used to measure the clarity of the fused image. The larger the AG , the higher the image definition, and the better the fusion quality.

$$SD = \sqrt{\frac{1}{MN} \sum_i^M \sum_j^N (F(i, j) - \mu)^2} \tag{11}$$

where F is the pixel value of the image at position (i, j) , which is the mean value. The larger the SD , the more information the image carries.

$$Q_{abf} = \frac{\sum_{i=0}^{M-1} \sum_{j=0}^{N-1} (Q_{(i,j)}^{af} W_{(i,j)}^a + Q_{(i,j)}^{bf} W_{(i,j)}^b)}{\sum_{i=0}^{M-1} \sum_{j=0}^{N-1} (W_{(i,j)}^a + W_{(i,j)}^b)} \tag{12}$$

where $Q_{(i,j)}^{af}$ and $Q_{(i,j)}^{bf}$ represent the edge information retention value from the source image a and image b to the fusion image f , respectively, and $W_{(i,j)}^a$ and $W_{(i,j)}^b$ represent the edge intensity function. Q_{abf} considers the gradient consistency between the fused image and the reference image. The higher the value, the better the quality of the fused image.

3.2. Comparison and Analysis of Experimental Results on the KAIST Dataset

Based on the fusion pre-training model, the fusion image is tested on the KAIST dataset and evaluated by subjective and objective evaluation indicators. Figure 6 shows the subjective comparison results of the three groups of images. From Figure 6a, it can be seen that DenseFuse, IFCNN, and U2Fusion show less clear sky edges and less obvious infrared information of houses than VTW-GAN. The SwinFusion fusion image has the lowest contrast, and the visible light details of light and shadow on the road are not clear enough. The clarity of TGFuse and VTW-GAN has been improved, but because of the introduction of the network structure of wavelet-guided pooling, VTW-GAN will not lose infrared thermal information while retaining the details of houses and roads in visible light. In Figure 6b, the difference between the bright and dark parts in the fusion image of DenseFuse, IFCNN, and SwinFusion is not obvious enough, and the contrast of the car is not reflected in the fusion image, resulting in the loss of local information. It can be seen from Figure 6c that although the “person” in the dark part of the fusion result of DenseFuse and SwinFusion is more prominent, the visible light texture of the figure’s clothing and

bicycle is seriously lost, and the high contrast in SwinFusion causes the image contour to be too sharp, making the fusion result unnatural. Compared with IFCNN and U2Fusion, TGFuse has a more prominent brightness difference between the eaves and walls in the fusion results, but the infrared characteristics of the characters are fuzzy, such as that the white infrared information of the head is not reflected. In contrast, the fused image of VTW-GAN contains more natural details, showing houses, roads, lights, and cars more clearly, retaining visible texture details while reducing the loss of local infrared information.

