applied sciences

MDPI

Article

Multi-Scale Feature Fusion with Attention Mechanism Based on CGAN Network for Infrared Image Colorization

Yibo Ai [1,2], Xiaoxi Liu [1], Haoyang Zhai [1], Jie Li [3], Shuangli Liu [4], Huilong An [3] and Weidong Zhang [1,*]

[1] National Center for Materials Service Safety, University of Science and Technology Beijing, Beijing 100083, China; ybai@ustb.edu.cn (Y.A.)
[2] Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Zhuhai 519082, China
[3] HBIS Materials Institute, No. 385 South Sports Street, Yuhua District, Shijiazhuang 050023, China
[4] Hesteel Group Tangsteel Company, No. 9 Binhe Road, Tangshan 063000, China
* Correspondence: zwd@ustb.edu.cn

Abstract: This paper proposes a colorization algorithm for infrared images based on a Conditional Generative Adversarial Network (CGAN) with multi-scale feature fusion and attention mechanisms, aiming to address issues such as color leakage and unclear semantics in existing infrared image coloring methods. Firstly, we improved the generator of the CGAN network by incorporating a multi-scale feature extraction module into the U-Net architecture to fuse features from different scales, thereby enhancing the network's ability to extract features and improving its semantic understanding, which improves the problems of color leakage and blurriness during colorization. Secondly, we enhanced the discriminator of the CGAN network by introducing an attention mechanism module, which includes channel attention and spatial attention modules, to better distinguish between real and generated images, thereby improving the semantic clarity of the resulting infrared images. Finally, we jointly improved the generator and discriminator of the CGAN network by incorporating both the multi-scale feature fusion module and attention mechanism module. We tested our method on a dataset containing both infrared and near-infrared images, which retains more detailed features while also preserving the advantages of existing infrared images. The experimental results show that our proposed method achieved a peak signal-to-noise ratio (PSNR) of 16.5342 dB and a structural similarity index (SSIM) of 0.6385 on an RGB-NIR (Red, Green, Blue-Near Infrared) testing dataset, representing a 5% and 13% improvement over the original CGAN network, respectively. These results demonstrate the effectiveness of our proposed algorithm in addressing the issues of color leakage and unclear semantics in the original network. The proposed method in this paper is not only applicable to infrared image colorization but can also be widely applied to the colorization of remote sensing and CT images.

Keywords: attention mechanism module; Generative Adversarial Network (GAN); image colorization; infrared images; multi-scale feature fusion

check for updates

Citation: Ai, Y.; Liu, X.; Zhai, H.; Li, J.; Liu, S.; An, H.; Zhang, W. Multi-Scale Feature Fusion with Attention Mechanism Based on CGAN Network for Infrared Image Colorization. *Appl. Sci.* **2023**, *13*, 4686. https://doi.org/10.3390/app13084686

Academic Editor: Christos Bouras

Received: 12 March 2023
Revised: 29 March 2023
Accepted: 3 April 2023
Published: 7 April 2023

Copyright: © 2023 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

Color is an essential information for understanding the world around us, as human eyes are significantly more sensitive to color than to black and white. Therefore, color images provide us with more abundant visual information. Image colorization refers to assigning new colors to each pixel of a grayscale image. Nowadays, image colorization has been extensively applied in various fields, such as artistic creation [1], remote sensing [2,3], medical imaging [4], comics [5], infrared imaging [6], etc. Infrared radiation is emitted by objects and can be captured by infrared cameras, producing infrared images. Infrared imaging technology originated in the military field, where it was used to manufacture a series of military products such as aiming devices and guidance heads. With the development and maturity of infrared imaging technology, the application scenarios of infrared images

have gradually expanded. Infrared cameras can be found in agriculture [7], medicine [8], electricity [9], and other fields. Infrared thermal imaging technology can be used for various types of detection, such as equipment failure detection and material defect detection [10], and is widely used in the steel industry. Infrared imaging captures information that is invisible to the human eye, which makes it valuable. However, infrared images captured at night are grayscale images that lack accuracy and details, limiting their interpretability and usability, and causing great difficulties in post-processing. To address the problem of background residue or target defect in the detection of infrared dark targets in complex marine environments [11], infrared image colorization can be used to separate targets from the background, thereby improving detection accuracy. Due to the potential of infrared images in the aforementioned fields of application, there has been a great deal of attention paid to infrared image colorization.

There are two main categories of traditional colorization methods: user-assisted techniques based on manual painting and reference example-based methods. The former method relies on manually applying partial colorization to the image, which can be subject to human subjective bias and ultimately affect the quality of the coloring results. The latter relies heavily on the reference image used, and using an inappropriate example can lead to colorization that is grossly mismatched with the actual colors in the image. Additionally, the complexity and unique characteristics of infrared images make it difficult for these methods to fully leverage the information within the image, resulting in low-quality colorization results. Therefore, these two traditional methods are not well-suited for the task of infrared image colorization.

In recent years, deep learning has been widely applied in the field of computer vision. This approach has greatly solved the problem of complex human–machine system interaction processes that exist in the first two methods. By training machines to autonomously complete the coloring task, the application of deep neural networks to image coloring problems has also received good feedback [12–15].

Colorizing infrared images presents unique challenges compared to colorizing grayscale images due to the need to estimate both luminance and chromaticity. Berg et al. [16] proposed two different methods to estimate color information for infrared images, with the first method estimating both luminance and chromaticity and the second method predicting chromaticity using grayscale image colorization after predicting the luminance. However, the lack of corresponding datasets for NIR and RGB images limited evaluations to traffic scenes. Nyberg et al. [17] addressed this issue by using unsupervised learning with CycleGAN to generate corresponding images, but the approach suffered from distortion issues. Xian et al. [18] tackled the challenge of modal differences between infrared and visible images by generating grayscale images as auxiliary information and using point-by-point transformation for single-channel infrared images.

