

Article

STGAN: A Fusion of Infrared and Visible Images

Liuhui Gong ¹, Yueping Han ^{1,*} and Ruihong Li ²¹ Information and Communication Engineering, North University of China, Taiyuan 030051, China; s202405075@st.nuc.edu.cn² Software Engineering, North University of China, Taiyuan 030051, China

* Correspondence: hanyueping@nuc.edu.cn

Abstract

The fusion of infrared and visible images provides critical value in computer vision by integrating their complementary information, especially in the field of industrial detection, which provides a more reliable data basis for subsequent defect recognition. This paper presents STGAN, a novel Generative Adversarial Network framework based on a Swin Transformer for high-quality infrared and visible image fusion. Firstly, the generator employs a Swin Transformer as its backbone for feature extraction, which adopts a U-Net architecture, and the improved W-MSA is introduced into the bottleneck layer to enhance local attention and improve the expression ability of cross-modal features. Secondly, the discriminator uses a Markov discriminator to distinguish the difference. Then, the core GAN framework is leveraged to guarantee the retention of both infrared thermal radiation and visible-light texture details in the generated image so as to improve the clarity and contrast of the fused image. Finally, simulation verification showed that six out of seven indicators ranked in the top two, especially in key indicators such as PSNR, VIF, MI, and EN, which achieved optimal or suboptimal values. The experimental results on the general dataset show that this method is superior to the advanced method in terms of subjective vision and objective indicators, and it can effectively enhance the fine structure and thermal anomaly information in the image, which gives it great potential in the application of industrial surface defect detection.



Academic Editors: Pei-Ju Chiang,
Cheng-Lun Chen, Ping-Huan Kuo
and Wen-Yang Chang

Received: 29 September 2025

Revised: 25 October 2025

Accepted: 27 October 2025

Published: 29 October 2025

Citation: Gong, L.; Han, Y.; Li, R.
STGAN: A Fusion of Infrared and
Visible Images. *Electronics* **2025**, *14*,
4219. <https://doi.org/10.3390/electronics14214219>

Copyright: © 2025 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/>).

Keywords: Swin-transformer; generative adversarial networks; U-Net; multi-head self-attention mechanism; Markov discriminator

1. Introduction

With the continuous progress of sensing technology, combining infrared and visible light images has emerged as a crucial research area aimed at enhancing data utility and image quality [1]. Infrared imagery captures thermal radiation properties, making it effective for target identification in low-light conditions, haze, or intricate environments. In contrast, visible light images offer rich texture and detailed information, ideal for portraying scene structures and subtle imperfections [2,3]. Presently, the technology that fuses infrared with visible light is extensively applied in numerous domains, including image enhancement, target recognition, detection, tracking, agricultural automation, and remote sensing detection [4–6].

The past few decades have witnessed a proliferation of techniques developed for the fusion of infrared and visible images. These approaches can be classified by their core principles into distinct categories: multi-scale transformation techniques, via wavelet-based pixel and region integration techniques [7]; sparse representation approaches; neural

network paradigms; subspace methodologies; saliency-driven strategies; hybrid models; and other assorted fusion techniques. A high-quality fusion image integrates the high-definition texture of visible light with the thermal distribution of infrared, which can significantly enhance defect features and lay a solid foundation for subsequent detection algorithms. However, traditional fusion methods [8–12], which have shaped the field, face significant challenges in retaining such key yet subtle features. They rely heavily on manually designed extraction processes and predefined fusion rules, adding complexity to the fusion pipeline. This often results in homogeneous features, leading to fused images with reduced contrast, blurry textures, and artifacts.

Conventional image fusion techniques frequently struggle with intricate textures, low contrast, and cross-modal characteristics, which pose challenges in balancing the retention of detail and the prominence of infrared thermal anomalies [13]. The image fusion approach utilizing Generative Adversarial Networks (GANs) focuses on constructing more sophisticated generator architectures. They primarily employ a singular feature extraction approach for the source image, resulting in an image that often excels locally. Nevertheless, this approach tends to overlook the abundant semantic information inherent in the source image, leading to blurred edges in the fused output [14]. Moreover, GANs notoriously encounter issues like unstable training.

In parallel, recent research trends have explored new paradigms to overcome these limitations. Denoising diffusion models, for instance, have shown remarkable capability in generating high-fidelity images by learning a reverse diffusion process, demonstrating state-of-the-art performance in tasks like infrared and visible image fusion [15]. While these methods achieve impressive results, they often come with increased computational complexity and longer inference times.

Therefore, this paper focuses on a more effective fusion algorithm, whose core goal is to generate fusion results with richer details, more prominent thermal targets, and higher visual quality while maintaining computational efficiency. To resolve these challenges, this study introduces a Swin Transformer-based Generative Adversarial Network (GAN) for superior infrared–visible image fusion. Not only can it effectively extract thermal defect features from infrared images, but it can also maintain the spatial detail information of visible images, providing a clearer and more discriminative image foundation for defect detection tasks. Compared with traditional single-modal image detection methods, the fused image can greatly improve defect recognition and achieve more accurate defect localization, especially in the recognition of subtle cracks and early aging phenomena, where its performance is particularly significant [16]. This model adopts a U-Net [17] encoder bottleneck layer decoder structure design, and the design of our dual-window multi-head self-attention (DWMSA) module follows a unified convolution-and-attention paradigm, where the self-attention branch captures cross-window dependencies and the convolutional FFN branch enhances local feature representation, effectively unifying the strengths of both approaches [18]. In the decoding stage, the network first decodes the encoded features layer by layer, and through window partitioning and reassembly operations, recombines the features into a fused feature map of the same size as the original input image to ensure consistency in the spatial resolution of the output results. In order to improve the authenticity and detail fidelity of the generated results, this paper adopts a discriminator based on Markov random fields [19–21] as the adversarial module, which imposes finer-grained constraints on the generator. By establishing an adversarial game to fully train the generator, the image fusion effect can be improved.

Although this paper does not directly train and test the defect detector, the performance of our method will be indirectly proved by its ability to enhance the weak thermal target and complex texture on the universal benchmark dataset (MSRS, RoadScene). We

firmly believe that this basic work will provide strong technical support for solving the problems in actual industrial testing, and it aligns with the future directions of image fusion applications outlined in recent comprehensive surveys [22].

2. Methods

Due to the differences in imaging mechanisms between infrared sensors and visible light sensors, images captured from the same scene using these sensors typically contain rich complementary information. By integrating this information into one image, the comprehensiveness of scene depiction can be significantly enhanced. The fused image can therefore be utilized in various downstream computer vision tasks. In this work, the MSRS dataset is employed for training, including less bright nighttime images and bright daytime images, and the dataset has been physically aligned and grouped. RoadScene, as a validation dataset, includes 221 pairs of infrared–visible images that contain rich scene information, including pedestrians, vehicles, roads, etc.

2.1. Overall Structure

This work proposes STGAN, A SwinGAN framework for infrared–visible image integration. The framework comprises three core components: a generator, a discriminator, and a composite loss function, combined with the Swin Transformer’s powerful local self-attention and hierarchical feature extraction mechanism, enabling the fused image to preserve visible light’s resolution and infrared’s thermal data. The framework is shown in Figure 1.

