



Article

Infrared and Visible Image Fusion with Significant Target Enhancement

Xing Huo ¹, Yinping Deng ¹  and Kun Shao ^{2,*} ¹ School of Mathematics, Hefei University of Technology, Hefei 230009, China² School of Software, Hefei University of Technology, Hefei 230009, China

* Correspondence: shaokun@hfut.edu.cn

Abstract: Existing fusion rules focus on retaining detailed information in the source image, but as the thermal radiation information in infrared images is mainly characterized by pixel intensity, these fusion rules are likely to result in reduced saliency of the target in the fused image. To address this problem, we propose an infrared and visible image fusion model based on significant target enhancement, aiming to inject thermal targets from infrared images into visible images to enhance target saliency while retaining important details in visible images. First, the source image is decomposed with multi-level Gaussian curvature filtering to obtain background information with high spatial resolution. Second, the large-scale layers are fused using ResNet50 and maximizing weights based on the average operator to improve detail retention. Finally, the base layers are fused by incorporating a new salient target detection method. The subjective and objective experimental results on TNO and MSRS datasets demonstrate that our method achieves better results compared to other traditional and deep learning-based methods.

Keywords: image fusion; infrared image; visible image; significant target enhancement; multi-level Gaussian curvature filtering; ResNet50



Citation: Huo, X.; Deng, Y.; Shao, K. Infrared and Visible Image Fusion with Significant Target Enhancement. *Entropy* **2022**, *24*, 1633. <https://doi.org/10.3390/e24111633>

Academic Editors: Jiayi Ma, Yu Liu, Junjun Jiang, Zheng Wang and Han Xu

Received: 14 October 2022

Accepted: 7 November 2022

Published: 10 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 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

Image fusion is an image enhancement technique that aims to integrate complementary and redundant information from multiple sensors, which can generate more reliable, robust, and informative images. Infrared and visible image fusion is an important field in image fusion [1]. Infrared images capture the thermal radiation emitted in the scene, which can effectively distinguish the target and background in the scene, and is not affected by occlusions and extreme weather. However, infrared images have shortcomings such as not obvious details [2]. According to this characteristic of infrared images, many different areas have been researched to identify prominent targets in infrared scenes, such as infrared image segmentation [3], pedestrian detection [4], etc. Visible images focus on capturing reflected light, which has quite high spatial resolution and contrast. Although visible images can effectively preserve the details of the scene, they are vulnerable to factors such as light intensity and extreme weather. The infrared and visible image can not only highlight the thermal radiation information in the infrared image, but also get the rich details in the visible image. Therefore, it is widely used in many fields such as military detection and public security [5].

In general, existing image fusion frameworks mainly include multi-scale transform [6], sparse representation [7], deep learning [8], etc. Currently, the most widely employed method for infrared and visible image fusion is the multi-scale transform, which can decompose the source image into multiple sub-images at different scales and effectively retain the detail features at different scales, so that the fused image can obtain a good visual effect. Traditional image fusion based on multi-scale transform includes pyramid transform [9], discrete wavelet transform [10], curvelet transform [11], and so on. These

multi-scale transform methods can fuse the source images at different scales, but there are still some defects. For example, the pyramid transform does not have translation invariance, which may lead to a false Gibbs phenomenon in the fused image. Wavelet and curvelet transform have been improved, but they do not fully take into account the spatial consistency, which may lead to the color brightness distortion of the fusion results [12]. To improve these problems, the edge preserving filters are proposed, such as guide filter [13] and bilateral filter [14], which are very effective in solving the spatial consistency problem and can reduce artifacts around edges [15]. In recent years, some new multi-scale methods have achieved good results. More specifically, Jia et al. [16] proposed a stretched multi-scale fusion method for infrared and visible images, which can improve the information content of fusion results. Li et al. [17] proposed a multi-exposure fusion method for alleviating the issue of lost details in fusion results that may be caused by traditional multi-scale. Meanwhile, hybrid filters have also been widely used in the field of image fusion. Do et al. [18] proposed a method combining Gaussian filtering (GF) and rolling guidance filter (RGF), which solved the problem that conventional multi-scale methods for image decomposition could only be performed from a single channel. Zhou et al. [9] proposed a hybrid multi-scale decomposition method combining GF and bilateral filtering, which improved the problem of unstable bilateral filtering weights. The Gaussian curvature filtering (GCF), an edge-preserving filter proposed in recent years, is superior to RGF and other edge-preserving filters because of its excellent parameter-free characteristics and high efficiency of fine-scale retention. Although the above methods have achieved certain effects, there are still some problems, such as not considering the spatial scale may lead to the loss of some details, and cannot be well extracted salient targets.

In recent years, deep learning methods have developed rapidly in the field of image fusion. Liu et al. [19] used CNN for multi-focus image fusion and achieved good fusion performance in both subjective and objective evaluation. However, this method only deployed the results of the last layer as image features, which may lose a lot of useful information obtained from the middle layer. Li et al. [20] applied VGG to an infrared and visible image fusion task, using VGG19 to reconstruct the detailed content and retain as much detailed information as possible. However, the deep learning fusion method based on the above does not make full use of the deep features, and with the deepening of these network layers, the performance will tend to saturate or even decline. To address these issues, Chen et al. [21] proposed an attention-guided progressive neural texture fusion model to suppress noise in the fusion results. Bai et al. [22] developed an end-to-end deep pre-dehazer model. Moreover, Li et al. [23] devised a meta learning-based deep framework for fusing infrared and visible images with different spatial resolutions. Ma et al. proposed DDcGAN [24], an end-to-end model for fusing infrared and visible images of different resolutions. Li et al. proposed DenseFuse [25] and RFN-Nest [26], infrared and visible image fusion frameworks based on self-Auto Encoder. Xu et al. proposed an unsupervised end-to-end image fusion network, U2Fusion [27], which achieved good results. However, these methods still have some defects in target saliency extraction, such as they cannot well highlight infrared salient targets.

To solve the above problems, a new infrared and visible image fusion framework is proposed, which preserves the details as much as possible on the basis of considering the spatial scale, reduces the generation of artifacts, and improves the saliency of the target. Since multi-level Gaussian filtering (MLGCF) [28] can better obtain high spatial resolution background information, we use it to decompose the source image into large-scale layer, small-scale layer, and base layer. For large-scale layers, ResNet50 is used to extract features, then ZCA and L1 norms are used to construct the fusion weights. The max-absolute rule is adopted for the small-scale layer. A new frequency tuning saliency detection method (FT++) is proposed for base layer fusion. The experimental results on TNO and MSRS datasets show that our method outperforms many state-of-the-art methods.

The main contributions of this paper are as follows:

- (1) We propose a novel infrared and visible fusion method, which can effectively retain detailed information and maintain the target saliency. This method can be widely used in the military, target detection, and other fields.
- (2) More abundant details can be obtained from the source image by employing MLGCF and ResNet50 to extract features.
- (3) A new approach to constructing saliency map (FT++) is proposed, which can productively retain the thermal radiation information. Extensive qualitative and quantitative experiments demonstrate the superiority of our method compared to the latest alternatives. Compared with other competitors, our approach could generate fused images looking like high-quality visible images with highlighted targets.

The remainder of this paper is organized as follows. Section 2 presents the theory related to residual networks and FT saliency detection. Section 3 describes the proposed strategy for infrared and visible image fusion in detail. In Section 4, our approach is compared with some existing state-of-the-art methods on two different datasets, and ablation experiments as well as fusion performance analysis are performed to verify the effectiveness of the proposed method. Finally, conclusions are presented in Section 5.

2. Correlation Theory

2.1. Residual Network

He et al. [29] proposed a residual network and introduced jumping connection lines on each residual block structure, which solved the problems of gradient disappearance caused by the increase of network layers and rapid decline after accuracy saturation. Compared with VGG19, the residual network has a superior ability to extract features. Considering the complexity of the task and the size of the data, we leverage ResNet50, a network of 50 weight layers, for the extraction of detailed features. The structure of the residual blocks in ResNet50 is shown in Figure 1:

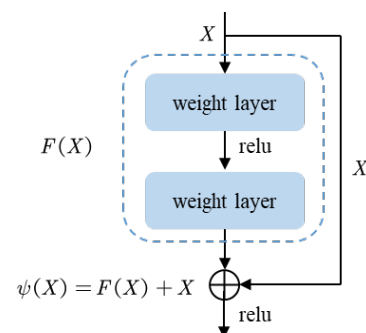


Figure 1. Residual block structure.