Traditional grayscale image coloring methods have a complicated workflow that cannot be applied to colorizing NIR images. The GAN network, with its unique network structure and training mechanism, has been widely applied by researchers in the field of infrared image coloring [19–22]. Wei et al. [19] proposed an improved Dual GAN architecture that uses two deep learning networks to establish the translation relationship between NIR and RGB images without the need for prior image pairing and labeling. Xu et al. [20] proposed a DenseUnet generative adversarial network for colorizing near-infrared (facial) images. The DenseNet effectively extracts facial features by increasing the network depth, while the Unet preserves important facial details through skip connections. With improvements to the network structure and loss constraints, this method can minimize facial shape distortion and enrich facial detail information in near-infrared facial images. Although there has been some progress in converting infrared images to color images, the problem of semantic coding entanglement and geometric distortion remains unresolved. Luo et al. [21] therefore proposed a top-down generative adversarial network called PearlGAN that aligns attention and gradient. By introducing attention-guided and loss modules, and adding a gradient alignment loss module, they reduced the ambiguity of semantic

coding and improved the edge consistency of input–output images. However, due to an incomplete understanding of image features and limited information acquisition in existing methods, issues such as color leakage and loss of details still exist. Li et al. [22] introduced a multi-scale attention mechanism into the task of infrared and visible light image fusion, extracting attention maps from multi-scale features of both types of images and adding them to the GAN network to preserve more details. Liu et al. [23] also addressed this issue by proposing a deep network for infrared and visible light image fusion, consisting of a feature learning module and a fusion learning mechanism, and designing an edge-guided attention mechanism on multi-scale features to guide attention to common structures during the fusion process. However, because the network is trained based on edge attention, during image reconstruction, the middle features tend to pay more attention to texture details, which makes it less effective when the source image contains a large amount of noise.

Building upon the previous methods, this paper proposes an infrared image coloring algorithm based on the CGAN with a multi-scale feature fusion and attention mechanism. CGAN is an extension of GAN [24], which controls the image generation process by adding conditional constraints, resulting in better image quality and more detailed output. In this work, we improved the generator architecture of CGAN by incorporating a multi-scale convolution module with three types of convolutions in the U-Net network to fuse different scale features, enhance the network's feature extraction ability, improve learning speed and semantic understanding ability, and address issues such as color leakage and blurring during the coloring process. We added an attention module to the discriminator, which contains both channel attention and spatial attention, to filter the feature layers from a channel perspective and select important regions on the feature map. This allows the network to focus on useful features and ignore unnecessary ones, while also improving the discriminator's effectiveness and efficiency. By combining the improvements to the generator and discriminator, a new network with multi-scale feature fusion and attention module is obtained. Finally, we tested our proposed method on a near-infrared image dataset that combines the advantages of both infrared and visible images by preserving more details, edges, and texture features from the visible light images while retaining the benefits of the infrared images.

2. Related Work

2.1. Generating Adversarial Network

GAN is designed from game theory, mainly the idea of two-person zero-sum game in game theory, and this idea is introduced into the training of generator (G) and discriminator (D), and the process of G and D training is the process of gaming these two networks [25].

The generator produces generated data that are close to the real data, and then these data are scored by the discriminator. The higher the score, the closer the discriminator thinks the image is to the real image. So, the generator will improve its ability to generate images that are as close to the real data as possible, so that the discriminator can give it a high score. The discriminator, moreover, will improve its ability to discriminate the image and maximize the distinction between the real data and the generated data. In the process of distinction, it will give the generated data a low score, as close to 0 as possible. In contrast, it will give the real data a high score, as close to 1 as possible.

The network structure of GAN is shown in Figure 1.

The input noise z passes through the generator G, which generates the data. At this point, the generator G wants to confuse the discriminator and get a higher score, that is, the generator's purpose is to output $D(G(z))$ as close as possible to 1. The role of discriminator D is to discriminate between true and false, and D wants to give the generated data a low score and the real data a high score, so the discriminator's purpose is to output $D(G(z))$ as close as possible to 0 and $D(\mathrm{x})$ as close as possible to 1, so as to achieve the purpose of correctly distinguishing the true and false data.

The ultimate goal of the network is to get a generator with good performance. This will ensure that, eventually, the data generated by the generator is close to the real data, when the discriminator has discriminated it.

The objective function of GAN is as follows [25]:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(x)}[\log(1-D(G(z)))] \quad (1)$$

where Z is noise, and X is real data.

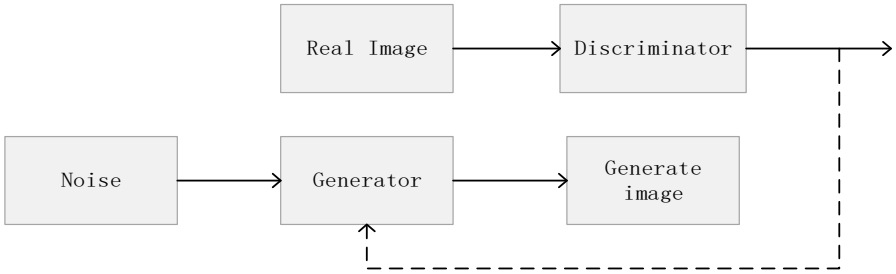


Figure 1. Generative adversarial network structure.

In Equation (1), the discriminator model expects to distinguish the real data from the generated data to the greatest extent possible. The image passes through the discriminator, the discriminator will score the image, the closer the score is to 1, the more the discriminator tends to think that the input is the real data; the closer the score is to 0, the more the discriminator tends to think that the input is the generated data, so to maximize $\log D(x)$ and $\log(1-D(G(z)))$. The generator aims to produce a generated image that is close to the real image, to let the discriminator give itself a high score. The expected score is close to 1, to minimize $\log(1-D(G(z)))$. That is, for the generator G to make V(G, D) as small as possible, the discriminator D has to make V(G, D) as large as possible.