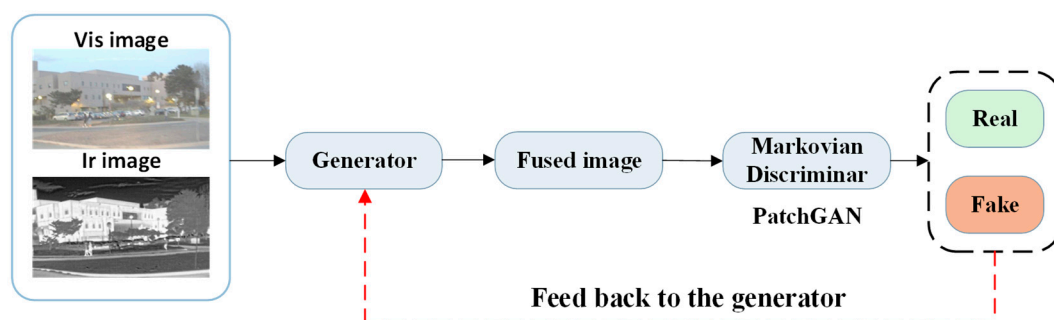


Figure 1. The model framework is based on the workflow of a visible (Vis) and infrared (Ir) image fusion system using Generative Adversarial Networks. It utilizes adversarial training with generators and Markov discriminators (using PatchGAN structure) to obtain high-quality fused images.

2.1.1. Generator

The generator uses a Swin Transformer as the basic unit of encoder and decoder, adopts a U-shaped structure, and uses an improved W-MSA for feature alignment and fusion at the bottleneck layer. We input the infrared image (IR) and the visible light image (VIS) and then concatenate the three-channel visible light image and the single-channel infrared image to form a four-channel input image. The input layer has four channels, and the encoder consists of patch segmentation, a linear mapping module, and three stages. Each stage is composed of a Swin TRM Block and down-sampling layer, which divides visible and infrared images into multiple patches through Patch Partition. Convolution is used instead of slicing to obtain 4×4 patches with a feature dimension of $C = 96$. These patches are then mapped to a new feature space using a linear embedding layer [23,24], and then they enter the down-sampling stage, which is composed of a Swin Transformer Block and a down-sampling layer. Swin Transformer Blocks consist of four stages, each containing multiple Swin Blocks (DWMSA + MLP + LN + skip). The default channel output of Swin is 768, and the bottleneck part (1×1 Conv) is down-converted and uniformly sent

to the decoder. At this time, the number of channels is 512. The decoder is equivalent to the inverse process of the encoder, and skip connections are added between the corresponding encoding and decoding layers to fuse the feature maps. The decoder gradually up-samples to restore the original image resolution and generates a fused image. The output is a three-channel fused image. Figure 2 illustrates the generator's fundamental design.

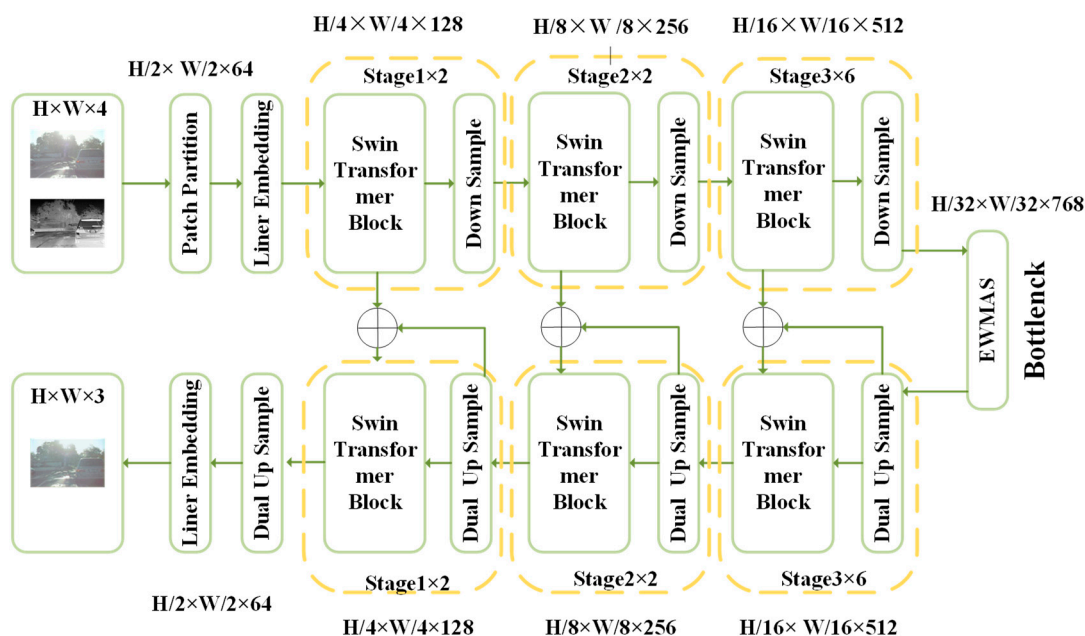


Figure 2. The generator uses a U-shaped structure to preprocess the input multimodal image, such as blocking and linear embedding. The features are extracted by the multi-stage subsampling module containing the Swin Transformer Block, and then the image resolution is gradually restored by the corresponding subsampling module. After combining with the EWMSA bottleneck and other modules, the fused reconstructed image is finally output.

We added a dual-window multi-head self-attention module to the bottleneck layer to improve the traditional window self-attention mechanism, W-MSA, and compensate for the shortcomings in processing local attention. Figure 3 illustrates the overall architecture.

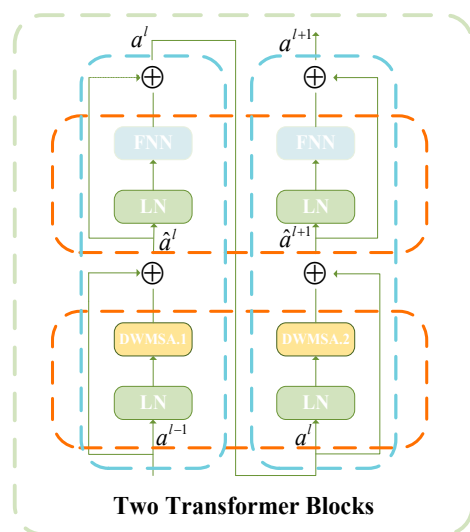


Figure 3. The module contains hierarchical computing units, including feedforward neural network (FNN), layer normalization (LN), dual-window multi-head self-attention (DWMSA, divided into DWMSA.1 and DWMSA.2), multi-layer perceptron (MLP), and other components.

Before the DWMSA module, the input feature map is divided into multiple windows and normalized. After normalization, the features are input into DWMSA. Unlike VIT, VIT adds a positional code when calculating embeddings and adds a relative positional code when calculating MSA. The total attention formula is as follows:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_K}} + B\right)V \quad (1)$$

Among them, B is the relative position encoding; d is the mapping matrix; and the matrices Q , K , and V represent the query, key, and value components, respectively, within each local window.