Where X represents the input of residual block structure, $F(X)$ represents the result of input X calculated by two weight layers, and $\psi(X) = F(X) + X$ is the feature learned by the residual block structure. The core idea of the network is to introduce a jump connection line to ensure that the output $\psi(X)$ can at least learn new features. When the gradient disappears, $\psi(X) = X$ is an identity map, and many such structures are stacked together, ensuring that the result of this network is at least as good as that of the shallow network.

2.2. FT Significance Detection

Achanta, R. et al. [30] proposed a saliency detection algorithm based on FT. The main principle of the algorithm is to discard the high-frequency information in the frequency domain and retain the low-frequency information such as the contour and basic composition of the image as much as possible. The mathematical expression of pixel saliency is as follows:

$$S(p) = \|I_{\mu} - I_{whc}(p)\|, \quad (1)$$

where I_μ is the average pixel value of input image I . $I_{whc}(p)$ is the pixel value of input image I at point p , that processed by GF with a window size of 5×5 , and $\|\cdot\|$ represents the L2-norm.

FT algorithm carries out saliency detection from the perspective of spatial frequency, which has simple and efficient advantages. However, because the original FT algorithm employs GF to process the input image, the strong and weak edge information of the image is easy to be blurred by GF, and the key information of the image cannot be fully extracted. To solve these problems, in recent years, Hou et al. improved the FT algorithm by guide filter [31]. As an edge-preserving filter, guide filter uses the mean and variance of pixel neighborhood as local estimation, which has a strong edge-preserving ability. However, considering that RGF as a hybrid filter, it combines the smoothing characteristics of GF and the edge retention characteristics of guide filter. Therefore, it can extract information from different scales and has a stronger edge preserving ability compared with GF and guide filter [15]. Based on the above analysis, we adopt RGF to replace GF to improve original FT (FT++).

The FT++ method is shown in Figure 2. First, the source image is processed with RGF and converted from RGB to LAB color space to obtain I_{rgf} , which can be represented as $I_{rgf}(x, y) = [l_{rgf}, a_{rgf}, b_{rgf}]^T$, where $l_{rgf}, a_{rgf}, b_{rgf}$ represent the three channels in LAB color space, respectively. Second, $I_\mu = [l_\mu, a_\mu, b_\mu]^T$ is obtained by calculating the average value for each channel. Finally, the pixel significant values are acquired according to Equation (7).

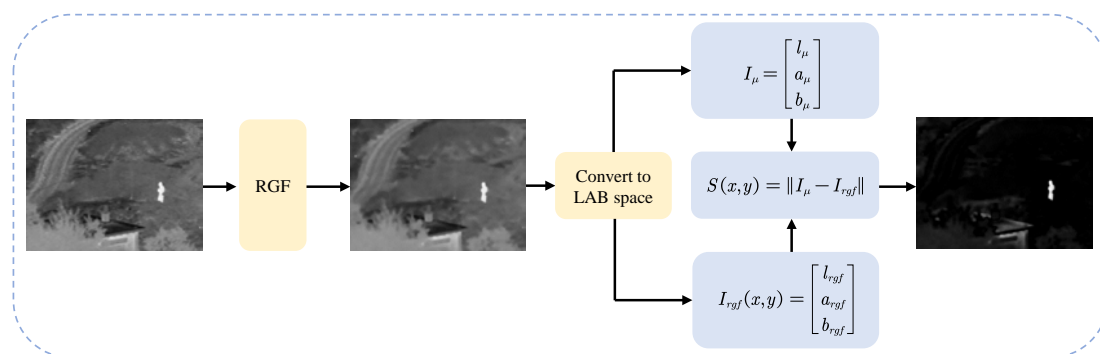


Figure 2. Flowchart for extracting salient targets with FT++.

We compare FT++ with the original FT method to verify the effectiveness of the method, see Figure 3. It can be seen from Figure 3d that there are many edge features such as people and houses in the difference map [32]. Therefore, FT++ can obtain more salient information than the original FT method.

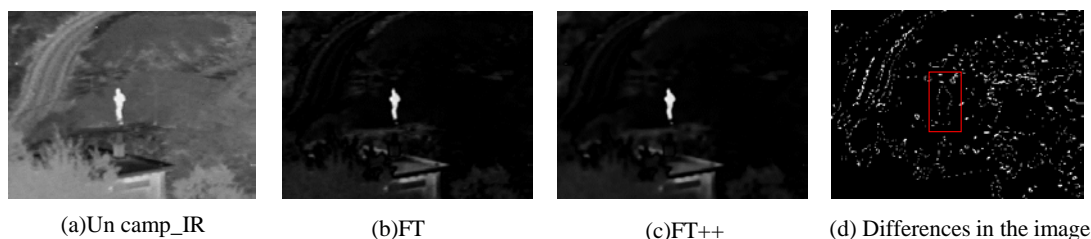


Figure 3. Comparison of different feature extraction methods for infrared images.

3. Proposed Fusion Framework

The infrared and visible image fusion based on multi-scale decomposition and FT saliency can effectively retain the texture details of the source image and enhance significant targets. After decomposing the source image using MLGCF, the base layer saliency map is extracted by FT++, the large-scale features are processed by introducing ResNet50 etc., and

finally the final fused image is reconstructed. The flow chart of the method in this paper is shown in Figure 4.

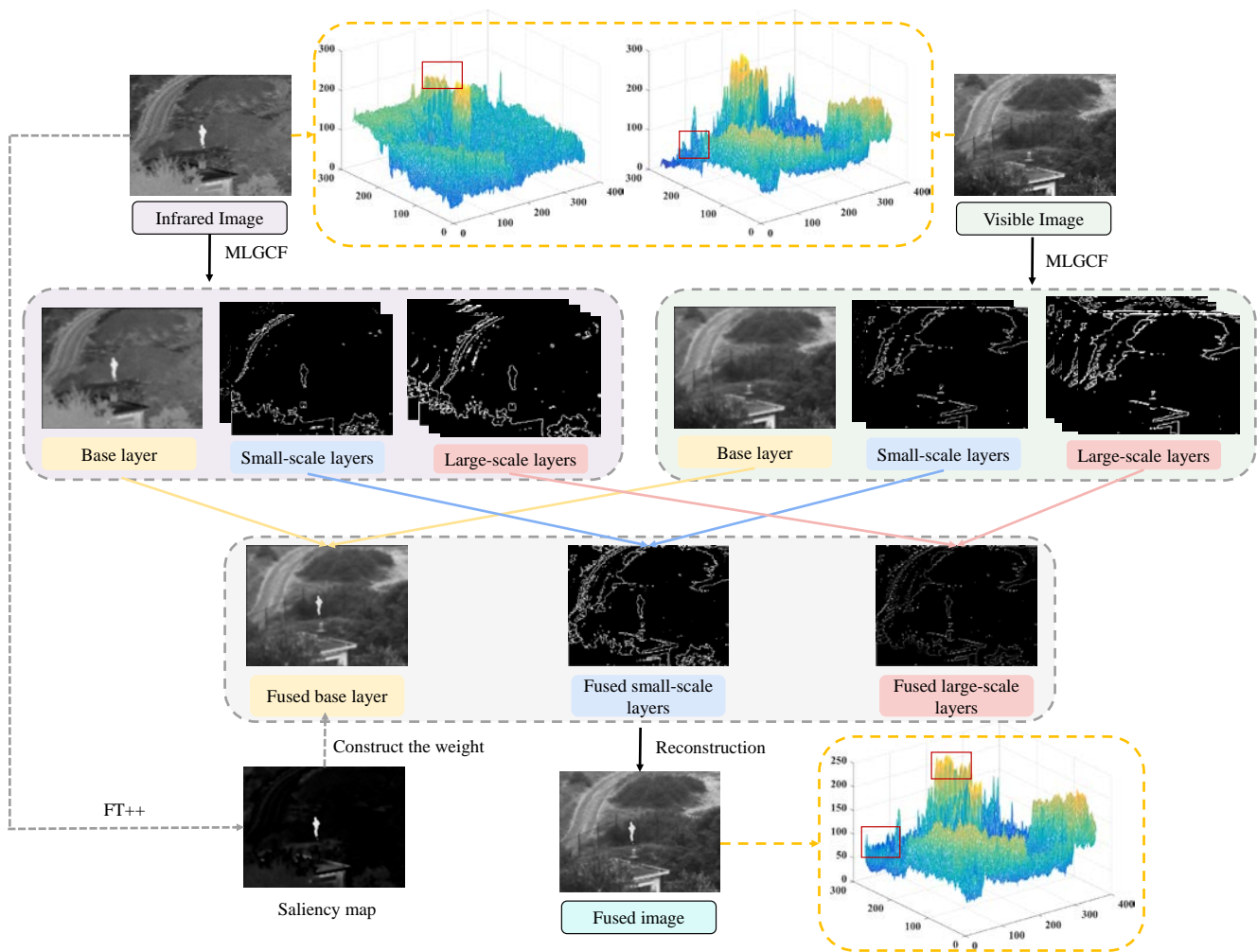


Figure 4. The image fusion framework of our method.