With the development of GAN, many variants of GAN have been developed, such as DCGAN, CGAN, and CycleGAN. In this paper, we chose CGAN to improve it.

2.2. Conditional Generative Adversarial Network

CGAN was proposed by Mirza et al. in 2014 [24], because the original GAN direction is not fixed and the training effect is not stable. For example, if the training dataset contains species with large differences such as flowers, birds, and trees, it is impossible to control the type of output at the time of testing. Therefore, people began to try to add some a priori conditions to GAN to let the direction of GAN be controlled. CGAN is able to add conditional information to the generative and discriminative models, the conditional information can be any information you want to add, such as class labels, etc. In the paper, the authors use y to represent the auxiliary information. CGAN makes the application of GAN more extensive, for example, by using CGAN to accomplish the task of converting input text into images, and controlling the output content by adding conditions. The structure of the network does not change in this process, and the training process is the same as that of GAN, only the input changes. The model structure of CGAN is such that the auxiliary information is input together with the noise in the generator, and the auxiliary information is input together with the real data in the discriminator.

The objective function of CGAN is basically the same as that of GAN, but CGAN will input the conditions at the same time in the input, and its objective function is as follows [24]:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x|y)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z|y)))] \quad (2)$$

CGAN network structure is shown in Table 1.

Table 1. Generator structure.

Convolution Layer	Convolution Kernel	Step Length	Number of Channels
Conv1	4 × 4	2	64
Conv2	4 × 4	2	128
Conv3	4 × 4	2	256
Conv4	4 × 4	2	512
Conv5	4 × 4	2	512
Conv6	4 × 4	2	512
Conv7	4 × 4	2	512
Conv8	4 × 4	2	512
ConvTranspose1	4 × 4	2	512
ConvTranspose2	4 × 4	2	512
ConvTranspose3	4 × 4	2	512
ConvTranspose4	4 × 4	2	512
ConvTranspose5	4 × 4	2	256
ConvTranspose6	4 × 4	2	128
ConvTranspose7	4 × 4	2	64
ConvTranspose8	4 × 4	2	2

The structure of the discriminator is shown in Table 2.

Table 2. Discriminator network structure.

Convolution Layer	Convolution Kernel	Step Length	Number of Channels
Conv1	4 × 4	2	64
Conv2	4 × 4	2	128
Conv3	4 × 4	2	256
Conv4	4 × 4	1	512
Conv5	4 × 4	1	1

3. Improve Image Colorization Generation Network

Since coloring techniques nowadays still suffer from problems such as color leakage and semantic errors, this paper will improve the generative adversarial network base generator by proposing to incorporate multi-scale convolutional modules to aggregate features of various scales.

3.1. U-Net Generation Network

The U-Net network is a model used for semantic segmentation [26]. It is called U-Net because of its U-shaped network structure. The U-Net is symmetrical on the left and right sides, with the left side performing downsampling for feature extraction, and the right side performing upsampling for feature reconstruction. Since infrared images usually contain some weak edge information, the U-Net network is better able to preserve these details in image colorization, resulting in higher image accuracy. Compared to VGG-16 and ResNet, the U-Net network has fewer parameters and trains faster. Moreover, in some infrared image colorization tasks, there may be a large imbalance in the number of pixels of different colors, but the structure of the U-Net network allows it to better handle such unbalanced datasets.

We compared the performance of U-Net with two other networks for colorizing near-infrared images. As shown in Figure 2, VGG-16 produces images with low saturation and color distortion after colorization. Although ResNet improves upon the colorization results of VGG-16, it still cannot fully restore the colors of the original image. Moreover, the images colorized by U-Net are closest to the real images, with clearer object edges. Therefore, we chose U-Net as the generator and made improvements to it.

As the number of convolutions increases, each convolutional layer loses some information from the previous layer's feature map. This loss of information can have a significant impact on the quality of the reconstructed image during upsampling. However, the U-Net

network utilizes skip connections to concatenate and merge low-level and high-level feature maps, resulting in feature maps that contain both low-level and high-level information. For U-Net, information from each scale is important as it can capture multi-scale feature information and thus improve feature extraction capability. The U-Net structure is shown in Figure 3.



Figure 2. Comparison of NIR image colorization for generator networks. (**a**) Real image; (**b**) NIR image; (**c**) VGG-16; (**d**) ResNet; and (**e**) U-Net.

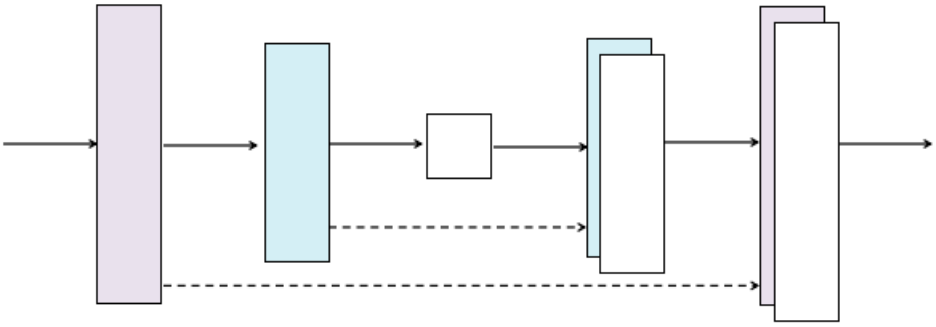


Figure 3. U-Net structure.