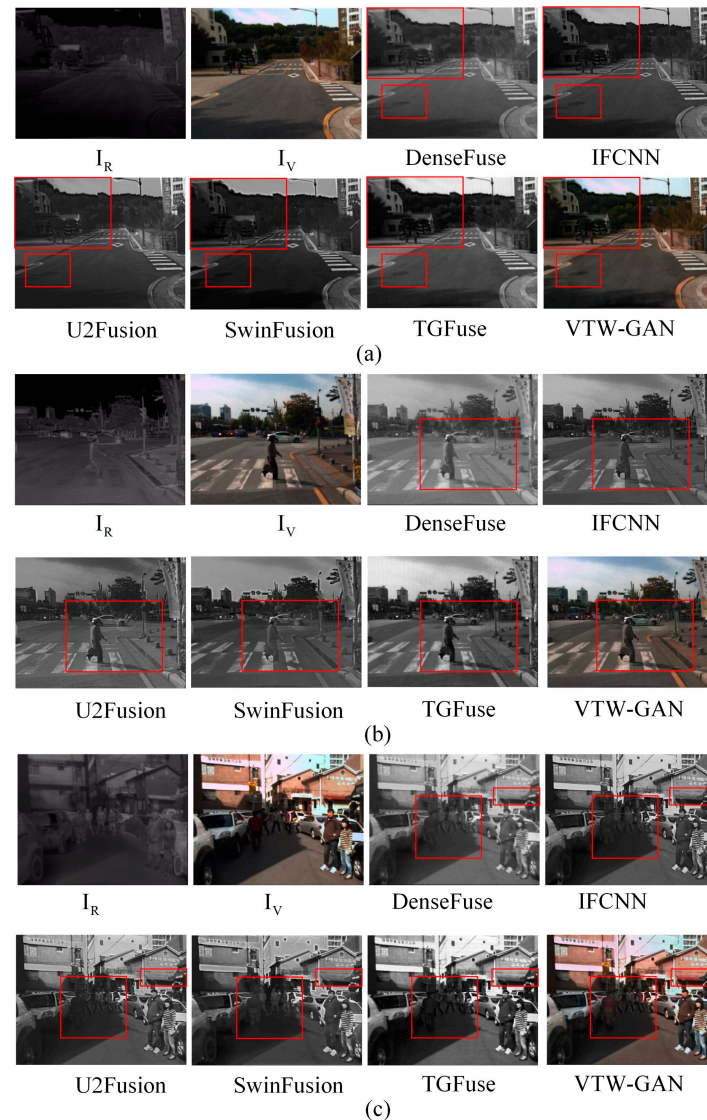


Figure 6. Experimental comparison results on the KAIST dataset: (a) street; (b) people; (c) a group of people. The marked red box is the Region of Interest (ROI) used for the result analysis.

Table 1 shows the objective evaluation of 128 pairs of test images of the KAIST dataset by each method. The optimal value and suboptimal value are shown in bold and underlined, respectively. It can be seen that the proposed method achieves the optimal performance in six indicators. Moreover, it performs well in EN , MI , and Q_{abf} , being 2.78%, 11.89%, and 10.45% higher than the suboptimal value, respectively. This shows that the VTW-GAN fusion image has more abundant information, and the correlation between different image sources in the fused image is stronger, which can better combine the characteristics of visible light and infrared light in the street with better consistency and coherence, and the fusion image has better clarity and edge retention, with better visual effect and detail presentation ability.

Table 1. Objective evaluation indicators of each method on 128 pairs of images in KAIST dataset.

Method	EN $\uparrow$	SF $\uparrow$	AG $\uparrow$	SD $\uparrow$	MI $\uparrow$	Q_{abf} $\uparrow$
DenseFuse [21]	6.949	8.098	2.941	47.442	2.538	0.476
IFCNN [22]	6.867	9.604	3.464	47.118	3.086	<u>0.622</u>
U2Fusion [23]	6.699	10.028	3.536	40.493	3.174	0.536
SwinFusion [24]	6.700	9.345	3.193	42.048	3.193	0.544
TGFuse [16]	<u>6.987</u>	<u>10.485</u>	<u>3.586</u>	<u>60.625</u>	<u>3.541</u>	0.592
VTW-GAN	7.187	10.517	3.610	61.563	3.962	0.687

“ $\uparrow$ ” shows that a larger value of the evaluation metric is better. The optimal value and suboptimal value are shown in bold and underlined.

3.3. Comparison and Analysis of Experimental Results on the Corsican Fire Dataset

Based on the strong feature extraction ability and image fusion performance of VTW-GAN on the KAIST dataset, in order to evaluate the effect of the model on the fire image, the model is trained and tested on the small dataset Corsican Fire, and the fusion result is shown in Figure 7. It can be seen that due to the scarcity of data, the model does not capture the potential structure and relationship of the image, the color and hierarchical details in the fusion result are missing and uneven, the sky and flame textures are too simple, the overall effect is unnatural, and the phenomenon of under fitting occurs.



Figure 7. Test results of VTW-GAN after training on Corsican Fire dataset: (a) infrared image; (b) visible image; (c) fusion results.

In order to solve the problem of under fitting and improve the generalization ability of the model on the Corsican Fire dataset, the pre-training model is migrated and learned. The fusion results of fire images before and after migration learning (fusion model) are shown in Figure 8, and some details in the fusion results are framed. From the comparison between (c) and (d) in the figure, it can be seen that compared with the fusion model, the fusion image generated by the transfer learning model is more detailed. The fused image can also better reflect the difference between the flame core and the edge. At the same time, the color of smoke and grass is closer to the visible light, the color and gradient of the sky are more coordinated, and the cloud and other information are effectively preserved and more natural.