3.1. Image Decomposition

The edge-preserving filter can improve the common problems in the multi-scale decomposition process, such as the tendency to produce halo artifacts [33], and does not take full account of space consistency, which can lead to distorted color brightness of the fusion results. To solve these problems, we adopt MLGCF to decompose the source image. GF is a classical tool for image degradation and smoothing. GCF is an effective edge-preserving filter with the advantage of being parameter-free and retaining fine detail. It is instructive to note that MLGCF takes use of the smoothing properties of the GF and the edge-preserving properties of the GCF to obtain features at different scales. The specific process is divided into the following three steps:

- (1) Using GF to smooth small structure information:

$$I_{k,g} = \text{Gaussian}(I_k, \sigma_s), \tag{2}$$

where $I_k (k \in 1, 2)$ is the input image. I_1 and I_2 are infrared image and visible image respectively. $I_{k,g}$ is the result of the input image processed by GF. σ_s is the standard deviation of GF, which is mainly used to smooth the texture details of the image.

- (2) Using GCF for the edge recovery process:

$$I_{k,gcf} = \text{GCF}(I_k, m), \tag{3}$$

- the parameter m is the number of iterations, and we set $m = 5$ based on experience.
- (3) Combining GF with GCF using a hybrid multiscale approach for a three-stage decomposition:

$$D_k^{i,1} = \begin{cases} I_k - I_{k,gcf}^1, & i = 1 \\ I_{k,g}^{i-1} - I_{k,gcf}^i, & i = 2, 3 \end{cases} \quad (4)$$

$$D_k^{i,2} = I_{k,gcf}^i - I_{k,g}^i, \quad i = 1, 2, 3, \quad (5)$$

$$B_k = I_{k,g}^3, \quad i = 3, \quad (6)$$

where $D_k^{i,j}$ ($j = 1, 2$) represents the texture detail and edge detail on the multi-scale decomposition of layer i , respectively, $i \in \{1, 2, 3\}$ denotes the number of decomposition layers. Record the result of the last GF decomposition as base layer $B_k \cdot I_{k,g}^i$ ($i = 1, 2, 3$) represents the result of the i -th GF process of the input image $I_k \cdot I_{k,gcf}^i$ ($i = 1, 2, 3$) denotes the result of I_k after the i -th GCF process. The parameter σ_s is the variance in the GF operation and is taken as $\sigma_s = 20$ in this paper.

To verify the advantages of the MLGCF decomposition, the MLGCF is compared with the RGF decomposition. As can be seen in Figure 5, the RGF decomposition has a halo around the target as the number of layers deepens, but this is barely visible with the MLGCF decomposition. Furthermore, the MLGCF decomposition results show that each scale layer contains the specific content of the current detail layer. Therefore, the MLGCF decomposition has the effect of suppressing halos and preserving the content of a specific scale of detail.

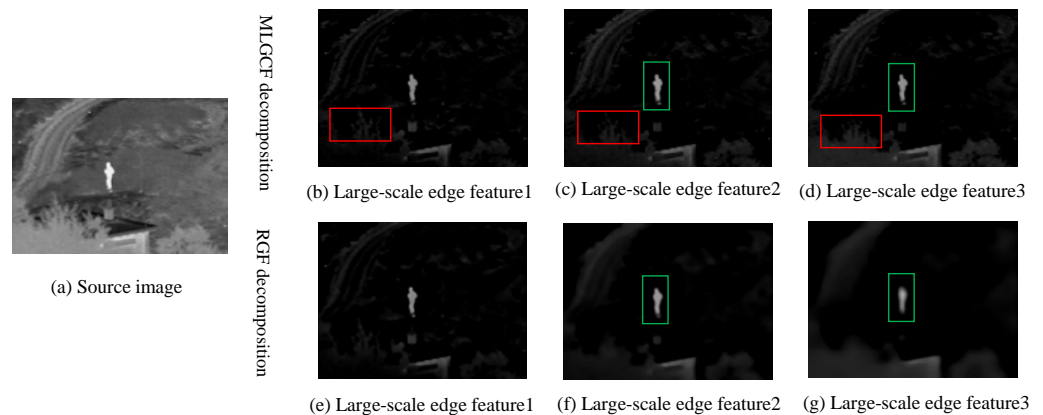


Figure 5. Detail comparison of MLGCF and RGF.

3.2. Image Fusion

3.2.1. Fusion Strategy for the Base Layer

In the past, most of the base layers were fused using simple averaging or weighted averaging, although these methods are simple to operate, they tend to lead to problems such as poor target saliency and low overall contrast of fusion results. To solve these problems, we adopt the FT++ method to process the infrared base layer and deploy its normalized result as the fusion weight. The specific steps are as follows:

- (1) FT++ method: The FT++ method in this paper only processes infrared images, so the input image for this process is the infrared image I_1 . An improvement is made using the RGF instead of the GF in the original FT algorithm, as shown in Figure 2.

Calculating saliency map:

$$S(p) = \left\| I_\mu - I_{rgf}(p) \right\|, \quad (7)$$

I_μ is the average pixel value of infrared image I_1 . $I_{rgf}(p)$ is the pixel value of I_1 at point p after RGF processing.

- (2) Normalizing the significance map to obtain the base layer fusion weights W_b :

$$W_b = \frac{S(p)}{\max(S(p))}. \quad (8)$$

- (3) Fusion of base layers using a weighted average strategy:

$$F_b = W_b \cdot B_1 + (1 - W_b) \cdot B_2, \quad (9)$$

where $B_k (k = 1, 2)$ is the base layer of the infrared and visible images respectively and F_b is the fusion result of the base layer.

3.2.2. Fusion Strategy for Small-Scale Layers

The texture information of the source image is contained in the small-scale layer. Usually, the larger the pixel value of the small-scale layer, the more texture information is retained [28], so for the fusion of the small-scale layer, we leverage the ‘‘Max-absolute’’ fusion method. The small-scale texture details and edge details are $D_k^{1,1}$, $D_k^{1,2}$, respectively.

Small-scale texture detail fusion results:

$$F_1^1(x, y) = \max(D_1^{1,1}(x, y), D_2^{1,1}(x, y)). \quad (10)$$

Small scale edge detail fusion results:

$$F_1^2(x, y) = \max(D_1^{1,2}(x, y), D_2^{1,2}(x, y)). \quad (11)$$

3.2.3. Fusion Strategy for Large-Scale Layers

The edge information and structural information of the source image are contained in the large-scale layer. Although deep learning fusion methods can effectively extract deep features, most of them only extract features without processing the extracted features, which may lead to the degradation of fusion results [34]. In order to make full use of and process the useful information in the deep network, and considering the complexity and data scale of the task, ResNet50 with ImageNet fixed training is used in this paper to extract large-scale layer features [29]. Then ZCA and L1-norm are employed to normalize the extracted features. Finally, the fusion weight is constructed to obtain the large-scale layer fusion results. The overall process is as follows:

- (1) Feature extraction: First, the large-scale layer $D_k^{i,j}$ ($i = 2, 3$) is input into ResNet50 to extract features. The texture features and edge features extracted to layer i ($i = 2, 3$) are denoted as $F_k^{i,j,t,c}$ ($j = 1, 2$), where t ($t = 1, 2, \dots, 5$) denotes the t -th convolutional block, and we take $t = 5$. c ($c = 1, 2, \dots, C$) denotes the c -th channel of the output feature, and C is the number of channels at level t , $C = 64 \times 2^{t-1}$.
- (2) The extracted features are ZCA processed to obtain the new features $\hat{F}_k^{i,j,t,c}$, then the L1-norm of $\hat{F}_k^{i,j,t,c}$ is calculated, and finally, we deploy the average operator to calculate the activity level measurement:

$$C_k^{i,j,t}(x, y) = \frac{\sum_{\beta=-r}^r \sum_{\theta=-r}^r \left\| \hat{F}_k^{i,j,t,1:C}(x + \beta, y + \theta) \right\|_1}{(2r + 1)^2}, \quad (12)$$

where the size of r determines the size of the extracted image block in the new feature $\left\| \hat{F}_k^{i,j,t,1:C} \right\|_1$. When r is too large, detail information may be lost [25], so we take $r = 1$.