3.2. Adding a Multi-Scale Feature Fusion Module to Improve the Generative Network U-Net

The existing algorithms for infrared image coloring still suffer from color leakage and semantic unclearness, leading to the loss of details, which is caused by insufficient extraction of features. Additionally, GAN includes two parts, generator G and discriminator D, where G generates data directly, and the proposed features by G in the downsampling process will directly affect the quality of reconstructed images. So, in this paper, in order to enhance the understanding of the semantic features of the image, the multi-scale feature fusion module is added to the generator. The improved network, by expanding the receptive field through atrous convolution, uses group convolution to motivate the network to have a good understanding of the features even with the information of only local channels. Shift convolution, again by permuting the channels, allows the network to have a deeper understanding of the features. Finally, the features are fused together to improve the feature extraction ability of the whole generator.

This subsection improves on the U-Net by adding multi-scale feature extraction to enhance the performance of the generator. The improved network structure is shown in Figure 4.

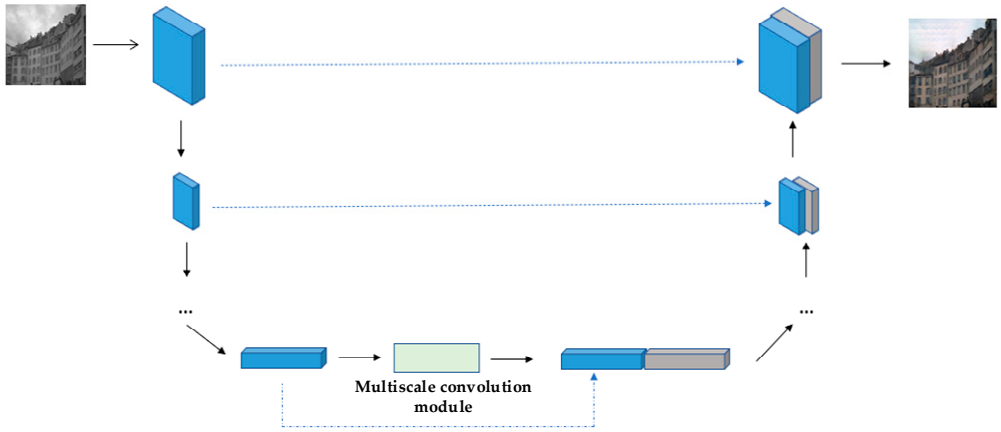


Figure 4. Improved U-Net.

The original U-Net includes 8 convolutional layers, with convolutional kernel size k = 4, step size s = 2, and padding layer padding = 1 used in the convolution operation. The size of the feature map is reduced by half for each convolution operation throughout the downsampling process. The number of channels is increased from 1 channel for the input, to 64 channels, and then exponentially to 512 channels for the hold. For upsampling, the transpose convolution operation is also performed 8 times. The transposed convolution kernel size k = 4, s = 2, and padding = 1.

In order to effectively prevent color leakage and deepen the network's understanding of semantic features, the multi-scale feature extraction module is added after the last layer of U-Net downsampling to fuse the features.

The multi-scale convolution module is shown specifically in Figure 5.

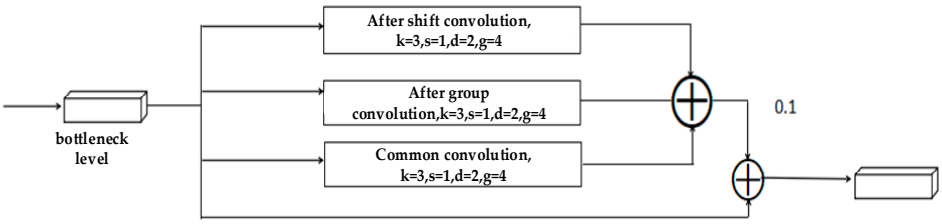


Figure 5. Multi-scale convolution module.

The U-Net is initially composed of eight convolutional layers with a downsampling structure, no pooling layer, and the activation function of the network is the LeakyReLU function. Then, the multi-scale convolution module is added after the bottleneck layer of U-Net, which has three convolutions in parallel. The first convolution employs shift convolution and separates the network's feature map into two parts, labeled A and B. The AB feature maps are transposed to create the BA combination, which is then subjected to group convolution and dilated convolution with g = 4 and d = 2, respectively. This increases the receptive field from 3 × 3 to 5 × 5. The second convolution is grouped convolution, which is divided into four groups and then convolved separately, and the last convolution layer is ordinary convolution. The three convolution operations are summed in parallel, then multiplied by a factor of 0.1 and summed with the input before convolution. This is equivalent to adding another residual operation to transpose the output of this module for convolution.

Group convolution improves the learning speed of the network by reducing the number of parameters, and can perform a whole action on different features. While ordinary convolution extracts only one kind of feature, grouped convolution can divide the feature layer into several parts, extract different features from them, and aggregate the different features. By using atrous convolution, the receptive field of the network is increased, and the loss of information in increasing the receptive field is avoided as much as

possible, and at the same time, multi-scale contextual information can be obtained, which enhances the ability of the network to extract features.

4. Improve Image Colorization Discriminant Network

The training process of generating adversarial networks is the process of gaming between the generator and the discriminator. Improving the discriminator's ability to discriminate images can also drive the generating ability of the generator to improve, so the discriminator will be modified in this section.