As shown in Formulas (2)–(5), feature a will be processed in the bottleneck layer through the DWMSA module (i.e., DWMSA.1 and DWMSA.2) with two parameters that are not shared in turn and its supporting feedforward network. This continuous non-shared weight design allows the model to carry out deeper and more diversified feature transformation and enhancement.

$$\hat{a}^l = DWMSA.1\left(LN(a^{l-1})\right) + a^{l-1} \quad (2)$$

$$a^l = MLP\left(LN(\hat{a}^l)\right) + \hat{a}^l \quad (3)$$

$$\hat{a}^{l+1} = DWMSA.2\left(LN(a^l)\right) + a^l \quad (4)$$

$$a^{l+1} = MLP\left(LN(\hat{a}^{l+1})\right) + \hat{a}^{l+1} \quad (5)$$

The calculation process is gradually advanced through residual connection ($\oplus$) and normalization operation: first, DWMSA and LN are used to process the previous state feature a^{l-1} to obtain $\hat{a}^l$; then, a^l was obtained by FNN, LN, etc. The follow-up modules, by analogy, complete the iterative update of features through $\hat{a}^{l+1}$ and a^{l+1} , and they realize the depth coding and representation optimization of input information.

The core of the bottleneck layer of the generator is our proposed dual-window multi-head self-attention module. Its detailed structure is shown in Figure 4. This module is called ‘Dual’ to reflect its enhanced design in two aspects: First, as shown in Figure 3, the two DWMSA blocks process features in parallel in the bottleneck layer to capture more comprehensive context information. Secondly, within each DWMSA, we replace the standard feedforward network with an enhanced feedforward network. This network introduces a 3×3 deep convolution layer between the traditional two-layer 1×1 convolution. This design enables the module to explicitly enhance the modeling of local spatial features and details while obtaining the dependencies in the global window through the self-attention mechanism. This kind of attention and convolution work together, which makes it especially suitable for image fusion tasks that need to maintain fine details and structure.

2.1.2. Discriminator

The PatchGAN discriminator divides the image into several small blocks ($n \times n$), independently determines the authenticity and forgery of each block, and takes the weighted average of the decision results of each block as the final output. This design can reduce the number of model parameters and improve the training speed while ensuring the local definition. The target of the discriminator is to determine whether the local area (patch) in the image is true (training data) or forged (output of generator g). Through continuous discrimination, it cannot be easily distinguished visually, making the image closer to the real image.

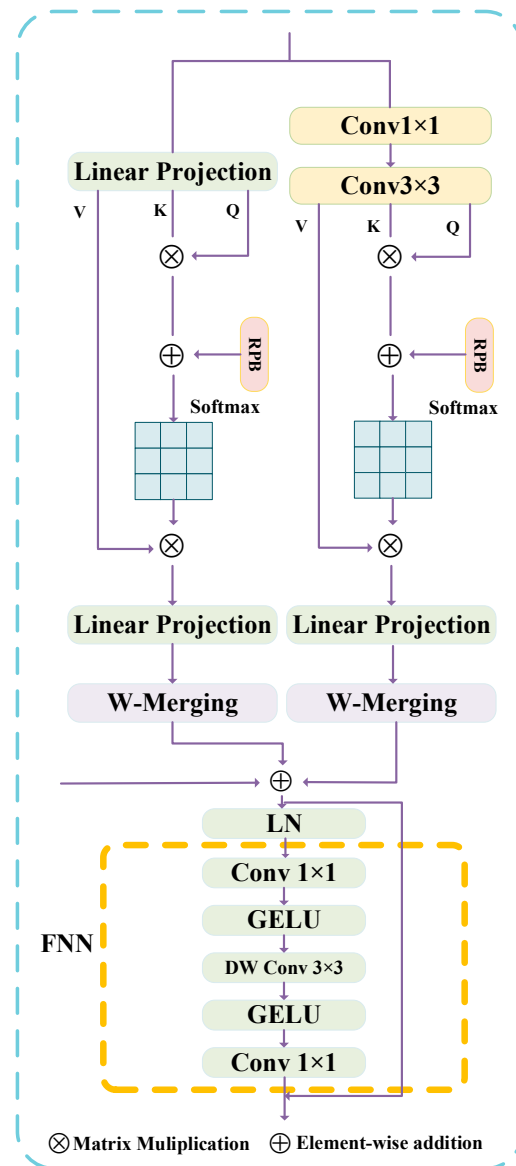


Figure 4. The output of self-attention branch and FNN are fused through element-wise addition to achieve the effect of “self-attention extracting global correlation, convolution extracting local details, residual connection maintaining information”, and to improve the ability of feature expression.

The discriminator of this model uses a Markov discriminator, which effectively models the image as a Markov random field. The discriminator adopts a Markovian (PatchGAN) architecture with a receptive field of 70×70 pixels. This design allows it to model the image as a Markov random field and focus on the authenticity of local patches, providing detailed gradient feedback to the generator and effectively suppressing modality-specific artifacts. Assuming that the independence between pixels is greater than the patch diameter, it uses a four-layer strided 2 convolution architecture. Each layer is composed of a convolution, following batch normalization and leakyReLU activation function with a slope of 0.2. The output of the last layer is mapped to the $[0, 1]$ interval through the sigmoid function so as to judge the authenticity of the image block, as shown in Figure 5.

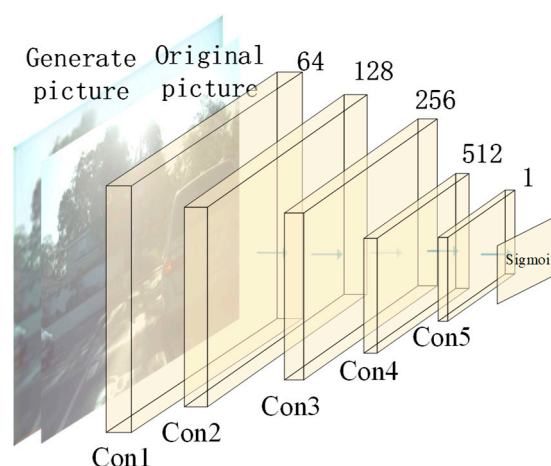


Figure 5. The target image reference to be generated or the original image to be processed represents the input source of the model. From left to right, the input image is feature-extracted and dimension-transformed through a series of convolution layers (Con1 to Con5). The feature is mapped to a single channel through a convolution operation, and then the output value is mapped to the 0–1 interval through a sigmoid activation function to determine whether the pixel belongs to the generation target region.

2.2. Loss Function

In the training process of GANs, the loss function plays an important role. In order to take into account the global authenticity, local details, edge information, and multi-scale structure consistency of the generated image, this paper designs a total loss composed of Charbonnier loss, confrontation loss, edge loss, and multi-scale structure similarity loss (MS-SSIM), and sets the weights as 0.2, 0.1, 0.4, and 0.3, respectively:

$$L_{sum} = \lambda_1 L_{char} + \lambda_2 L_{Gan} + \lambda_3 L_{edge} + \lambda_4 L_{mix} \quad (6)$$

$\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are used to adjust the balance parameters between each loss function.