To verify the effectiveness of transfer learning, the fusion model and transfer learning model are objectively evaluated and analyzed, and the results are shown in Table 2. By comparison, it is found that the indicators of the infrared image generated by the transfer learning model are improved compared with the fusion model, in which *SF* and *SD* are increased by 10.69% and 11.73%, respectively, which indicates that the color contrast of the image fused by the transfer learning model is stronger and retains a lot of information content of the original image, the clarity and brightness are improved, and the details are richer. At the same time, *EN*, *AG*, *MI*, and Q_{abf} have also been improved to varying degrees. It can be seen that the fusion image generated by the model after transfer learning has more information, higher image quality, and high consistency and relevance. On the whole, the images generated by the transfer learning model are clearer and more natural and more reflect the thermal characteristics of the fire.

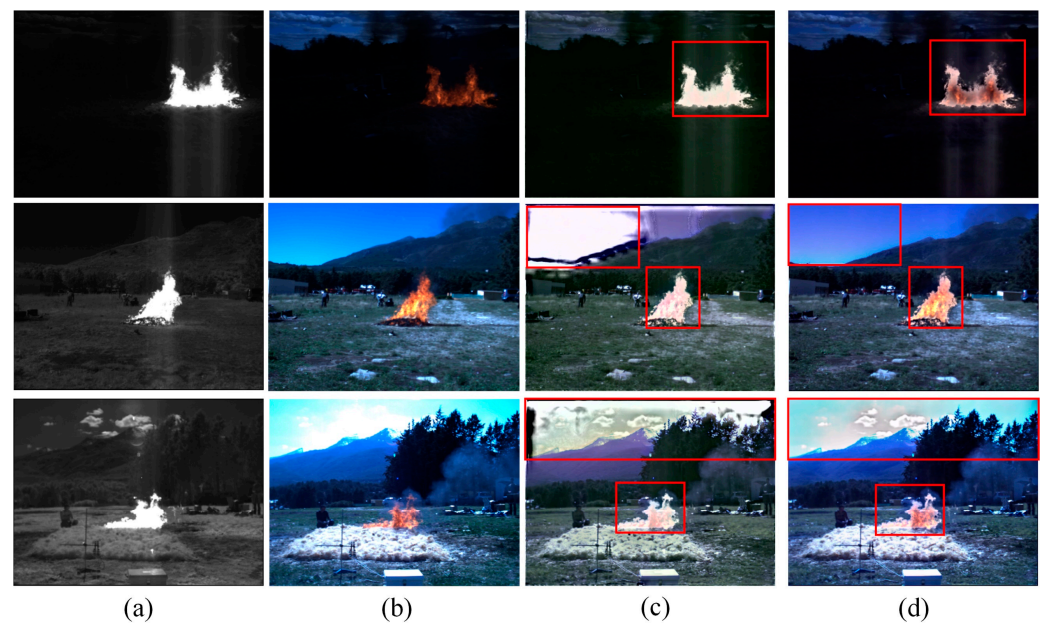


Figure 8. Subjective comparison results of fused images before and after transfer learning: (a) infrared image; (b) visible image; (c) fusion model; (d) transfer learning model. The marked red box is the ROI used for the result analysis.

Table 2. Evaluation indicators of fused images before and after transfer learning.

Framework	$EN \uparrow$	$SF \uparrow$	$AG \uparrow$	$SD \uparrow$	$MI \uparrow$	$Q_{abf} \uparrow$
Fusion Model	6.689	9.842	3.604	38.686	3.749	0.586
Transfer Learning Model	6.867	10.894	3.827	43.223	4.085	0.660

" $\uparrow$ " shows that a larger value of the evaluation metric is better. The optimal value is shown in bold.

Testing the transfer learning models on the Corsican Fire dataset, a comparison evaluation was conducted using both subjective and objective evaluation metrics. Figure 9 illustrates the subjective comparison results of three image sets. From Figure 9a,b, it can be observed that the fused images produced by DenseFuse are relatively clear but lack overall contrast, and the brightness contrast of the flame core is poor. In comparison to VTW-GAN, IFCNN, U2Fusion, and SwinFusion show less distinct differences in the flame core and edges, while TGFuse exhibits low clarity in the shape and texture of clouds. In Figure 9c, DenseFuse has lower clarity compared to other methods. SwinFusion captures finer texture details in the flame core, making the difference in grayscale within the flame barely visible in the infrared image. U2Fusion effectively preserves the infrared information but represents the shape of tree branches too sharply. TGFuse tends to emphasize visible light details over others, with blurry shadows in the white smoke and ambiguous brightness contrast between the white box and the flame edges, resulting in the loss of infrared details. In contrast, the fusion images generated by VTW-GAN exhibit a more natural fusion of trees while retaining more texture information of the flame core, smoke shadows, and white boxes.