4.1. Discriminant Network PatchGAN

The "PatchGAN" is a commonly used discriminator in GAN. When GAN was first proposed, the discriminator only output a single evaluation score to judge whether the generated image was real or fake, and this score was obtained by evaluating the entire image data. PatchGAN uses the full convolution form, and the final output is not a single evaluation value, but the input image is convolved layer by layer to map the input image into the form of an N × N matrix. This matrix replaces the one evaluation value of the previous GAN into a separate evaluation of N × N regions, which is what Patch means. While the original GAN uses only one value to evaluate the image, PatchGAN produces a multi-value evaluation by evaluating N × N regions, and obviously PatchGAN focuses on more details. It was demonstrated in [27] that the chosen N can be much smaller than the size of the global image and still give good results. This experimental result also demonstrates that N can be chosen to be a smaller number, which also reduces the number of parameters and works for images of arbitrary size. The pattern of PatchGAN for patch scoring is shown in Figure 6.

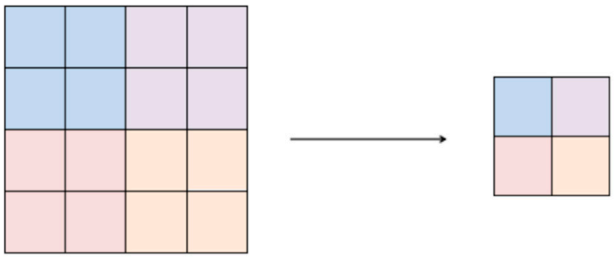


Figure 6. The pattern of PatchGAN for patch scoring.

4.2. Attention Mechanism

Convolutional Block Attention Module (CBAM) is a simple and effective attention module [28], which consists of two parts: channel attention and spatial attention. CBAM calculates the attention weights corresponding to the two dimensions. It uses the attention weights to weigh the input features so that the feature map can highlight the important features more. Each channel is a feature, and channel attention allows the network to focus more on meaningful features. Not all regions of a feature map are necessarily useful information, and each region has a different level of importance. Therefore, spatial attention lets the network find important regions to focus on.

The structure diagram of channel attention is shown in Figure 7.

The operation process of CBAM is divided into two steps, the first step is to calculate the channel attention *Mc*, and then to calculate the spatial attention *Ms*. As shown in Figure 8, F denotes the input feature map; the input feature map is pooled by maximum pooling and mean pooling, and after pooling, the summation operation is performed through the fully connected layer, and finally, the channel attention *Mc* is obtained through the sigmod function. The feature map *F* is subjected to the operation of *Mc*, resulting in a further feature map. The feature map is then subjected to maximum pooling and mean pooling, and the features are stitched together after pooling, and then obtain the spatial attention *Ms* by convolution and sigmod. Finally, *Ms* is multiplied with the feature map

to get the features that pass through the attention channel. The whole process can be formulated as follows:

$$F' = Mc(F) \otimes F \tag{3}$$

$$F'' = Ms(F') \otimes F' \tag{4}$$

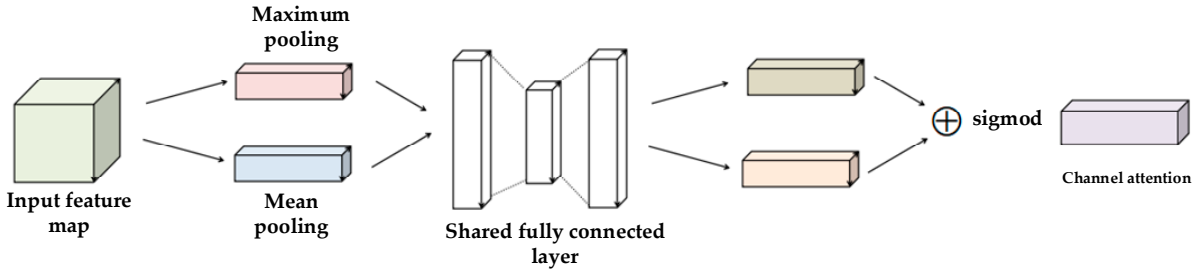


Figure 7. The structure diagram of channel attention.

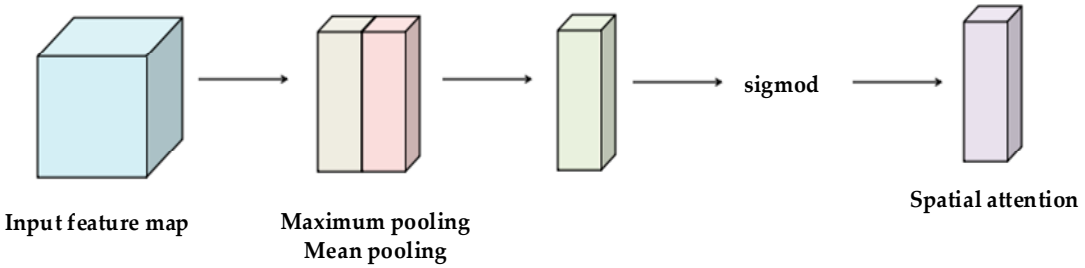


Figure 8. The structural diagram of spatial attention.

4.3. Improved Discriminative Network PatchGAN with Added Attention Module

The GAN network consists of two modules, the generator module and the discriminator module. In this paper, we first improved the generator by adding a multi-scale feature fusion module to the generator U-Net network to improve the feature extraction ability of the generator and improve the coloring effect of the infrared images. In the GAN network, the generator and the discriminator promote each other through confrontation. When the ability of the generator network improves, the discriminator must improve its own discriminatory ability to not be confused by the generator.

Therefore, in this section, we improved the discriminator by adding an attention module to the discriminator network PatchGAN. When the input passes through, it will first pass through the channel attention and then the spatial attention. Through the attention module, we obtain the channel attention weight, and then multiply the input features with the channel attention weight to get a new feature map, which highlights the features and ignores the unimportant features to complete the filtering of the feature layer in the channel perspective. Then, the spatial attention module is used to filter the important regions on the feature map as well. The attention module allows PatchGAN to focus more on useful features and ignore unnecessary features, and also improves the learning efficiency of the network.