- (3) Construction of initial weight maps using Softmax:

$$\hat{C}_k^{i,j,t}(x,y) = \frac{C_k^{i,j,t}(x,y)}{\sum_{k=1}^2 C_k^{i,j,t}(x,y)}. \quad (13)$$

- (4) Using a maximum weight construction method based on average operator (MWAO) method: In order to obtain as much detail information as possible, the largest pixel value in Equation (13) is taken on each large-scale layer as the fusion weight for that layer. Finally, the obtained weight is used to reconstruct the large-scale layer of fusion image:

$$W_k^{i,j,t} = \max(\hat{C}_k^{i,j,t}(x,y)). \quad (14)$$

$$F_i^j = W_1^{i,j,t} \cdot D_1^{i,j} + W_2^{i,j,t} \cdot D_2^{i,j} (i = 2, 3, j = 1, 2). \quad (15)$$

The MWAO method is compared with the method of constructing weight map [34] to verify its superiority. As shown in Table 1, it can be seen that the method of selecting the maximum weight to construct the fusion weight has more advantages than the original scheme of using the weight map in objective evaluation.

Table 1. Validation of large-scale construction method.

	SD	MI	AG	CE
Weight Map	56.8479	14.5510	3.6330	0.4770
MWAO	57.2314	14.5519	4.0634	0.4578

3.3. Reconstructing Fusion Image

Reconstruction of the fused image using the obtained fused base layer F_b and the detail layer $F_i^j (i = 1, 2, 3, j = 1, 2)$:

$$F = F_b + \sum_{j=1}^2 \sum_{i=1}^3 F_i^j. \quad (16)$$

4. Experimental Results and Comparisons

This section first introduces the datasets and evaluation metrics, as shown in Sections 4.1 and 4.2, respectively. Then we make a quantitative and qualitative comparison with the state-of-the-art methods, as shown in Sections 4.3 and 4.4 respectively. Finally, in Sections 4.5 and 4.6, the rationality and superiority of the method were proved by the ablation experiment and fusion performance analysis. We will introduce each part in detail in the following. The CPU used for the experiment is Intel Core i7-11800H, the graphics card is NVIDIA RTX 3060, the operating system is Windows 10, and the programming software is Matlab2016b.

4.1. Experimental Datasets

Subjective and objective evaluations of our method were carried out on two different datasets. The datasets are derived from TNO [35] and MSRS [36], and the selected images are aligned. Among them, the TNO dataset contains infrared and visible images of different military scenes, and MSRS dataset contains multiple infrared and visible images of multi-spectral road scenarios. In the subjective evaluation, five groups of representative infrared and visible images were selected for comparison on the two datasets, among which Un Camp, Kaptein_1123, Bench, Tree, and Pavilion were selected for TNO and 00352D, 00196D, 00417D, 00427D, 00545D were selected for MSRS, as shown in Figures 6 and 7. In the objective evaluation, 20 groups of registered infrared and visible images were selected from

each of the two datasets to calculate the corresponding evaluation metrics values. The detailed experimental results are shown in the Figures 8–11.

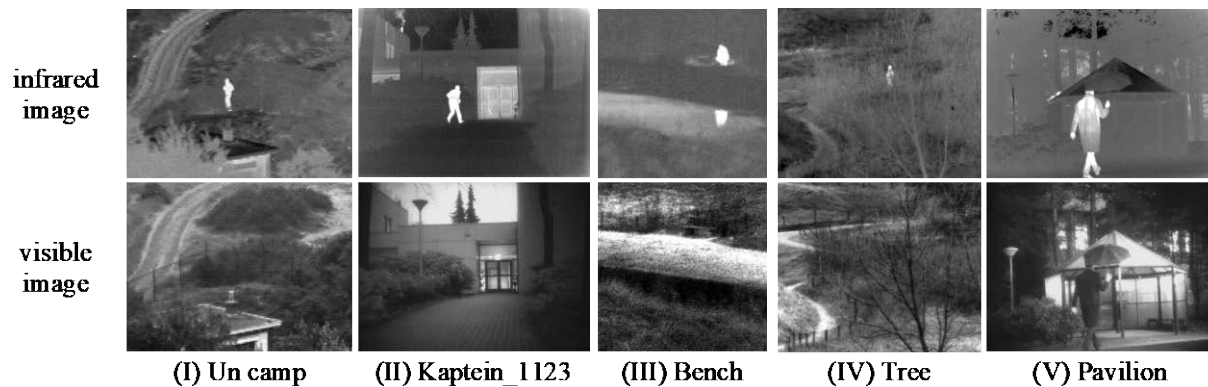


Figure 6. Five typical infrared and visible images on the TNO dataset.



Figure 7. Five typical infrared and visible images on the MSRS dataset.

4.2. Fusion Metrics

In order to reduce the interference of human consciousness in subjective evaluation, we chose six evaluation metrics, namely Entropy (EN) [37], Standard Deviation (SD) [38], Average Gradient (AG) [39], Visual Information Fidelity (VIF) [40], Mutual Information (MI) [41], and Cross Entropy (CE) [42], to validate the validity and superiority of our proposed method.

EN computes the amount of information contained in the fused image based on information theory. The higher the EN, the richer the information contained in the fused image:

$$EN = - \sum_{i=0}^{L-1} p_i \log_2 p_i, \quad (17)$$

where L is the number of grey levels, p_i is the normalized histogram of the corresponding gray level in the fused image.

SD is used to describe the statistical distribution and contrast features of images. The larger the SD, the higher the image contrast:

$$SD = \sqrt{\sum_{i=1}^M \sum_{j=1}^N (F(i,j) - \mu)^2}, \quad (18)$$

where F is the image fusion result, and μ is the average pixel value of the fusion result.

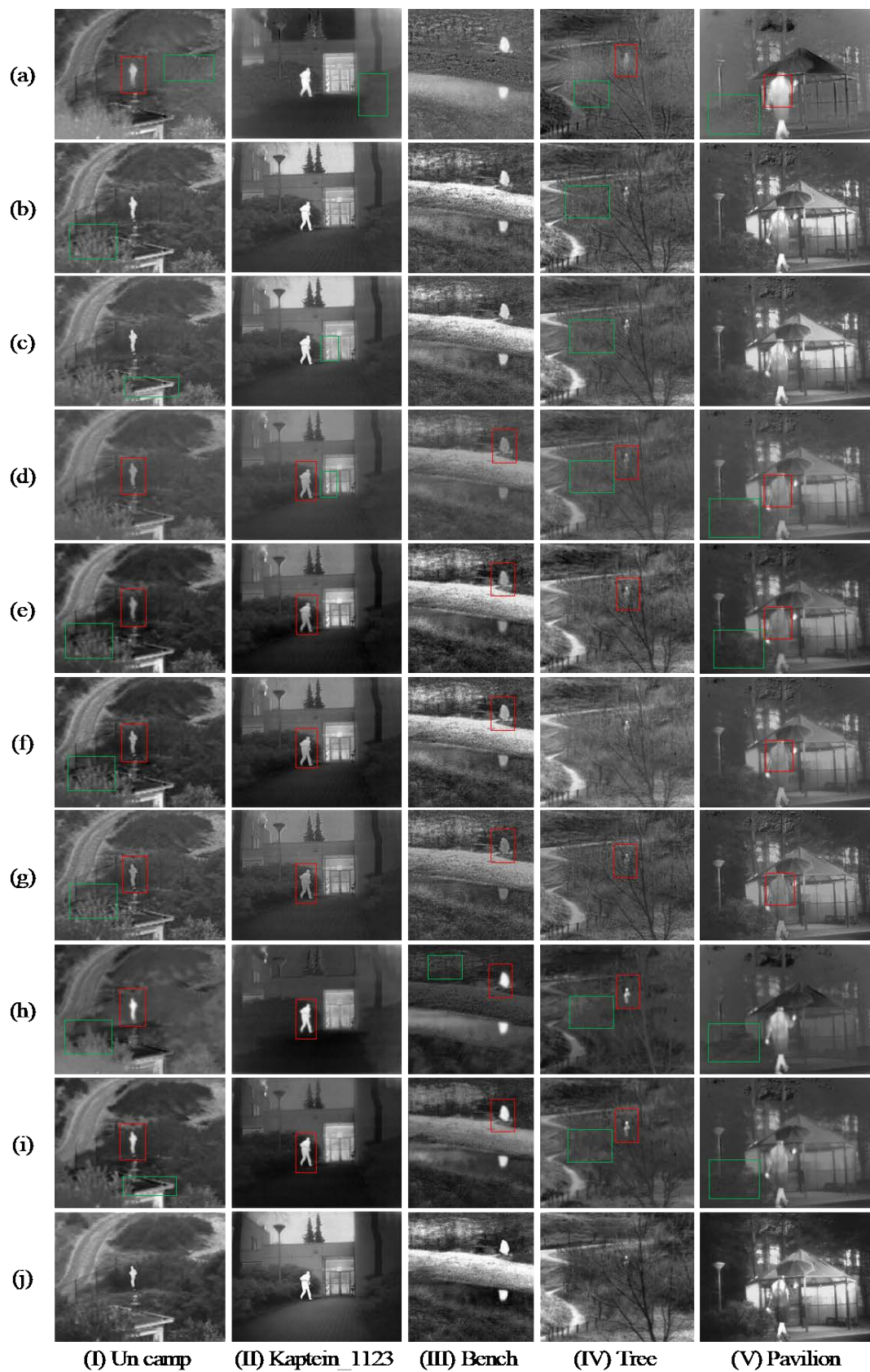


Figure 8. Typical fusion results of five infrared and visible images on the TNO dataset. (a) GTF, (b) WLS, (c) MLGCF, (d) ResNet50, (e) RFN-Nest, (f) DenseFuse, (g) U2Fusion, (h) FusionGAN, (i) GANMcC, (j) Ours.