The objective evaluation results on the Corsican Fire dataset are presented in Table 3, where the optimal and suboptimal values are indicated in bold and underlined, respectively. The proposed method achieves the optimal average values across all six evaluation metrics. Particularly, it demonstrates good performance in terms of AG , SD , and MI , with improvements of 3.43%, 4.84%, and 4.21% over the suboptimal values, respectively. This indicates that the edge information of the fire fusion images generated by VTW-GAN has been enhanced, allowing for better contours and structural features of the flames and background. Moreover, the increase in Standard Deviation enhances the contrast of fire images, enriching details such as smoke, and the improvement in Mutual Information indicates

that the fusion incorporates more details and features, resulting in a more comprehensive representation of both the infrared and visible light aspects of the fire and better consistency in the results.

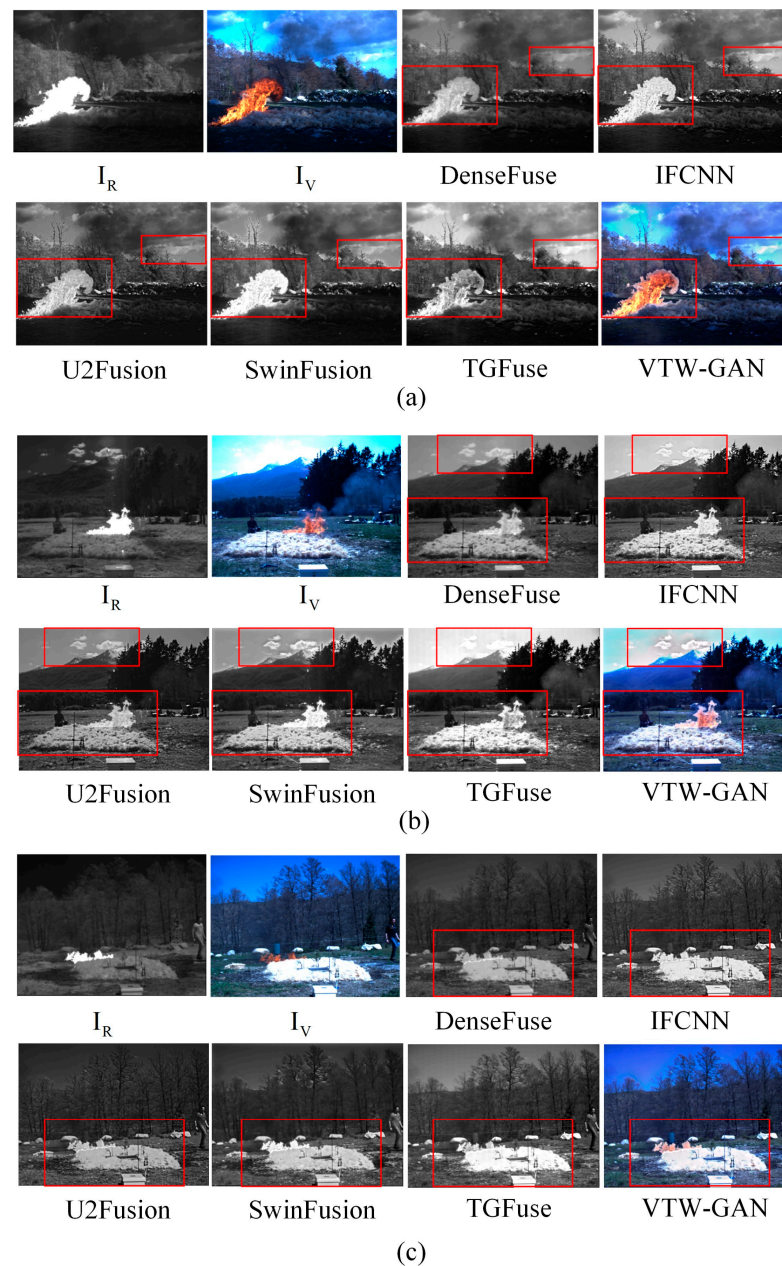


Figure 9. Experimental comparison results on the Corsican Fire dataset: (a) fire in the forest; (b) fire and clouds; (c) fire and smoke. The marked red box is the ROI used for the result analysis.