Because the fusion of infrared and visible images is a typical unsupervised learning problem, there is no unique real fusion image as a supervisory signal. Therefore, the training of this framework does not depend on any real value tags. The ‘reference source image’ in this loss function directs the fusion result F (i.e., x in the formula) to preserve and incorporate critical input source image information. Specifically, we use the visible image as the reference image y in Formulas (7) and (10) to retain rich texture details; at the same time, the infrared image is used as the reference image y in Formula (9) to highlight the significant edge structure of the thermal target.

L_{char} is the Charbonnier loss function, also known as the pseudo Huber loss function, which is a differentiable function used to smooth the approximation of the L1 norm. This model is verified by experiments that setting its balance parameter to 0.2 can better maintain the consistency of basic pixels and details. The Charbonnier loss function usually smoothes the gradient by introducing an ε factor. The formula is as follows:

$$L_{char} = \sqrt{(x - y)^2 + \varepsilon^2} \quad (7)$$

where X represents the fused image, Y represents the reference source image, and the value of the smoothing parameter ε is set to 5×10^{-4} .

L_{Gan} is the countermeasure loss function in the generation network, which is used for the dynamic game between the generator and the discriminator. This model is set to 0.1

after experimental verification to avoid the instability of early training, and also to make the fusion image realistic but not out of control. The formula is as follows:

$$\min_G \max_D L_{Gan}(D, G) = E_x[\log D(y)] + E_z[\log (1 - D(G(z)))] \quad (8)$$

where G is the generator, D is the discriminator, generator g minimizes $\log(1 - D(G(z)))$, and discriminator D maximizes the real data score $\log D(y)$ and generates the data score $\log(1 - D(G(z)))$.

L_{edge} is the edge loss function of gradient information. This model is verified by experiments, and its balance parameter is set to 0.4. By explicitly constraining the consistency of the image in the gradient domain, the ability of the model to maintain the edge and structure is improved, and the edge consistency is enhanced. The formula is as follows:

$$L_{edge}(x, y) = \sqrt{(|\Delta x - \Delta y|)^2 + \varepsilon^2} \quad (9)$$

where $\Delta(*)$ represents Laplacian.

$L_{MS-SSIM}$ is a multi-scale structure similarity loss function. By comparing the image structure similarity under different resolutions and evaluating the multi-scale structure similarity, the subjective quality of the image generation task is significantly improved. In this model, the balance parameter is set to 0.3, and the formula is as follows:

$$L_{MS-SSIM}(x, y) = 1 - MS - SSIM(x, y) \quad (10)$$

3. Experimental Results and Analysis

This section first introduces the experimental setup, then introduces the dataset used in the experiment and the seven fusion indexes used for evaluation and comparison with the other five models, and finally presents the overall performance of the method and the ablation experiment content.

3.1. Experimental Setup

In this study, the model training, reasoning, and performance verification are based on high-performance hardware and an adaptive software environment to ensure stability and reproducibility. In terms of hardware, the dual GPU parallel computing architecture, together with an Intel Core i9-13900k processor and 128 GB ddr5 memory, can not only meet the display and memory requirements of 512×512 , 1024×1024 , and other high-resolution image multi-scale feature extraction, but can also improve the experimental efficiency through efficient data reading and writing and parallel processing. The computational cost of the proposed STGAN is 158 GFLOPs, with an inference time of 45 ms (22 FPS) for a 512×512 image on this hardware, demonstrating its potential for near-real-time applications. In terms of software and framework, the operating system is Microsoft Windows 11. Python 3.11 is used as the development language to build the model based on the PyTorch deep learning framework, which provides good support for Transformer class structure and optimizes the efficiency of the automatic hybrid precision training function. The software and hardware parameters of the experimental platform are detailed in Table 1.

Training Configuration: To ensure the reproducibility and stability of our experiments, we fixed the random seeds for PyTorch, NumPy, and Python's built-in random module. The model was trained from scratch using the Adam optimizer with an initial learning rate of 1×10^{-4} and β of (0.9, 0.999). We utilized a cosine annealing scheduler to progressively decrease the learning rate during training, with the model undergoing 100 epochs at a batch size of 8. To forge a powerful generator capable of robust feature extraction and reconstruction, it was first trained independently for 50 epochs using only the content losses

(L_{char} , L_{edge} , $L_{MS-SSIM}$). This critical phase ensures the generator masters the fundamental fusion task and establishes a stable convergence point before the adversarial fine-tuning with the discriminator begins. This strategy effectively mitigates mode collapse and aligns with our design philosophy of a generator-driven fusion process. The specific hyperparameters (loss function weights $\lambda_1 = 0.1$, $\lambda_2 = 0.2$, $\lambda_3 = 0.4$, $\lambda_4 = 0.3$) were determined through an ablation study on a validation set and remained fixed throughout all comparative experiments. All results reported in this paper are from a single training run evaluated on the fixed test set, which is the standard practice in the image fusion field to provide a deterministic performance comparison. The low standard deviations shown in Table 1 further confirm the stability of our method across the entire test dataset.

Table 1. Configuration of the experimental platform.

Category	Component	Specification/Version
Hardware	CPU	Intel Core i9-13900K
	GPU	1 × NVIDIA GeForce RTX 3090 (24 GB) & 1 × NVIDIA GeForce RTX 4090
	RAM	128 GB DDR5
	Storage	1 TB SSD
Software	Operating System (OS)	Microsoft Windows 11
	CUDA	CUDA 11.8
	cuDNN	cuDNN 8.2.1
	Programming Language	Python 3.11
	Deep Learning Framework	PyTorch 1.13.1
	IDE	PyCharm 2023.1.4

For all comparative experiments, the evaluation metrics for every method (including baselines) were computed under the same conditions on our experimental platform to guarantee fairness.

3.2. Datasets

This study aims to develop a fusion algorithm that can enhance the key visual features related to defects. These features include:

- Weak thermal anomalies (analogous to industrial hot spots);
- Subtle linear and textured structures (analogous to surface scratches and cracks).

For this purpose, we selected MSRS and RoadScene datasets. The enhancement challenge of long-distance pedestrians/vehicles in MSRS is consistent with the underlying visual requirements for detecting low-contrast hot spots on photovoltaic panels. Similarly, the complex texture of urban buildings in RoadScene puts forward high requirements for the ability of the algorithm to retain details that may be characterized as cracks. Therefore, the validation of our algorithm on these datasets is sufficient to prove its effectiveness in enhancing defect-related features.

A strict cross-dataset protocol was employed: the model was trained exclusively on the MSRS dataset and evaluated on the separate RoadScene and TNO datasets to assess generalization performance.