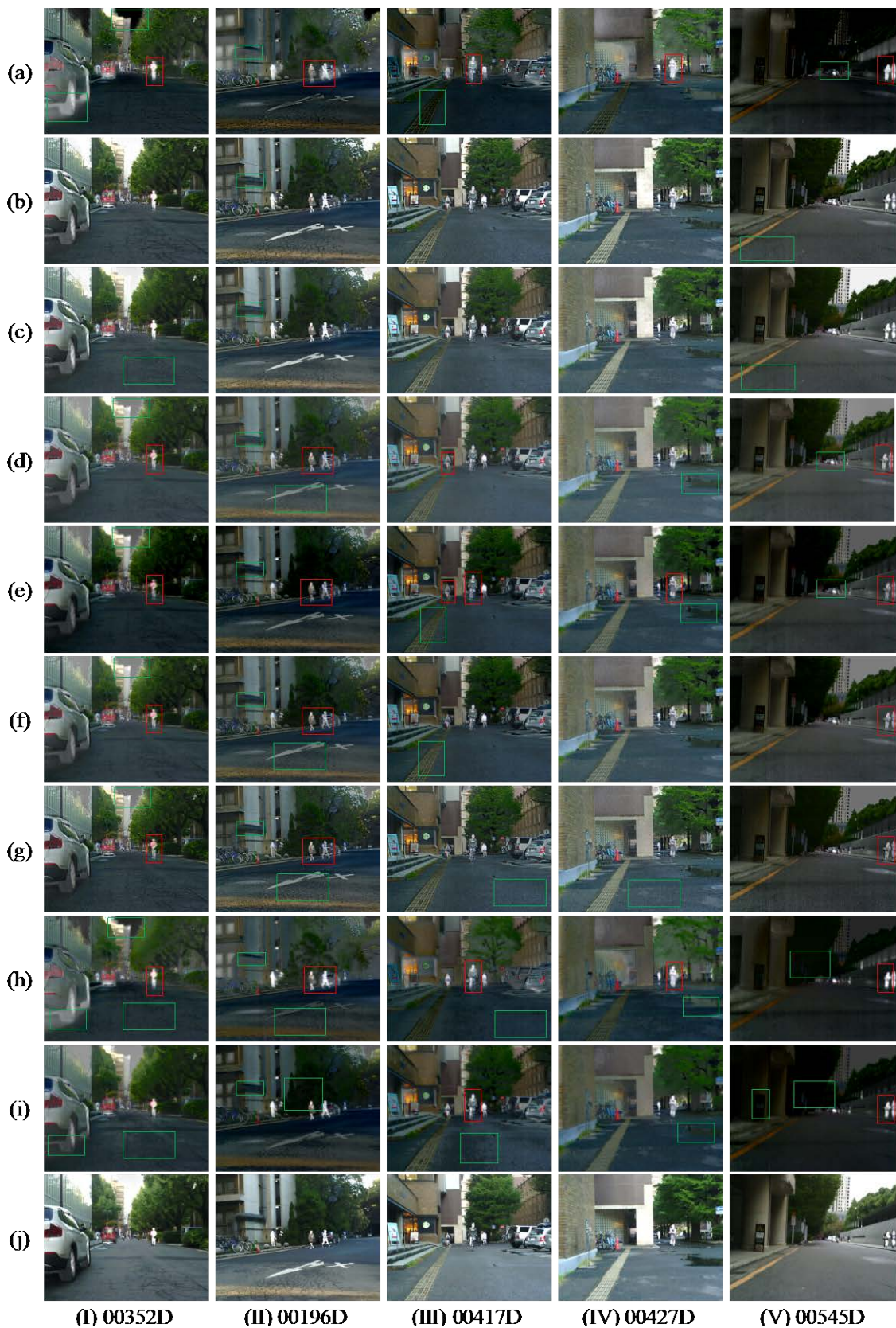


Figure 9. Typical fusion results of five infrared and visible images on the MSRS dataset. (a) GTF, (b) WLS, (c) MLGCF, (d) ResNet50, (e) RFN-Nest, (f) DenseFuse, (g) U2Fusion, (h) FusionGAN, (i) GANMcC, (j) Ours.

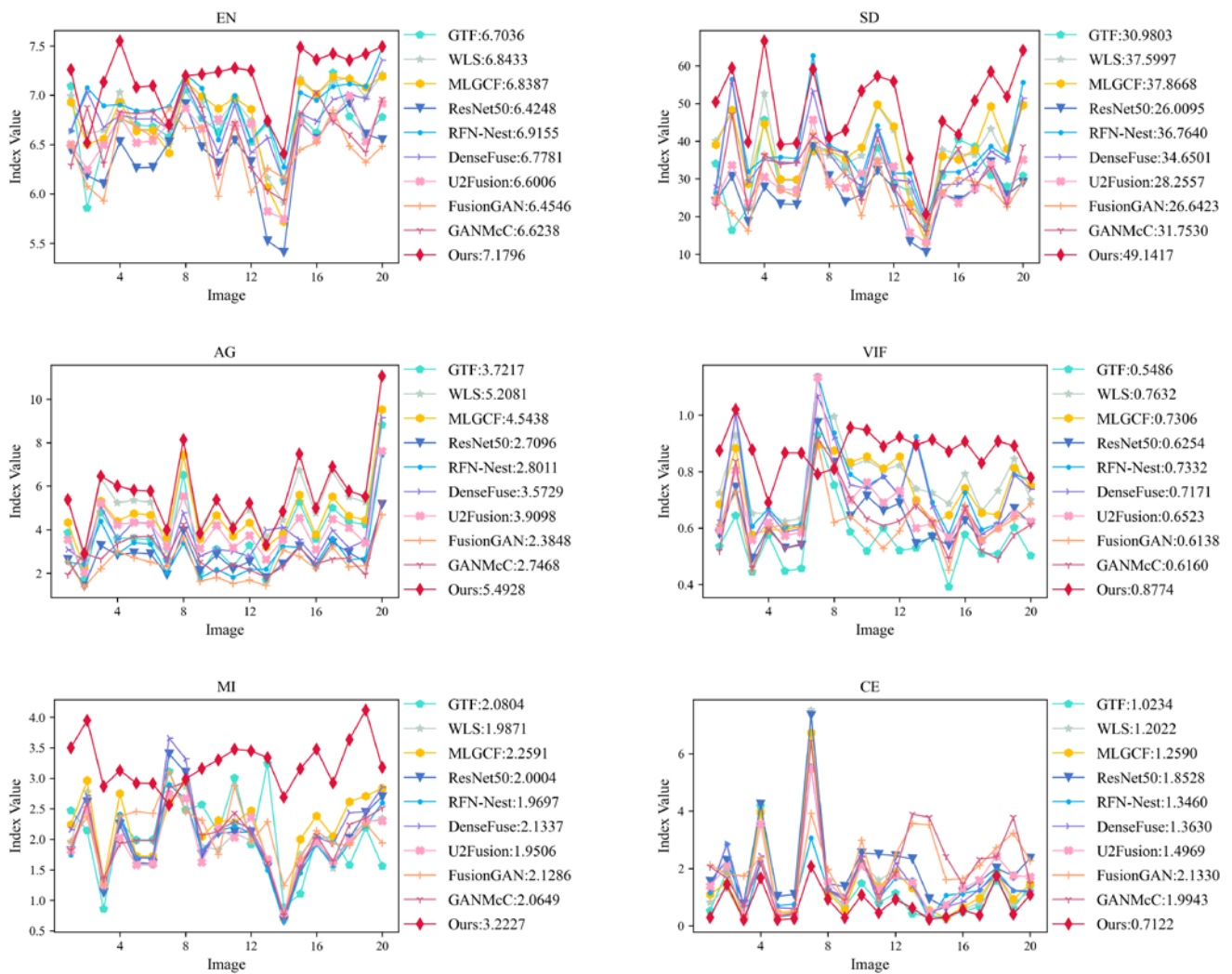


Figure 10. Twenty groups of infrared images and visible image different objective evaluation metric statistical broken line graph on the TNO dataset.