Table 3. Experimental comparison results on the Corsican Fire dataset.

Method	EN ↑	SF ↑	AG ↑	SD ↑	MI ↑	Q_{abf} ↑
DenseFuse [21]	6.548	6.375	2.517	36.300	3.280	0.367
IFCNN [22]	6.599	10.103	3.633	40.824	3.513	0.603
U2Fusion [23]	5.969	9.598	3.371	33.555	2.743	0.497
SwinFusion [24]	6.467	10.603	3.501	41.164	3.115	0.582
TGFuse [16]	<u>6.765</u>	<u>10.691</u>	<u>3.700</u>	<u>41.227</u>	<u>3.920</u>	<u>0.642</u>
VTW-GAN	6.867	10.894	3.827	43.223	4.085	0.660

“↑” shows that a larger value of the evaluation metric is better. The optimal value and suboptimal value are shown in bold and underlined.

The model's timeliness for forest-fire detection is evident in the comparison of different models' runtime results, as shown in Table 4. VTW-GAN, compared to DenseFuse, has similar performance on STD but significantly boosts efficiency by 78.3% in the mean value after transfer learning. Its Frames Per Second (FPS) value also increases from 1.449 to 6.667. Compared with TGFuse, there is a slight increase in mean runtime and a decrease in the FPS value, but the STD difference is minimal. This could be because during transfer learning fine-tuning, the model requires extra training iterations on the fire dataset, prolonging the time needed for convergence and adding computational overhead, thus reducing runtime speed. Overall, VTW-GAN maintains high efficiency while enhancing image detail features.

Table 4. Running time and FPS of different models on 128 pairs of images.

Method	DenseFuse21	IFCNN22	U2Fusion23	SwinFusion24	TGFuse16	VTW-GAN
Mean (s)	0.690	2.974	1.500	0.765	0.042	0.150
STD (s)	0.025	0.243	0.007	0.052	0.026	0.027
FPS	1.449	0.336	0.667	1.307	23.810	6.667

3.4. Ablation Study

When designing the network, a network structure was incorporated into the generator's Transformer. Wavelet-guided pooling and unpooling were added to the Spatial Transformer in the model training. To validate the effectiveness of this structure and strategy, comparative experiments were conducted based on pre-trained models using the Corsican Fire dataset as the test set. As shown in Figure 10, subjective comparisons were made among the improved Channel Transformer (VTW-GAN (c)), the improved Spatial Transformer (VTW-GAN (s)), and both improved Transformers (VTW-GAN (cs)). It can be observed that in VTW-GAN (c), due to only improvements made to the Channel Transformer, the ViT fusion module initially takes the number of image channels as input tokens for the encoder, ignoring spatial structures and local features, thus lacking the relative relationships between pixels in each channel. Although wavelet-guided pooling enhances the details in each channel, it still struggles to capture spatial relationships. The Spatial Transformer remains unchanged, resulting in the loss of some minor textures present in the infrared image, leading to the omission of tiny details in the result. On the other hand, in VTW-GAN (s), while the Channel Transformer remains unchanged and learns cross-channel relationships, the improved Spatial Transformer takes the number of image blocks as input tokens for the encoder, preserving spatial positional information and local features of different regions. Meanwhile, wavelet-guided pooling enhances the texture details of the image, making details such as clouds and flames more prominent compared to the other two cases. In VTW-GAN (cs), further improvements were made to the Spatial Transformer based on VTW-GAN (c), resulting in the loss of more infrared and visible light details, with significant differences in color compared to visible light. Objective experimental results are shown in Table 5, with bold numbers indicating the best results. From the table, it can be seen that the fusion model employed achieves the optimal values in all metrics, indicating that when only wavelet-guided pooling and unpooling are incorporated into the Spatial Transformer, the fused image contains richer information, exhibiting better image quality, clarity, and fidelity in the fusion results.