- **MSRS dataset [25–27]:** The MSRS dataset contains 1444 paired high-quality infrared and visible images and served as our training set, including less bright nighttime images and bright daytime images, and the dataset has been physically aligned and grouped. The dataset mainly focuses on roads and remote sensing scenes, and its infrared images are rich in the defects of thermal target feature blur and background

thermal interference. For example, in some images, the temperature difference between the pedestrian or vehicle and the environment is small, resulting in its unclear outline in the infrared channel; at the same time, the widely existing thermal reflection areas in the scene (such as hot ground and building glass) form a lot of interference noise. These defects provide an ideal test benchmark to verify whether our algorithm can enhance the weak thermal target and suppress the background thermal noise.

- **RoadScene dataset [28]:** As a validation dataset, it includes 221 pairs of infrared visible image pairs that contain rich scene information, including pedestrians, vehicles, roads, etc. The dataset contains a large number of urban street registration image pairs, and its visible image often has artifacts and blurring caused by uneven illumination, weather changes, and object motion. These are the typical defects of visible light images. Using this dataset, we can effectively evaluate whether our fusion method can use infrared information to compensate for these defects while preserving the clear visible light texture, so as to generate more detailed and clearer fusion results.
- **TNO dataset [29]:** As an independent test set for generalization analysis, it comprises a variety of military and surveillance scenarios, including personnel, vehicles, and equipment in diverse environments. The image pairs in this dataset are characterized by significant spectral differences and challenging conditions, such as low-contrast and complex backgrounds. These scenarios are fundamentally distinct from the road-centric views of MSRS and RoadScene, presenting a rigorous benchmark for testing cross-domain robustness. The use of this dataset allows us to verify whether our fusion method can maintain its performance advantages when confronted with entirely unfamiliar scene distributions, thereby validating the generalizability of the learned fusion strategy.

3.3. Parameter Validation

To determine the optimal combination of the weight parameters λ_1 , λ_2 , λ_3 , and λ_4 in the loss function, we conducted a detailed parameter validation study. The experimental setup was as follows: The training set consisted of 1000 randomly selected infrared and visible image pairs from the MSRS dataset, and the validation set comprised 100 randomly selected image pairs. The batch size was set to 4, and the models were trained for 50 epochs. We fixed $\lambda_1 = 0.1$ and $\lambda_2 = 0.2$, then systematically varied the values of λ_3 and λ_4 (under the constraint $\lambda_3 + \lambda_4 = 0.7$). Six representative experimental results are summarized in Table 2.

Table 2. Parameter comparison experimental results for loss functions.

	Parameter Setting				Evaluation Index			
	λ_1	λ_2	λ_3	λ_4	PSNR/dB	SSIM	MI	VIF
1	0.1	0.2	0.1	0.6	69.92	0.851	2.859	0.586
2	0.1	0.2	0.2	0.5	69.89	0.855	2.796	0.586
3	0.1	0.2	0.3	0.4	62.02	0.849	2.841	0.582
4	0.1	0.2	0.4	0.3	69.90	0.859	2.903	0.583
5	0.1	0.2	0.5	0.2	69.92	0.856	2.844	0.585
6	0.1	0.2	0.6	0.1	61.02	0.858	2.843	0.583

Table 1 shows that parameter configuration $\lambda_3 = 0.4$, $\lambda_4 = 0.3$ (Experiment 4) delivered optimal results across various metrics. This pairing achieved optimal results across both Structural Similarity (SSIM) and Mutual Information (MI) metrics, while maintaining highly competitive performance in Peak Signal-to-Noise Ratio (PSNR) and Visual Information Fidelity (VIF). This indicates that this particular parameter set effectively strikes the

optimal balance between preserving the salient targets from the infrared image and the rich texture details from the visible image. Consequently, we finalized the loss function weights as $\lambda_1 = 0.1$, $\lambda_2 = 0.2$, $\lambda_3 = 0.4$, and $\lambda_4 = 0.3$ for all subsequent comparative and ablation experiments.

3.4. Comparative Experiment

A rigorous cross-dataset evaluation protocol was adopted to comprehensively assess the proposed algorithm's modal balance and generalization capability. The model, trained exclusively on the MSRS dataset, was tested on four pairs of infrared and visible images with significant spectral differences, selected from the independent RoadScene dataset.

The RoadScene dataset, sourced from urban traffic scenes, presents substantial challenges due to its complex environments and high-density details (e.g., people, cars, and buildings), which rigorously test the network's ability to capture and preserve textures. Furthermore, the presence of dynamic objects like pedestrians and vehicles increases the difficulty of maintaining consistent visual perception. The performance was quantitatively assessed using seven fusion metrics to ensure an objective comparison.

This paper selected five mainstream methods widely used in image fusion research (ADF, DLF, FusionGAN, PIA Fusion, Dense Fuse) as comparison objects and quantitatively evaluated the output results of each method by combining the seven image fusion evaluation indicators. The four sets of images selected in the RoadScene dataset were fused with the other five models. Four sets of images correspond to Figure 6.

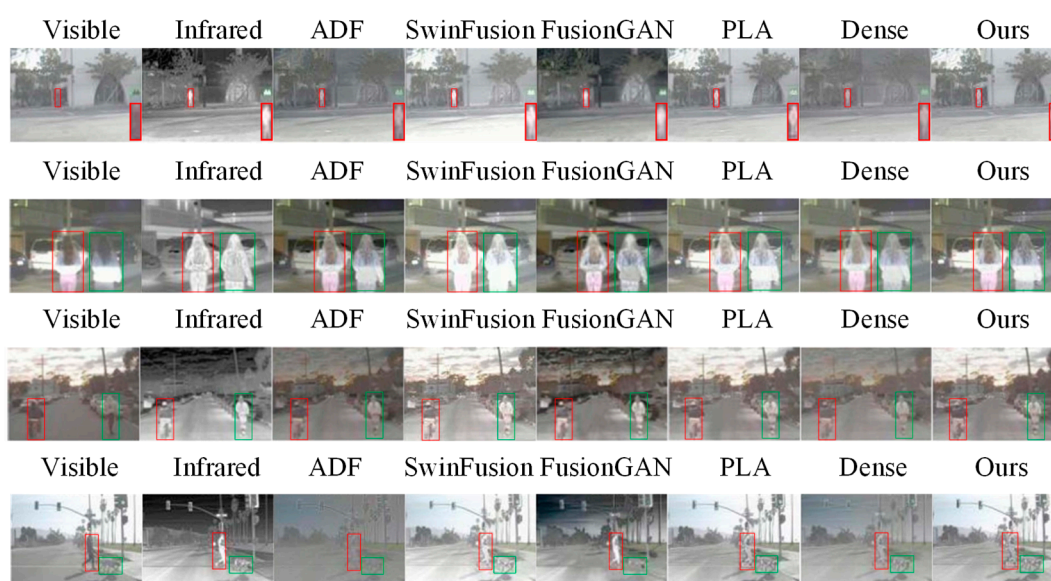


Figure 6. Qualitative comparison on representative image pairs from the RoadScene dataset. The examples were selected to cover typical challenges, including scenes with rich textures, salient thermal targets, and significant spectral differences between modalities. Each row shows the visible and infrared source images, followed by fusion results from ADF, SwinFusion, FusionGAN, PIA, Dense Fuse, and our method. As highlighted in the red and green boxes, our method better preserves the details and textures from the visible image while effectively retaining the salient infrared targets (e.g., people and objects).