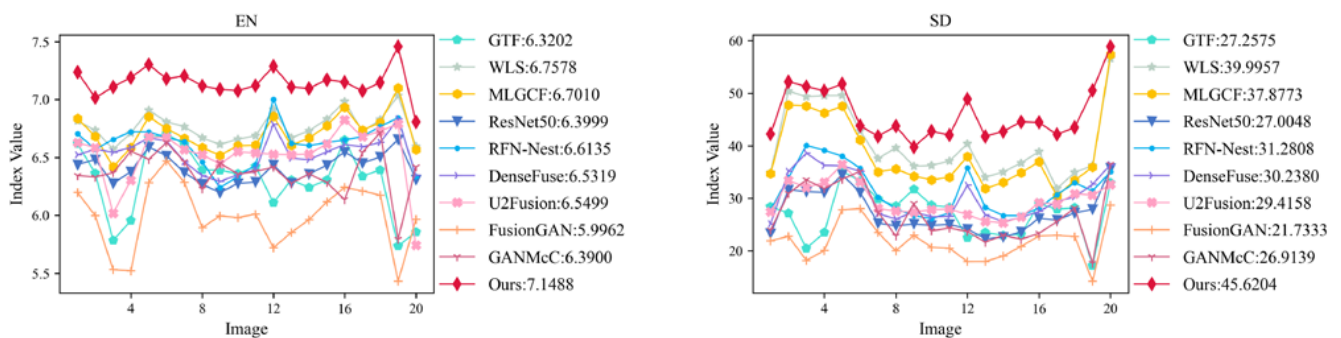


Figure 11. Cont.

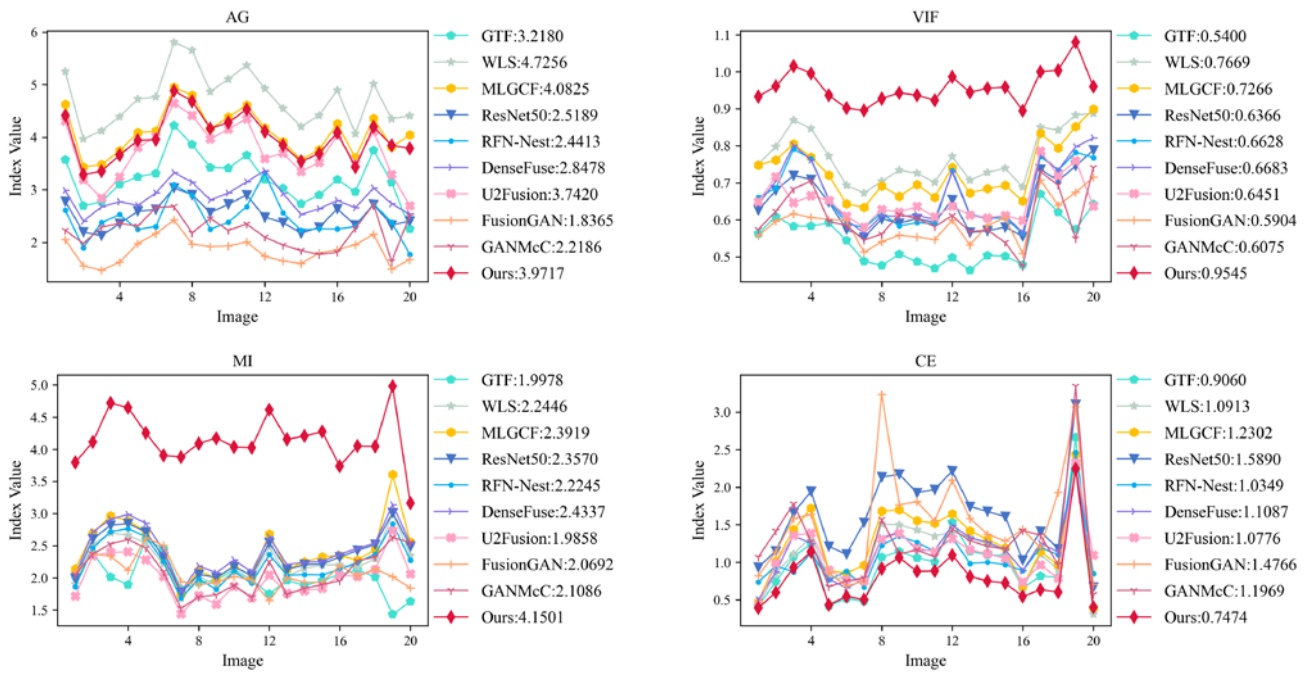


Figure 11. Twenty groups of infrared images and visible image different objective evaluation metric statistical broken line graph on the MSRS dataset.

AG is used to reflect the sharpness of the image. The larger the AG, the clearer the texture of the details contained in the image:

$$AG = \frac{1}{M \cdot N} \sum_{i=1}^M \sum_{j=1}^N \sqrt{\frac{\nabla F_x^2(i,j) + \nabla F_y^2(i,j)}{2}}, \tag{19}$$

$$\nabla F_x(i,j) = F(i,j) - F(i+1,j),$$

$$\nabla F_y(i,j) = F(i,j) - F(i,j+1).$$

VIF can objectively express people’s feelings when observing images. The larger the VIF, the better the visual effect of the images. The building process is divided into four steps: First, the two source images and their fusion results are divided into blocks; second, evaluating the visual information of the first step’s block results, both with and without distortion; third, calculating the VIF of each sub-band; fourth, calculate the overall indicators.

MI indicates the amount of information obtained from the source image, the larger the MI, the more information is obtained.

$$MI_{X,F} = \sum_{x,f} p_{X,F}(x,f) \log \frac{p_{X,F}(x,f)}{p_X(x)p_F(f)}, \tag{20}$$

$$MI = MI_{I,F} + MI_{V,F}, \tag{21}$$

where $p_{X,F}$ represents the joint probability density, p_X , p_F represent the edge probability density, $MI_{I,F}$ and $MI_{V,F}$ represent the information content of infrared image and visible image respectively.

CE is the average value of the relative entropy between the two source images and the fused image, which is used to characterize the pixel difference at the corresponding position between the two source images and the fused image [42]. The smaller the CE, the better the image fusion effect:

$$D(p \parallel q) = \sum_{i=0}^{L-1} p(i) \log_2 \frac{p(i)}{q(i)}, \tag{22}$$

$$CE(I, V, F) = \frac{D(h_I \parallel h_F) + D(h_V \parallel h_F)}{2},$$

where $p(i)$ and $q(i)$ are two probability distribution functions, h_I , h_V , and h_F are the normalized histograms of infrared image, visible image and source image, respectively.

4.3. Subjective Evaluation

Figures 8 and 9 show the comparison results of our method with the nine methods on TNO and MSRS datasets, respectively. The nine methods are as follows: (a) GTF [43], (b) WLS [44], (c) MLGCF [28], (d) ResNet50 [34], (e) RFN -Nest [26], (f) DenseFuse [25], (g) U2Fusion [27], (h) FusionGAN [45], (i) GANMcC [46]. Where (a–c) are representative of traditional methods, (d–i) are advanced deep learning methods in recent years, and (j) is our method. Targets that are not significant in the comparison methods are marked with red rectangles, and poor details are marked with green rectangles.

4.3.1. Subjective Evaluation on the TNO Datasets

The subjective evaluation of fusion results on the TNO dataset is shown in Figure 8. As can be seen, in Figure 8I, the low contrast of the person in (a,d,e–g) indicates that some infrared information was lost during the fusion process. The color of the leaves in (b,e–h) are inconsistent and there are obvious abrupt changes. The eaves and fences of (a,c,i) are not clear, indicating that some details are missing. The overall visual fidelity in (a,h,i) is low, resulting in distortion of the person. By contrast, our method (j) not only can better highlight the target person, but also retain detail information such as leaves and fences.

In Figure 8II–V, the (d–g) have less significant target, and the (a) is more toward the infrared images. Specifically, in Figure 8IV, the bright grass in (a) is not extracted in the place circled by the red box, the dark trees in (b–d,h,i) are not highlighted, which makes it difficult to see the specific background information as a whole. In Figure 8V, the leaves of bushes in (a,d,e,h,i) are not clear and do not highlight the gradually bright characteristics of bushes from the lower left to the upper right. Compared with other methods, our method performs well, especially in detail extraction and target saliency. However, the infrared targets in (b,c,f) of Figure 8II,IV are more natural compared to (j), so further validation of the fusion effect using objective quality evaluation is needed.