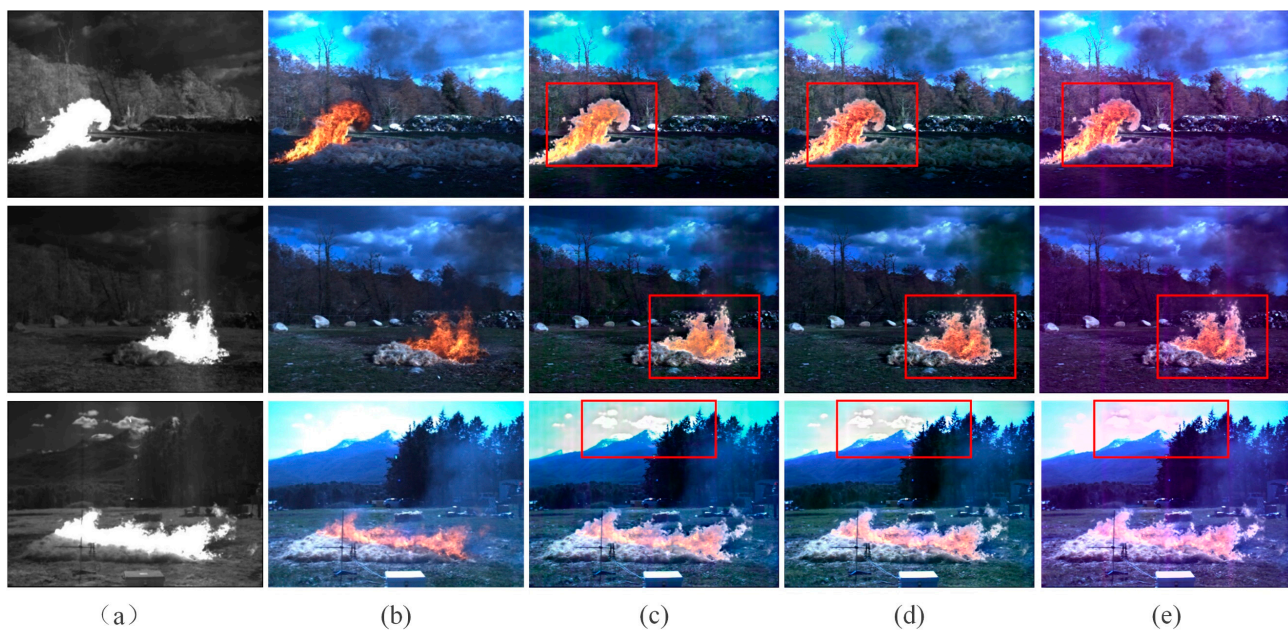


Figure 10. Results of ablation experiments on Corsican Fire dataset: (a) infrared image; (b) visible light image; (c) VTW-GAN (c); (d) VTW-GAN (s); (e) VTW-GAN (cs). The marked red box is the ROI used for the result analysis.

Table 5. Results of ablation study.

	Channel	Spatial	$EN \uparrow$	$SF \uparrow$	$AG \uparrow$	$SD \uparrow$	$MI \uparrow$	$Q_{abf} \uparrow$
Wavelet-guided pooling and unpooling	✓		6.803	10.719	3.752	41.801	3.536	0.643
		✓	6.837	10.809	3.812	42.834	3.825	0.651
	✓	✓	6.832	10.714	3.786	42.572	3.745	0.634

“↑” shows that a larger value of the evaluation metric is better. The optimal value is shown in bold.

Forest fires are highly destructive disasters that pose a serious threat to human life, ecological environments, and natural resources. Currently, most fire-monitoring methods rely on visible light for detection, but single-modal fire detection performs poorly in complex scenarios. Visible light images are susceptible to environmental factors such as low light and vegetation obstruction, leading to crucial information loss. However, VTW-GAN integrates visible light and infrared images, compensating for the deficiencies of single-image information. Experimental results show that the fused images not only contain richer visible light and infrared detail information but also exhibit better target features.

Despite progress, this study has limitations. Compared to TGFuse, VTW-GAN significantly improves image granularity expression and various objective evaluation metrics on the KAST dataset and Corsican Fire dataset; however, these improvements come at the expense of increased inference time. In future research, techniques such as model compression, quantization, and acceleration will be considered to achieve faster inference speeds, thereby enhancing the model’s performance and applicability. Additionally, while the fused images generated by VTW-GAN retain the color information of visible light images, the model’s input requires obtaining the YUV components of the images, adding complexity to the fusion process. Future research will explore how to optimize the model structure to enable the direct processing and output of three-channel fused images, thereby simplifying the fusion process and improving efficiency.