The structure of PatchGAN is composed of five convolutional layers, using LeakyReLU as the activation function. At the end of the fourth convolution, the attention mechanism module is added, which is composed of two parts, channel attention and spatial attention, in series. The structure of the improved network is shown in Figure 9. The channel attention module goes through a mean pooling and maximum pooling first, which will go through a convolution kernel of 1×1 and a ReLU activation function, and then a convolution kernel of 1×1 and a Sigmod activation function. The output is fed to the spatial attention module, which goes through a convolution with a convolution kernel of 7×7 and a Sigmod activation function, as shown in Figure 10.

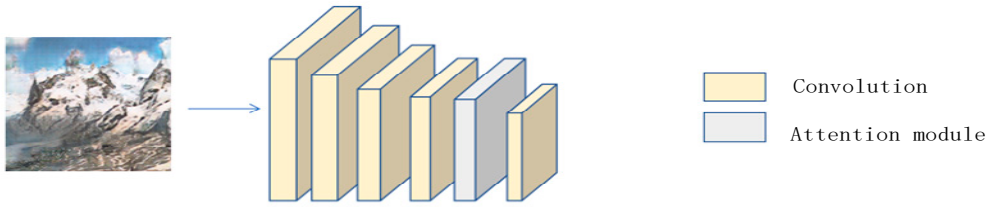


Figure 9. Improved PatchGAN.

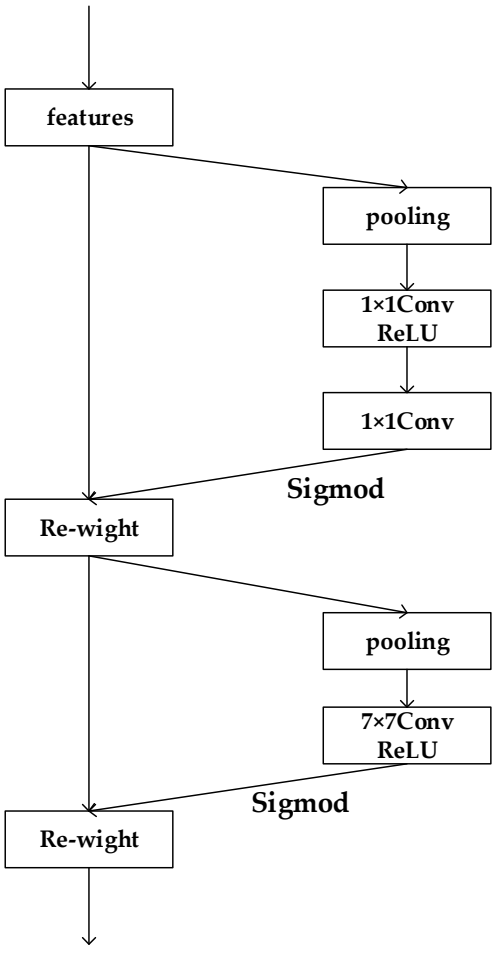


Figure 10. Attention mechanism module.

Finally, the generator and discriminator are improved simultaneously to reconstruct the network. The improvement of the network consists of two parts, the first part improves the generator by adding a multi-scale feature fusion module to it to improve the semantic understanding of the network, and the second part improves the discriminator by adding an attention mechanism to the discriminator network to make the network more focused on useful features and improve the network recognition ability. Both networks have further improvements in feature extraction ability, so the problems of semantic uncertainty and color leakage for infrared image coloring are improved, and the next experimental validation will be performed on the NIR image dataset.

5. Experimental Results and Evaluation

5.1. Experimental Environment and Settings

The data used in this experiment are a scene dataset of RGB-NIR released by Brown in 2011 [29]. This dataset includes 477 RGB images and 477 NIR images, and is rich in scenes including 9 categories such as field, forest, and indoor, as Table 3 shows. The experiments

in this paper are conducted based on this multi-category dataset rather than training for one of the categories individually. The training and test sets are divided in the ratio of 8:2, where the training set is 381 images and the test set is 96 images, and the images are uniformly sized to 256 × 256 pixels.

Table 3. Data set introduction.

Category	Total	Number of Training Sets	Number of Test Sets
country	52	43	9
field	51	40	11
forest	53	45	8
indoor	56	42	14
mountain	55	44	11
old building	51	39	12
street	50	44	6
urban	59	47	11
water	51	37	14

The Adam algorithm [30] was used to optimize the loss function with initial learning rates $lr = 0.0001$, $\beta_1 = 0.5$, $\beta_2 = 0.999$, weight decay of 0.005, batch_size = 1, and epoch = 300.

5.2. Evaluation Indicators

What is sought in image coloring is not a fixed color, but as realistic a color as possible, because, in reality, objects will inherently have multiple colors, and it is reasonable for a car to be red or black, as long as it looks realistic. Therefore, the evaluation of image coloring, in general, is subjective, and there is no uniform quantitative standard to evaluate colorized images [17]. In this paper, two metrics were selected to quantitatively analyze the experiment; peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) are commonly used image evaluation metrics.

(1) PSNR

PSNR is based on the error between pixel points in dB, and the value of PSNR is negatively correlated with the distortion of the image. Its calculation expression is shown in Equations (5) and (6) and represents the mean square error between two images.

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\|I(i,j) - k(i,j)\|^2 \tag{5}$$

$$PSNR = 10\log_{10}\left(\frac{MAX_I^2}{MSE}\right) = 20\log_{10}\left(\frac{MAX_I}{MSE}\right) \tag{6}$$

where m is the total number of pixels of image I; n is the total number of pixels of image K; and I (i, j) is the pixel value of image I at (i, j). k (i, j) is the pixel value of image K at (i, j); MAX_I is the maximum value of the image I color.