4.3.2. Subjective Evaluation on the MSRS Datasets

Figure 9 shows our results on MSRS. It can be seen that the target significance of (a) and (d–g) is low. The sky colors of (a) and (d–h) in Figure 9I show abrupt changes, which do not conform to human visual observation. The pipe details of (a–i) and the ground texture of (d) and (f–h) in Figure 9II are not clear. In Figure 9III, there is obvious noise in (g), and the road color in (a,e,f,h,i) is closer to the infrared image. In Figure 9IV, the contrast between (d,e,h,i) is low, and the water beach is not obvious. In Figure 9V, the road surface (b,c) appears noisy, and the vehicles (a,d,e) are blurred. In contrast, our method shows great fusion effects, especially in terms of contrast enhancement and highlighting the target.

4.4. Objective Evaluation

4.4.1. Objective Evaluation on the TNO Datasets

We employ the six evaluation metrics mentioned in Section 4.2 to objectively evaluate 20 groups of infrared and visible images on the TNO dataset, and the results are shown in Figure 10. The values in the legend represent the average of the metrics after removing the maximum and minimum values.

Figure 10 shows that our method is superior to other comparison methods, which is consistent with the subjective evaluation results. The significant improvement in EN indicates that our method performs well in terms of information retention. The improvement of SD and VIF indicates that the fusion results of our method have high contrast and good visual effect. This is because we develop FT++ to obtain the saliency map of the infrared base layer. The increase in the AG indicates an improvement in the clarity of the fusion results. The improvement in the MI indicates that the fusion results are rich in information

from the source images. The reduction in CE indicates that the difference between our fusion results and the two source images is smaller. All these are due to the combined use of our decomposition method and fusion rules, which gives the overall method a significant advantage in terms of both information retention and target saliency enhancement.

4.4.2. Objective Evaluation on the MSRS Datasets

Figure 11 shows our objective evaluation results on the MSRS dataset. It can be seen that our method is significantly better than the other seven methods in the six evaluation metrics of EN, SD, AG, VIF, MI and CE, which indicates that our method has high contrast and obtains rich information from the source image. Although our method is not optimal in AG, it is still better than most methods. Combined with the results of the subjective evaluation, our method is visually excellent and shows an outstanding competitive advantage.

4.5. Ablation Experiments

To demonstrate that the methods can produce beneficial effects, we conduct ablation experiments for the base and large-scale layer methods separately. The experiment consists of six parts: (i) Removing the large-scale layer (ResNet50 and MWA0, etc.) fusion method from this paper and using FT++ for the base layer; (ii) removing the large-scale layer fusion method and using FT for the base layer; (iii) removing all three methods; (iv) keeping the large-scale layer fusion method from this paper only; (v) keeping the large-scale layer fusion method and using FT for the base layer; (vi) keeping the large-scale layer fusion method and using FT++ for the base layer (our method), where the removed methods are replaced with those of the corresponding scales from the Ref. [28]. Table 2 shows the objective evaluation metric values after the average value of the 20 groups of image fusion results on TNO datasets.

Table 2. Ablation experiment setup and the average value of the evaluation metric of a fused image.

	ResNet50	FT++	FT	EN	SD	AG	VIF	MI	CE
(i)	-	√	-	7.1781	46.2157	4.7924	0.9593	3.9047	0.9390
(ii)	-	-	√	7.1761	48.6513	4.9052	0.9387	3.9015	0.7440
(iii)	-	-	-	6.8387	37.4755	4.5438	0.7360	2.2591	1.2590
(iv)	√	-	-	6.8216	35.9339	4.2283	0.7613	2.2544	1.5441
(v)	√	-	√	7.1637	48.2857	4.6031	0.9209	3.6287	0.7625
(vi)	√	√	-	7.1796	49.1417	5.4928	0.8774	3.2227	0.7122

It can be seen that our method achieves the optimum in EN, SD, AG, and CE metrics, which shows that our method has great advantages in information and contrast. The main reason for the lack of advantages in VIF and MI is that our method discarded unnecessary redundant information in the fusion process, which led to the reduction of some reference-based evaluation metrics. However, combined with the subjective evaluation results, our method has clear details and prominent targets, so it provides a good visual effect. In addition, it can be seen from the values of each evaluation metric that the FT and FT++ methods have higher metric values, and among these methods, the combination of ResNet50 and FT++ methods have the best overall performance. This phenomenon shows that the method achieves a better fusion effect overall. It can therefore be shown that the FT++ method and the MWA0 method based on the average operator help to improve the image fusion quality.

4.6. Fusion Performance Analysis

Based on the above subjective and objective evaluation, it can be seen that our method is significantly better than other methods, which proves that our method can more effectively obtain high-quality fused images.

Because the hybrid multi-scale decomposition and ResNet50 fusion rule are relatively time-consuming, the running time of our method is slightly longer than that of other

traditional methods. However, in terms of fusion effect, compared with the optimal method in the comparison method on TNO dataset, the EN, SD, AG, VIF, MI, and CE of our method are improved by 3.82%, 29.78%, 5.47%, 14.96%, 3.82%, and 30.41% on average, respectively. On MSRS dataset, the EN, SD, VIF, MI, and CE of our method are improved by 5.79%, 14.06%, 24.46%, 5.79%, and 17.51% on average, respectively. Analyzing the above time and performance together, our method is far superior in performance to other comparative methods. Therefore, the cost of a reasonable increase in running time is feasible and worthwhile in order to obtain better fusion results for precise and widespread application in various fields.

5. Conclusions

In this study, we design a novel infrared and visible image fusion method based on significant target enhancement. The proposed method solves the problem regarding the preservation of thermal radiation features. MLGCF is deployed to decompose the source image and extract useful information accurately. To provide smooth and prominent base layers for fusion results, we propose a new method (FT++) to construct the fusion weights for the base layer. Large-scale features are extracted and processed by ResNet50 and ZCA in such a way as to preserve useful details in the source images effectively. The subjective and objective comparison results on TNO and MSRS datasets demonstrate that our method achieves better results compared to the traditional and deep learning-based alternatives. Although our method has shortcomings in terms of running efficiency, the fusion results are improved significantly over other approaches. In upcoming future research, we will further improve this method by reducing the time consumption and deploy it to the target detection task.

Author Contributions: Conceptualization Y.D. and X.H.; methodology, Y.D.; software, Y.D.; validation, X.H., Y.D. and K.S.; data curation, K.S.; writing—original draft preparation, Y.D.; writing—review and editing, X.H., Y.D. and K.S.; funding acquisition, X.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 61872407.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Tang, L.F.; Yuan, J.T.; Ma, J.Y. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Inf. Fusion* **2022**, *82*, 28–42. [[CrossRef](#)]
2. Chen, J.; Li, X.J.; Luo, L.B.; Mei, X.G.; Ma, J.Y. Infrared and visible image fusion based on target-enhanced multiscale transform decomposition. *Inf. Sci.* **2020**, *508*, 64–78. [[CrossRef](#)]
3. Zhou, S.Y.; Yang, P.X.; Xie, W.L. Infrared image segmentation based on Otsu and genetic algorithm. In Proceedings of the 2011 International Conference on Multimedia Technology, Hangzhou, China, 26–28 July 2011; pp. 5421–5424.
4. Zhao, Y.F.; Cheng, J.C.; Zhou, W.; Zhang, C.X.; Pan, X. Infrared pedestrian detection with converted temperature map. In Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 18–21 November 2019; pp. 2025–2031.
5. Ma, J.Y.; Tang, L.F.; Fan, F.; Huang, J.; Mei, X.G.; 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](#)]
6. Li, G.F.; Lin, Y.J.; Qu, X.D. An infrared and visible image fusion method based on multi-scale transformation and norm optimization. *Inf. Fusion* **2021**, *71*, 109–129. [[CrossRef](#)]
7. Liu, Y.; Chen, X.; Liu, A.P.; Ward, R.K.; Wang, Z.J. Recent advances in sparse representation based medical image fusion. *IEEE Instrum. Meas. Mag.* **2021**, *24*, 45–53. [[CrossRef](#)]
8. Wang, D.; Liu, J.Y.; Fan, X.; Liu, R.S. Unsupervised Misaligned Infrared and Visible Image Fusion via Cross-Modality Image Generation and Registration. *arXiv* **2022**, arXiv:2205.11876.