4. Conclusions

The VTW-GAN method proposed addresses the issues of detail loss, insufficient matching datasets, and low fusion accuracy in infrared/visible light fire image fusion, with

applicability to fire-recognition scenarios. This approach primarily employs a generator and discriminator network embedded with Transformer modules to achieve the fusion of infrared/visible light fire images. The model combines the efficient global representation capability of Transformers with wavelet-guided pooling for detail enhancement, extracting finer-grained features. Transfer learning is utilized to mitigate the impact of limited datasets. Through comparisons with public datasets and ablation experiments, the method's advantages in infrared and visible light image fusion for fire scenes are validated. The experimental results indicate that VTW-GAN outperforms DenseFuse, IFCNN, U2Fusion, SwinFusion, and TGFuse algorithms subjectively, with the fused images effectively enhancing clarity while retaining the thermal information characteristics of the infrared images. Objectively, on the KAIST dataset, the fused images perform well on EN , MI , and Q_{abf} , with improvements of 2.78%, 11.89%, and 10.45%, respectively, over the suboptimal values. On the Corsican Fire dataset, compared to the data-constrained fusion models, the transferred fused images show enhancements of 10.69% and 11.73% in SD and MI , respectively, with clearer detail presentation. Compared to other methods, the transferred fused images exhibit good performance in AG , SD , and MI , with improvements of 3.43%, 4.84%, and 4.21%, respectively, over the suboptimal values, while achieving a 78.3% increase in efficiency compared to DenseFuse. This technology can be applied to fire monitoring to improve the performance and accuracy of fire-detection systems, promoting the development of fire safety management technology. For example, in the future, drones equipped with VTW-GAN fusion image technology can be used to monitor forest-fire scenes in real time and transmit information back to rescue centers, enabling automated fire detection and rescue, thereby reducing the losses and hazards caused by fires.

Author Contributions: Conceptualization, H.W. and X.F.; methodology, H.W. and X.F.; software, X.F.; validation, X.F. and Z.W.; formal analysis, X.F. and Z.W.; investigation, X.F. and Z.W.; resources, H.W. and J.Z.; data curation, X.F. and Z.W.; writing—original draft preparation, X.F.; writing—review and editing, H.W., X.F., Z.W. and J.Z.; visualization, X.F. and Z.W.; supervision, H.W. and J.Z.; project administration, H.W., X.F. and J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of Ningxia (2022AAC03006); the National Natural Science Foundation of China (No. 62361001); the Leading Talent Project Plan of the State Ethnic Affairs Commission; the Ningxia Technology Innovative Team of Advanced Intelligent Perception and Control, Leading talents in scientific and technological innovation of Ningxia; The Ningxia Autonomous Region Graduate Education Reform Project “Research on the Cultivation Model of Graduate Innovation Ability Based on Tutor Team Collaboration”, (No. YJG202104); the Graduate Student Innovation Project of North Minzu University (No. YCX23141); the Ningxia 2021 Industry University Collaborative Education Project “Construction and Exploration of the Four in One Practice Platform under the Background of New Engineering”, (No. cxy2021017); and North Minzu University for special funds for basic scientific research operations of central universities [grant number 2021JCYJ10].

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Li, S.; Kang, X.; Fang, L.; Hu, J.; Yin, H. Pixel-level image fusion: A survey of the state of the art. *Inf. Fusion* **2017**, *33*, 100–112. [\[CrossRef\]](#)
2. Ma, J.; Ma, Y.; Li, C. Infrared and Visible Image Fusion Methods and Applications: A Survey. *Inf. Fusion* **2019**, *45*, 153–178. [\[CrossRef\]](#)
3. Yin, H.; Xiao, J. Laplacian pyramid generative adversarial network for infrared and visible image fusion. *IEEE Signal Process. Lett.* **2022**, *29*, 1988–1992. [\[CrossRef\]](#)
4. Mallat, S.G. A Theory for Multiresolution Signal Decomposition—The Wavelet Representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **1989**, *11*, 674–693. [\[CrossRef\]](#)