(2) SSIM

SSIM evaluates the similarity of two images in terms of brightness, contrast, and structure. The formula for calculating structural similarity is shown in Equation (7).

$$\mu_x = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} X(i,j) \tag{7}$$

$$\sigma_x = \left(\frac{1}{H \times W - 1}\sum_{i=1}^{H}\sum_{j=1}^{W}\left(X(i,j) - {\mu_x}^2\right)^{\frac{1}{2}}\right) \tag{8}$$

$$\sigma_{XY} = \frac{1}{H \times W - 1}\sum_{i=1}^{H}\sum_{j=1}^{W}(X(i,j) - \mu_X)(Y(i,j) - \mu_Y) \tag{9}$$

SSIM is evaluated in three aspects.

$$l(X,Y) = \frac{2\mu_X\mu_Y + C_1}{\mu_X^2 + \mu_Y^2 + C_1} \tag{10}$$

$$c(X,Y) = \frac{2\sigma_X\sigma_Y + C_2}{\sigma_X^2 + \sigma_Y^2 + C_2} \tag{11}$$

$$s(X,Y) = \frac{\sigma_{XY} + C_3}{\sigma_X\sigma_Y + C_3} \tag{12}$$

$$SSIM(X,Y) = l(X,Y) * c(X,Y) * s(X,Y) \tag{13}$$

where μ_x is the pixel mean of image X, and σ_x is the variance of image X.

5.3. Experiment Results

5.3.1. Experimental Results of NIR Image Colorization Based on Multi-Scale Features

The comparison experiments are first conducted on the generative network U-Net after adding the improved multi-scale feature fusion module. Previously, we added the multi-scale feature fusion module to the generator composed of U-Net, and this improvement enhanced the network's understanding of semantic features and effectively improved the color leakage. As the experimental results in Figure 11 show, the images generated by the multi-scale feature-based method do not have the problem of color leakage, and the colors of the images look natural. The comparison images in this experiment are colored by the CGAN network pix2pix.



(**a**) (**b**) (**c**) (**d**)

Figure 11. Experimental results of NIR image colorization based on multi-scale features. (**a**) Real image; (**b**) NIR image; (**c**) base; and (**d**) multi-scale feature fusion.

The image obtained from the base network in the first group, the sky color leaks during the coloring process, resulting in a blue window color in the red frame; adding

multi-scale feature fusion network on the image coloring looks very natural, and there is no color spillover situation. In addition, compared with the coloring of the baseline network, the outline of the vehicle in front of the house is clearer and the color contrast has been improved very well. The second group of images colored by the baseline network shows that the house in the red box is very blurred due to the shadow, and it is not even possible to distinguish the object as a house. The outline of the house after being colored with the improved network of multi-scale feature fusion is very clear, and the shadow color is also clearly distinguished from the near part, which shows that the improved network is better than the previous network for semantic extraction. The third group of images shows that the images colored with multi-scale feature fusion are clearer, and the boundaries of trees and fine branches are clearly represented. However, the images colored with baseline have blurred contours, white color, and blue leakage. From the above comparison, although the network with the multi-scale feature fusion module still cannot fully restore the color of the real image, it obviously improves the problem of color leakage that exists in other methods and makes the contour boundary of the object clearer.

5.3.2. Experimental Results of NIR Image Colorization Based on Attention Mechanism and Multi-Scale Features

(1) Qualitative analysis

The experimental objects are the following sets of images, where (a) are real images, (b) are NIR images, (c) are images colored with the Pix2PixCGAN network, (d) are the coloring effects of adding channel attention and spatial attention modules to the discriminator, and (e) are the coloring effects of the network after adding multi-scale feature fusion to the generator and improving the discriminator with the attention mechanism at the same time.

As can be seen from the Figure 12, the image (c) in the first group has a lot of noise in the red box part, and the contour and edge information of the image is not clear enough, in addition to the obvious color leakage at the edges of the columns. However, in the images (d) and (e) obtained by the improved network, the image clarity is improved and the edges are more obvious. After adding the attention mechanism, (d) has a deeper outline than (c), and the situation where the tip of the tower is not well distinguished from the background has been improved, and the overall color of the building theme is more harmonious, without the problem of the bluish color caused by color leakage. After improving both the generator and the discriminator, (e) has better results than before, the distinction between the building and the background is more obvious, the building color is closer to the real color, the overall color is more consistent than the first two methods of coloring, which effectively improves the blurred parts of the image, and the overall style is not coordinated. It can be seen from the second group of images in figure (c) that, color leakage in the background of the red box section caused color distortion in the tower after colorization, so that the top of the tower has some blue. And due to the influence of the semantic being unclear when coloring, the tower and the sky border is relatively blurred. The difference between the object and the background in (d) and (e) is more obvious and the contrast is a bit greater. The objects in the figure have more detailed information and more obvious features. The infrared image is able to show the objects that the visible image cannot present due to shadow exposure and other reasons, and the object outline is clearer after colorization of the NIR image by the method in this paper, as shown in the red box in (e), and it gives the color that is consistent with the real image. After the experiments, it was proved that the improved network in this paper has improved both color leakage and semantic ambiguity, and the colorized images have good results visible to the naked eye.

(2) Quantitative analysis

The experimental results are evaluated using the two previously mentioned metrics of peak signal-to-noise ratio and structural similarity. The PSNR focuses on the distortion level of the color image obtained after using network coloring, and its value is negatively correlated with the distortion level, with larger values indicating that the distortion level is lower and the image obtained is more closely matched to the real image. SSIM measures

the similarity between the generated image and the real image from three aspects: brightness, contrast, and structure, which is calculated and evaluated based on the structural information of the image and is more in line with the subjective human visual perception. Accordingly, the closer the value of SSIM is to 1, the more consistent the two image structures are, and the more similar the image after network coloring is to the real one, which proves that the network is more effective.