9. Burt, P.J.; Adelson, E.H. The Laplacian pyramid as a compact image code. In *Readings in Computer Vision*; Morgan Kaufmann: Burlington, MA, USA, 1987; pp. 671–679.
10. Chipman, L.J.; Orr, T.M.; Graham, L.N. Wavelets and image fusion. *Int. Conf. Image Process.* **1995**, *3*, 248–251.
11. Shao, Z.F.; Liu, J.; Cheng, Q.M. Fusion of infrared and visible images based on focus measure operators in the curvelet domain. *Appl. Opt.* **2012**, *51*, 1910–1921.
12. Li, H.; Manjunath, B.S.; Mitra, S.K. Multisensor image fusion using the wavelet transform. *Graph. Model. Image Process.* **1995**, *57*, 235–245. [[CrossRef](#)]
13. Li, S.T.; Kang, X.D.; Hu, J.W. Image fusion with guided filtering. *IEEE Trans. Image Process.* **2013**, *22*, 2864–2875.
14. Zhou, Z.Q.; Wang, B.; Li, S.; Dong, M.J. Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with Gaussian and bilateral filters. *Inf. Fusion* **2016**, *30*, 15–26. [[CrossRef](#)]
15. He, K.M.; Sun, J.; Tang, X.O. Guided image filtering. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 1397–1409. [[CrossRef](#)] [[PubMed](#)]
16. Jia, W.B.; Song, Z.H.; Li, Z.G. Multi-scale Fusion of Stretched Infrared and Visible Images. *Sensors* **2022**, *22*, 6660. [[CrossRef](#)] [[PubMed](#)]
17. Li, H.; Chan, T.N.; Qi, X.B.; Xie, W.Y. Detail-preserving multi-exposure fusion with edge-preserving structural patch decomposition. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 4293–4304. [[CrossRef](#)]
18. Do, M.N.; Vetterli, M. The contourlet transform: An efficient directional multiresolution image representation. *IEEE Trans. Image Process.* **2005**, *14*, 2091–2106. [[CrossRef](#)] [[PubMed](#)]
19. Liu, Y.; Chen, X.; Peng, H.; Wang, Z.F. Multi-focus image fusion with a deep convolutional neural network. *Inf. Fusion* **2017**, *36*, 191–207. [[CrossRef](#)]
20. Li, H.; Wu, X.J.; Kittler, J. Infrared and visible image fusion using a deep learning framework. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 2705–2710.
21. Chen, J.; Yang, Z.F.; Chan, T.N.; Li, H.; Hou, J.H.; Chau, L.P. Attention-Guided Progressive Neural Texture Fusion for High Dynamic Range Image Restoration. *IEEE Trans. Image Process.* **2022**, *31*, 2661–2672. [[CrossRef](#)]
22. Bai, H.R.; Pan, J.S.; Xiang, X.G.; Tang, J.H. Self-Guided Image Dehazing Using Progressive Feature Fusion. *IEEE Trans. Image Process.* **2022**, *31*, 1217–1229. [[CrossRef](#)] [[PubMed](#)]
23. Li, H.F.; Cen, Y.L.; Liu, Y.; Chen, X.; Yu, Z.T. Different Input Resolutions and Arbitrary Output Resolution: A Meta Learning-Based Deep Framework for Infrared and Visible Image Fusion. *IEEE Trans. Image Process.* **2021**, *30*, 4070–4083. [[CrossRef](#)]
24. Ma, J.Y.; Xu, H.; Jiang, J.J.; Mei, X.G.; Zhang, X.P. DDcGAN: A Dual-Discriminator Conditional Generative Adversarial Network for Multi-Resolution Image Fusion. *IEEE Trans. Image Process.* **2020**, *29*, 4980–4995. [[CrossRef](#)]
25. Li, H.; Wu, X.J. DenseFuse: A fusion approach to infrared and visible images. *IEEE Trans. Image Process.* **2018**, *28*, 2614–2623. [[CrossRef](#)] [[PubMed](#)]
26. Li, H.; Wu, X.J.; Kittler, J. RFN-Nest: An end-to-end residual fusion network for infrared and visible images. *Inf. Fusion* **2021**, *73*, 72–86. [[CrossRef](#)]
27. Xu, H.; Ma, J.Y.; Jiang, J.J.; Guo, X.J.; Ling, H.B. U2Fusion: A unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 502–518. [[CrossRef](#)] [[PubMed](#)]
28. Tan, W.; Zhou, H.X.; Song, J.L.Q.; Li, H.; Yu, Y.; Du, J. Infrared and visible image perceptive fusion through multi-level Gaussian curvature filtering image decomposition. *Appl. Opt.* **2019**, *58*, 3064–3073. [[CrossRef](#)] [[PubMed](#)]
29. He, K.M.; Zhang, X.Y.; Ren, S.Q.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
30. Achanta, R.; Hemami, S.; Estrada, F.; Susstrunk, S. Frequency-tuned salient region detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 1597–1604.
31. Hou, R.C.; Nie, R.C.; Zhou, D.M.; Cao, J.D.; Liu, D. Infrared and visible images fusion using visual saliency and optimized spiking cortical model in non-subsampled shearlet transform domain. *Multimed. Tools Appl.* **2019**, *78*, 28609–28632. [[CrossRef](#)]
32. Wang, S.; Ai, H.; He, K. Difference-image-based multiple motion targets detection and tracking. *J. Image Graph.* **1999**, *4*, 470–475.
33. Ochotorena, C.N.; Yamashita, Y. Anisotropic guided filtering. *IEEE Trans. Image Process.* **2019**, *29*, 1397–1412. [[CrossRef](#)]
34. Li, H.; Wu, X.; Durrani, T.S. Infrared and visible image fusion with ResNet and zero-phase component analysis. *Infrared Phys. Technol.* **2019**, *102*, 103039. [[CrossRef](#)]
35. Available online: https://figshare.com/articles/TNO_Image_Fusion_Dataset/1008029 (accessed on 28 September 2022).
36. Tang, L.F.; Yuan, J.T.; Zhang, H.; Jiang, X.Y.; Ma, J.Y. PIAFusion: A progressive infrared and visible image fusion network based on illumination aware. *Inf. Fusion* **2022**, *83*, 79–92. [[CrossRef](#)]
37. Roberts, J.W.; Van Aardt, J.A.; Ahmed, F.B. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *J. Appl. Remote Sens.* **2008**, *2*, 023522.
38. Jin, X.; Jiang, Q.; Yao, S.W.; Zhou, D.M.; Nie, R.; Lee, S.; He, K.J. Infrared and visual image fusion method based on discrete cosine transform and local spatial frequency in discrete stationary wavelet transform domain. *Infrared Phys. Technol.* **2018**, *88*, 1–12. [[CrossRef](#)]
39. Cui, G.M.; Feng, H.J.; Xu, Z.H.; Li, Q.; Chen, Y.T. Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition. *Opt. Commun.* **2015**, *341*, 199–209. [[CrossRef](#)]

40. Li, H.; Wu, X.J.; Kittler, J. MDLatLRR: A novel decomposition method for infrared and visible image fusion. *IEEE Trans. Image Process.* **2020**, *29*, 4733–4746. [[CrossRef](#)]
41. Qu, G.H.; Zhang, D.L.; Yan, P.F. Information measure for performance of image fusion. *Electron. Lett.* **2002**, *38*, 313–315. [[CrossRef](#)]
42. Haghghat, M.B.A.; Aghagolzadeh, A.; Seyedarabi, H. A non-reference image fusion metric based on mutual information of image features. *Comput. Electr. Eng.* **2011**, *37*, 744–756. [[CrossRef](#)]
43. Ma, J.Y.; Chen, C.; Li, C.; Huang, J. Infrared and visible image fusion via gradient transfer and total variation minimization. *Inf. Fusion* **2016**, *31*, 100–109. [[CrossRef](#)]
44. Ma, J.L.; Zhou, Z.Q.; Wang, B.; Zong, H. Infrared and visible image fusion based on visual saliency map and weighted least square optimization. *Infrared Phys. Technol.* **2017**, *82*, 8–17. [[CrossRef](#)]
45. Ma, J.Y.; Yu, W.; Liang, P.W.; Li, C.; Jiang, J.J. FusionGAN: A generative adversarial network for infrared and visible image fusion. *Inf. Fusion* **2019**, *48*, 11–26. [[CrossRef](#)]
46. Ma, J.Y.; Zhang, H.; Shao, Z.F.; Liang, P.W.; Xu, H. GANMcC: A Generative Adversarial Network with Multi-classification Constraints for Infrared and Visible Image Fusion. *IEEE Trans. Instrum. Meas.* **2020**, *70*, 1–14. [[CrossRef](#)]