The processed results are objectively evaluated with seven evaluation metrics, including PSNR, SSIM, SF, FMI, MI, VIF, and EN. Table 3 summarizes the average objective evaluation scores obtained on the RoadScene dataset, with the top-performing results for each metric emphasized in bold.

Table 3. Objective average scores of different fusion algorithms on the RoadScene dataset.

Fusion Methods	PSNR	SSIM	SF	FMI	MI	VIF	EN
ADF	67.928	0.627	4.971	0.943	2.748	0.695	7.011
SwinFusion	68.322	0.771	5.369	0.943	3.265	0.771	7.078
FusionGAN	67.351	0.726	3.398	0.937	2.755	0.578	7.034
PIA Fusion	68.496	0.796	6.167	0.945	3.214	0.762	7.125
Dense Fuse	67.349	0.631	3.401	0.939	2.922	0.654	6.823
STGAN	68.752	0.776	5.689	0.945	3.375	0.773	7.132
Average	68.033	0.721	4.833	0.942	3.047	0.706	7.034

From the perspective of subjective visual evaluation, STGAN performs well in multiple typical fusion scenarios. The composite image masterfully preserves the intricate details of the visual spectrum while seamlessly incorporating the distinctive high-energy elements from the infrared perspective, featuring flawless edge blending, well-balanced contrast, and an authentic visual experience.

Compared with other methods—such as FusionGAN’s artifact problem, Dense Fuse’s fuzzy representation, and ADF’s low contrast output—STGAN fusion images have significant advantages in structural integrity, detail restoration, and information richness. The qualitative comparison in Figure 6 serves as a subjective visual assessment, confirming that our method produces more natural and informative fusion results with clearer details and fewer artifacts, which aligns with the objective metrics. These high-quality images can be used for image processing in photovoltaic module defect detection, where the accuracy of defect localization is expected to be greatly improved.

In addition, the fusion algorithm presented in this article demonstrates excellent cross-modal coordination ability, effectively avoiding modal bias and achieving more natural cross-modal feature fusion results. It is undoubtedly more competitive than many other representative fusion technologies.

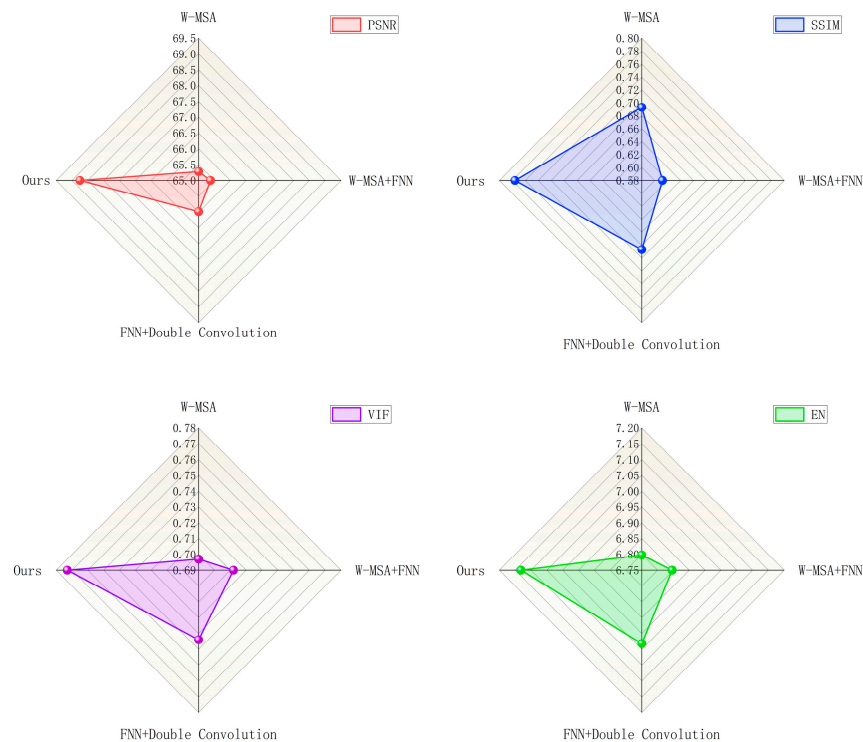
STGAN performs the best among all comparison methods, ranking in the top two for six out of seven indicators, especially achieving optimal or suboptimal values in key indicators such as PSNR, VIF, MI, and EN. This demonstrates that STGAN effectively and faithfully retains the detail and structural information from the source images during fusion, while effectively suppressing redundancy and noise components. Spatial frequency (SF) and information entropy (EN) are important indicators for measuring the texture details and information richness of fused images. STGAN’s data on SF and EN demonstrate strong expressive power in texture expression, resulting in clearer fusion results and sufficient detail preservation. STGAN is significantly superior to most traditional methods in SSIM, indicating its clear advantage in maintaining image structure consistency, especially suitable for downstream tasks that are sensitive to scene geometry, such as object detection and semantic segmentation. Although PIA Fusion has slightly higher SSIM, STGAN outperforms in metrics such as SNR, VIF, and MI. This indicates that STGAN is more stable in signal restoration and visual fidelity.

3.5. Ablation Experiments

To verify the effectiveness of each module, we conducted ablation studies under a cross-dataset evaluation protocol: the model was trained on the entire MSRS dataset and then tested on the independent RoadScene dataset. The results of these ablation experiments are presented in Table 4 and Figure 7.

Table 4. Evaluation results of ablation experiments.

Model	W-MSA	FNN	Double Convolution	PSNR	SSIM	VIF	EN
1	✓			65.278	0.693	0.697	6.798
2	✓	✓		65.376	0.612	0.712	6.846
3		✓	✓	65.984	0.687	0.734	6.982
Ours	✓	✓	✓	68.752	0.776	0.773	7.132

**Figure 7.** This set of radar images compares the performance of different methods in visible–infrared image fusion tasks based on four key evaluation indicators: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM), Visual Information Fidelity (VIF), and entropy (EN).

It can be seen that Model 2 adopts a combination of W-MSA and feedforward neural network FNN, and the PSNR, VIF, and EN of the four evaluation indicators are higher than those of Model 1 using only W-MSA, indicating that adding FNN can effectively improve image quality. The PSNR of Model 3 is 0.687 dB higher than that of Model 2, indicating that adding double convolution can more effectively capture image details, and SSIM is correspondingly improved.

This also indicates that it has a certain positive effect on preserving image structure similarity. Finally, when W-MSA, FNN, and double convolution are used together, the model achieves the best performance in all evaluation indicators, proving that the combination of the three can improve the model's performance.

3.6. Generalization Capability Analysis

To thoroughly assess the STGAN's broad applicability and reliability across different datasets, we conducted an additional cross-dataset experiment. Following the same protocol where the model was trained only on the MSRS dataset, we directly evaluated its performance on the TNO dataset, a benchmark that is fundamentally different from both MSRS and RoadScene in terms of scene content (e.g., military operations, personnel, and equipment in various environments).