5. Li, L.; Ma, H. Pulse coupled neural network-based multimodal medical image fusion via guided filtering and WSEML in NSCT domain. *Entropy* **2021**, *23*, 591. [[CrossRef](#)] [[PubMed](#)]
6. Liu, Y.; Liu, S.P.; Wang, Z.F. A General Framework for Image Fusion Based on Multi-Scale Transform and Sparse Representation. *Inf. Fusion* **2015**, *24*, 147–164. [[CrossRef](#)]
7. Liu, Y.; Chen, X.; Wang, Z.; Wang, Z.J.; Ward, R.K.; Wang, X. Deep learning for pixel-level image fusion: Recent advances and future prospects. *Inf. Fusion* **2018**, *42*, 158–173. [[CrossRef](#)]
8. Pang, S.; Huo, H.; Yang, X.; Li, J.; Liu, X. Infrared and visible image fusion based on double fluid pyramids and multi-scale gradient residual block. *Infrared Phys. Technol.* **2023**, *131*, 104702. [[CrossRef](#)]
9. Li, G.; Qian, X.; Qu, X. SOSMaskFuse: An infrared and visible image fusion architecture based on salient object segmentation mask. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 10118–10137. [[CrossRef](#)]
10. Ding, Z.; Li, H.; Zhou, D.; Liu, Y.; Hou, R. A robust infrared and visible image fusion framework via multi-receptive-field attention and color visual perception. *Appl. Intell.* **2023**, *53*, 8114–8132. [[CrossRef](#)]
11. Jin, Q.; Tan, S.; Zhang, G.; Yang, Z.; Wen, Y.; Xiao, H.; Wu, X. Visible and Infrared Image Fusion of Forest Fire Scenes Based on Generative Adversarial Networks with Multi-Classification and Multi-Level Constraints. *Forests* **2023**, *14*, 1952. [[CrossRef](#)]
12. Rao, Y.; Wu, D.; Han, M.; Wang, T.; Yang, Y.; Lei, T.; Zhou, C.; Bai, H.; Xing, L. AT-GAN: A generative adversarial network with attention and transition for infrared and visible image fusion. *Inf. Fusion* **2023**, *92*, 336–349. [[CrossRef](#)]
13. Huang, S.; Song, Z.; Yang, Y.; Wan, W.; Kong, X. MAGAN: Multi-Attention Generative Adversarial Network for Infrared and Visible Image Fusion. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 1–14.
14. Wang, Z.; Chen, Y.; Shao, W.; Li, H.; Zhang, L. SwinFuse: A residual swin transformer fusion network for infrared and visible images. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–12. [[CrossRef](#)]
15. Tang, W.; He, F.; Liu, Y.; Duan, Y. MATR: Multimodal medical image fusion via multiscale adaptive transformer. *IEEE Trans. Image Process.* **2022**, *31*, 5134–5149. [[CrossRef](#)]
16. Rao, D.; Xu, T.; Wu, X.J. Tgfuse: An infrared and visible image fusion approach based on transformer and generative adversarial network [Early Access]. *IEEE Trans. Image Process.* **2023**. [[CrossRef](#)]
17. Yoo, J.; Uh, Y.; Chun, S.; Kang, B.; Ha, J. Photorealistic style transfer via wavelet transforms. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019.
18. Hwang, J.; Yu, C.; Shin, Y. SAR-to-optical image translation using SSIM and perceptual loss based cycle-consistent GAN. In Proceedings of the 2020 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, Republic of Korea, 21–23 October 2020.
19. Hwang, S.; Park, J.; Kim, N.; Choi, Y.; Kweon, I.S. Multispectral pedestrian detection: Benchmark dataset and baseline. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–15 June 2015.
20. Toulouse, T.; Rossi, L.; Campana, A.; Celik, T.; Akhloufi, M. Computer vision for wildfire research: An evolving image dataset for processing and analysis. *Fire Saf. J.* **2017**, *92*, 188–194. [[CrossRef](#)]
21. Li, H.; Wu, X.J. DenseFuse: A fusion approach to infrared and visible images. *IEEE Trans. Image Process.* **2018**, *28*, 2614–2623. [[CrossRef](#)] [[PubMed](#)]
22. Zhang, Y.; Liu, Y.; Sun, P.; Yan, H.; Zhao, X.; Zhang, L. IFCNN: A general image fusion framework based on convolutional neural network. *Inf. Fusion* **2020**, *54*, 99–118. [[CrossRef](#)]
23. Xu, H.; Ma, J.; Jiang, J.; Guo, X.; Ling, H. U2Fusion: A unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 502–518. [[CrossRef](#)]
24. Ma, J.; Tang, L.; Fan, F.; Huang, J.; Mei, X.; Ma, Y. SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA J. Autom. Sin.* **2022**, *9*, 1200–1217. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.