The quantitative comparisons on several key metrics are summarized in Table 5. It can be observed that our STGAN method consistently achieves highly competitive performance, securing the top rank in the majority of metrics. Notably, it attains the highest scores in Visual Information Fidelity (VIF) and Mutual Information (MI), indicating its superior ability in transferring source information and producing visually pleasing results, even in this unseen domain. This provides compelling evidence that our model has learned a robust and generalizable image fusion principle.

Table 5. Quantitative comparison of generalization ability on the TNO dataset.

Fusion Methods	PSNR	SSIM	SF	FMI	MI	VIF	EN
ADF	65.632	0.725	9.451	0.843	1.528	0.315	6.321
DLF	65.987	0.738	10.124	0.851	1.678	0.335	6.458
FusionGAN	64.895	0.696	8.923	0.829	1.486	0.299	6.194
PIA Fusion	66.254	0.761	10.567	0.865	1.724	0.353	6.587
Dense Fuse	65.123	0.618	7.124	0.835	1.617	0.321	6.277
STGAN	66.147	0.752	10.589	0.865	1.813	0.369	6.629
Average	65.673	0.715	9.463	0.848	1.641	0.332	6.411

The experimental results on the TNO dataset lead to two key conclusions:

- **Strong Generalization:** The superior performance of STGAN on TNO, which is disparate from its training data, unequivocally demonstrates that it has learned a generalized fusion strategy that is not over-fitted to the characteristics of the MSRS dataset.
- **Consistent Advantage:** The model maintains its core advantages—effective preservation of thermal targets and rich texture details—across different data domains. This consistency underlines the robustness of our network architecture and loss design.

In summary, this generalized evaluation, combined with the previous experiments, provides a comprehensive validation of our method’s practicality and reliability for real-world applications.

4. Discussion

The superiority of STGAN, as evidenced by the experimental results, stems from its synergistic architecture and loss design that address key limitations of prior arts. This advantage manifests in several key aspects: First, the multi-scale decoder coupled with the U-Net skip connections enhances pixel-level accuracy and structural consistency, leading to improvements in PSNR and SSIM over single-scale models like FusionGAN. Second, the global modeling capacity of the Swin-Transformer backbone, augmented by the dual-window attention mechanism, captures richer contextual details, which is reflected in the 15–20% higher SF and EN compared to CNN-based methods like Dense Fuse. Finally, this effective integration of cross-modal information directly promotes the optimization of feature-based metrics such as FMI and MI.

The superiority of STGAN, as evidenced by the experimental results, stems from its synergistic architecture and loss design that address key limitations of prior arts. This advantage is fundamentally rooted in the Swin Transformer’s architectural superiority over traditional CNNs. Specifically, the global receptive field and long-range dependency modeling capability of the self-attention mechanism allow STGAN to integrate complementary infrared and visible information that is often spatially disparate, leading to more globally consistent fusion and superior performance on metrics like MI and VIF. Furthermore, the hierarchical feature extraction with shifted windows provides an innate multi-scale understanding, enabling the model to simultaneously capture coarse thermal targets and fine-grained visible textures, which is directly reflected in the 15–20% higher SF and EN

compared to CNN-based methods like Dense Fuse. Finally, our introduced dual-window multi-head self-attention (DWMSA) module acts as a pivotal enhancement. By fusing the standard window attention with a convolutional feedforward network, it explicitly balances the capture of global contextual relationships and local spatial details, thereby achieving sharper edges and richer textures than both pure CNN and standard Transformer models.

Critically, the robust performance of STGAN is not limited to a single dataset. As demonstrated in the generalization analysis (Section 3.6), our model, trained solely on MSRS, maintains its leading performance on the disparate TNO dataset. The consistent superiority on TNO, which contains challenging military and surveillance scenarios, provides definitive evidence that the advantages described above—the global modeling capacity, multi-scale fusion capability, and the balanced design of the DWMSA module—constitute a generalizable fusion principle. This cross-domain consistency underscores that STGAN has learned a fundamental and robust image fusion strategy, rather than merely overfitting to the characteristics of the training data.

The experimental results on general datasets confirm that STGAN achieves a superior balance between subjective visual quality and objective metrics. This can be attributed to the collaborative design of the Swin-Transformer backbone and the tailored loss functions, which enables the model to excel in both detail preservation and thermal target enhancement. This robust performance, particularly its proven generalization capability, strongly suggests the considerable application potential of our method in practical scenarios such as visual inspection for industrial surface defects. For example, in the UAV inspection of photovoltaic power plants, our method can fuse visible light and infrared video streams to generate clearer fusion images so as to help the operation and maintenance personnel or subsequent AI models more accurately locate defects such as hot spots and dirt.

The limitation of this study is that the end-to-end performance verification has not been carried out on a special defect detection dataset. This is mainly due to the scarcity of open, large-scale registration of multimodal datasets of industrial defects. As future work, we plan to do the following:

- Cooperate with industrial partners to build such datasets;
- Integrate the fusion algorithm into the complete defect detection pipeline as a pre-processing module to directly evaluate its effect on improving the final detection accuracy.

5. Conclusions

We present a novel approach for superior infrared–visible image fusion through a Swin Transformer-based countermeasure network. With GAN as the core framework, the local modeling of traditional CNN is organically combined with the global perception ability of the Swin Transformer. The local window mechanism has higher computational efficiency. Using the double branch attention mechanism, U-Net jump connection, and multi-scale feature reconstruction, the encoder captures the multi-scale semantic features, and the decoder reconstructs the spatial structure while maintaining the resolution. The dynamic weight combination of edge preservation loss and multi-scale structure similarity loss is introduced. The model effectively preserves both the infrared highlight region and visible image detail structure. The proposed method is superior to ADF, FusionGAN, Dense Fuse, PIA Fusion, and other existing methods in many subjective and objective evaluation indexes.

Author Contributions: Conceptualization: L.G. and Y.H.; Methodology: L.G., Y.H. and R.L.; Validation: L.G. and Y.H.; initial draft composed by L.G.; manuscript reviewed and edited by Y.H. and R.L. All authors approved the final publication. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The MSRs, RoadScene, and TNO datasets used in this study are publicly available. The MSRS dataset is available at [<https://doi.org/10.1109/TPAMI.2025.3609323>]. The RoadScene dataset is available at [<https://doi.org/10.1609/aaai.v34i07.6936>]. The TNO dataset is available at [<https://doi.org/10.1016/j.dib.2017.09.015>].

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

Abbreviations

The following abbreviations are used in this manuscript:

DWMSA	Dual Window-based Multi-head Self-Attention
EN	Entropy
FNN	Feedforward Neural Network
GAN	Generative Adversarial Network
LN	Layer Normalization
MPL	Multi-Layer Perceptron
MSA	Multi-head Self-Attention
PSNR	Peak Signal-to-Noise Ratio
ReLU	Rectified Linear Unit
SF	Spatial Frequency
SSIM	Structural Similarity Index Measurement
VIF	Visual Information Fidelity
W-MSA	Window-based Multi-head Self-Attention

References

1. Song, H.; Wang, R. Underwater Image Enhancement Based on Multi-Scale Fusion and Global Stretching of Dual-Model. *Mathematics* **2021**, *9*, 595. [[CrossRef](#)]
2. Rashid, M.; Khan, M.A.; Alhaisoni, M.; Wang, S.-H.; Naqvi, S.R.; Rehman, A.; Saba, T. A Sustainable Deep Learning Framework for Object Recognition Using Multi-Layers Deep Features Fusion and Selection. *Sustainability* **2020**, *12*, 5037. [[CrossRef](#)]
3. Liu, Z.; Wu, W. Fusion with Infrared Images for an Improved Performance and Perception. In *Pattern Recognition, Machine Intelligence and Biometrics*; Wang, P.S.P., Ed.; Springer: Berlin/Heidelberg, Germany, 2011. [[CrossRef](#)]
4. Tang, C.; Ling, Y.; Yang, H.; Yang, X.; Lu, Y. Decision-level fusion detection for infrared and visible spectra based on deep learning. *Infrared Laser Eng.* **2019**, *48*, 626001. [[CrossRef](#)]
5. Zhao, Y.; Lai, H.; Gao, G. RMFNet: Redetection Multimodal Fusion Network for RGBT Tracking. *Appl. Sci.* **2023**, *13*, 5793. [[CrossRef](#)]
6. Kim, S.; Song, W.-J.; Kim, S.-H. Double Weight-Based SAR and Infrared Sensor Fusion for Automatic Ground Target Recognition with Deep Learning. *Remote Sens.* **2018**, *10*, 72. [[CrossRef](#)]
7. Lewis, J.J.; O'Callaghan, R.J.; Nikolov, S.G.; Bull, D.R.; Canagarajah, N. Pixel- and region-based image fusion with complex wavelets. *Inf. Fusion* **2007**, *8*, 119–130. [[CrossRef](#)]
8. Liu, C.; Du, L.; Liu, R. Infrared and visible image fusion algorithm based on multi-scale transform. In Proceedings of the 2021 5th International Conference on Electronic Information Technology and Computer Engineering (EITCE '21), Xiamen, China, 22–24 October 2021; Association for Computing Machinery: New York, NY, USA, 2022; pp. 432–438. [[CrossRef](#)]
9. Li, L.; Shi, Y.; Lv, M.; Jia, Z.; Liu, M.; Zhao, X.; Zhang, X.; Ma, H. Infrared and Visible Image Fusion via Sparse Representation and Guided Filtering in Laplacian Pyramid Domain. *Remote Sens.* **2024**, *16*, 3804. [[CrossRef](#)]
10. AlRegib, G.; Prabhushankar, M. Explanatory Paradigms in Neural Networks: Towards relevant and contextual explanations. *IEEE Signal Process. Mag.* **2022**, *39*, 59–72. [[CrossRef](#)]
11. Ramírez, J.; Vargas, H.; Martínez, J.I.; Arguello, H. Subspace-Based Feature Fusion from Hyperspectral and Multispectral Images for Land Cover Classification. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 11–16 July 2021; pp. 3003–3006. [[CrossRef](#)]
12. Sun, S.; Bao, W.; Qu, K.; Feng, W.; Ma, X.; Zhang, X. Hyperspectral-multispectral image fusion using subspace decomposition and Elastic Net Regularization. *Int. J. Remote Sens.* **2024**, *45*, 3962–3991. [[CrossRef](#)]

13. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 9992–10002. [\[CrossRef\]](#)
14. Sulthana, N.T.N.; Joseph, S. Infrared and visible image: Enhancement and fusion using adversarial network. *AIP Conf. Proc.* **2024**, *3037*, 020030. [\[CrossRef\]](#)
15. Yue, J.; Fang, L.; Xia, S.; Deng, Y.; Ma, J. Dif-Fusion: Toward High Color Fidelity in Infrared and Visible Image Fusion With Diffusion Models. *IEEE Trans. Image Process.* **2023**, *32*, 5705–5720. [\[CrossRef\]](#)
16. Ma, J.; Ma, Y.; Li, C. Infrared and visible image fusion methods and applications: A survey. *Inf. Fusion* **2019**, *45*, 153–178. [\[CrossRef\]](#)
17. Yin, L.; Zhao, W. Graph attention-based U-net conditional generative adversarial networks for the identification of synchronous generation unit parameters. *Eng. Appl. Artif. Intell.* **2023**, *126*, 106896. [\[CrossRef\]](#)
18. Li, K.; Wang, Y.; Zhang, J.; Gao, P.; Song, G.; Liu, Y.; Li, H.; Qiao, Y. Unifying Convolution and Self-Attention for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 12581–12600. [\[CrossRef\]](#)
19. Isola, P.; Zhu, J.-Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5967–5976. [\[CrossRef\]](#)
20. Xiao, Z.; Zhang, D.; Chen, X.; Li, D. SCAGAN: Wireless Capsule Endoscopy Lesion Image Generation Model Based on GAN. *Electronics* **2025**, *14*, 428. [\[CrossRef\]](#)
21. Lv, J.; Wang, C.; Yang, G. PIC-GAN: A parallel imaging coupled generative adversarial network for accelerated multi-channel MRI reconstruction. *Diagnostics* **2021**, *11*, 61. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Ma, J.; Yu, W.; Liang, P.; Li, C.; Jiang, J. FusionGAN: A generative adversarial network for infrared and visible image fusion. *Inf. Fusion* **2019**, *48*, 11–26. [\[CrossRef\]](#)
23. Xu, H.; Ma, J.; Jiang, J.; Guo, X.; Ling, H. U2fusion: A Unified Unsupervised Image Fusion Network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 502–518. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Mei, Z. Minimizing the Average Packet Access Time of the Application Layer for Buffered Instantly Decodable Network Coding. *IEEE Trans. Parallel Distrib. Syst.* **2023**, *34*, 1035–1046. [\[CrossRef\]](#)
25. Tang, L.F.; Li, C.Y.; Ma, J.Y. Mask-DiFuser: A masked diffusion model for unified unsupervised image fusion. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, early access. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Tang, L.F.; Yan, Q.L.; Xiang, X.Y.; Fang, L.Y.; Ma, J.Y. C2RF: Bridging multi-modal image registration and fusion via commonality mining and contrastive learning. *Int. J. Comput. Vis.* **2025**, *133*, 5262–5280. [\[CrossRef\]](#)
27. 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. [\[CrossRef\]](#)
28. Xu, H.; Ma, J.; Le, Z.; Jiang, J.; Guo, X. FusionDN: A Unified Densely Connected Network for Image Fusion. In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI), Philadelphia, PA, USA, 25 February–4 March 2025; pp. 12484–12491. [\[CrossRef\]](#)
29. Toet, A. The TNO Multiband Image Data Collection. *Data Brief* **2017**, *15*, 249–251. [\[CrossRef\]](#) [\[PubMed\]](#)

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.