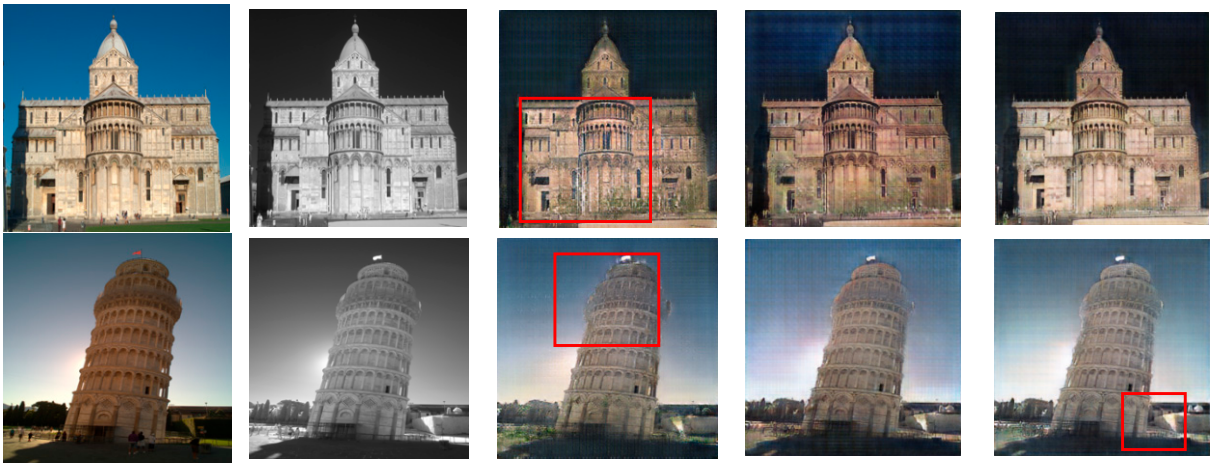


Figure 12. Experimental results of NIR image colorization based on attention mechanism and multi-scale features. (**a**) Real image; (**b**) NIR image; (**c**) pix2pixCGAN; (**d**) attention mechanism; and (**e**) attention + multi-scale feature fusion.

From Table 4, it can be observed that both the generator-improved network and the discriminator-improved network show some improvement in the above-mentioned two metrics after colorizing NIR images. This suggests that the corresponding issues have been addressed to some extent by the improved networks. The network that adds both attention and multi-scale feature fusion works best, and both metrics have higher values. The experiments show that, both in terms of subjective human perception and objective evaluation metrics, the method used significantly improves the color leakage and semantic uncertainty of the original network for infrared coloring.

Table 4. Comparison of evaluation indexes for colorization of NIR images.

	PSNR (dB)	SSIM
Pix2pixCGAN	15.6913	0.5633
Add attention mechanism	15.9534	0.6111
Multi-scale feature fusion	15.9724	0.6161
Add attention and multi-scale feature fusion	16.5342	0.6385

6. Discussion

In this paper, we proposed a method to improve the infrared image colorization problem by using a CGAN network that integrates multi-scale features into the generator and adds an attention mechanism to the discriminator. Prior literature has also attempted to improve networks using multi-scale feature modules and attention mechanisms [22,23,31]. For example, Ref. [31] proposed a multi-scale residual attention network (MsRAN), while Ref. [22] integrated multi-scale attention mechanisms into the generator and discriminator of a GAN to fuse infrared and visible light images (AttentionFGAN) and [23] proposed a deep network that concatenates feature learning modules and fusion learning mechanisms for infrared and visible light image fusion.

The main improvement direction for infrared image colorization is to focus the network's attention on the most important areas of the image and retain more texture infor-

mation. We also used multi-scale feature modules and attention mechanism modules but, unlike previous studies, we chose to improve the CGAN network. The CGAN network is more suitable for tasks that require generating images with specified conditional information than the GAN network. Infrared image colorization tasks require generating color images, and images generated solely by the GAN generator may have issues such as unnatural colors, blurriness, and distortion. The CGAN generator can generate corresponding color images based on the input infrared images, and the input infrared images as conditional information can help the generator produce more accurate corresponding color images, thus solving the above issues. We then added the multi-scale feature module and attention mechanism to the generator and discriminator separately, rather than adding both modules to the generator or discriminator simultaneously. Our goal was to use the game theory of GAN networks, allowing the generator and discriminator networks to compete and promote each other using different methods.

We selected a dataset that included many images with texture details, such as buildings with tightly arranged windows and dense trees. However, due to our focus on making the edges of the subject object smaller and clearer to solve color leakage and edge blur problems, there may be some deviation in the background color of the sky in some images. In future work, we will consider pre-processing the images before inputting them into the CGAN network to enhance the image quality and color restoration ability. After generating color images, post-processing can also be applied to the output images, such as denoising, smoothing, and enhancing contrast, to improve the quality and realism of the output images. The quality and quantity of the dataset are also crucial for the effectiveness of infrared image colorization. Therefore, future research can try to collect more high-quality infrared image datasets and conduct more in-depth studies based on the dataset.

7. Conclusions

In our study on infrared image colorization, we have identified issues with existing networks such as color leakage and semantic ambiguity. In this paper, we proposed a solution to address these issues by improving the generator and discriminator positions. Specifically, we added a multi-scale feature fusion module after the bottleneck layer of the U-Net generator network, which enhances the network's understanding of features and improves its semantic recognition capabilities. For the discriminator, we added an attention mechanism module to the network, allowing it to focus more on useful features and improve its recognition capabilities. We also paid attention to dataset selection and chose scene datasets that are most likely to be applied in actual infrared image scenarios, and near-infrared image datasets that retain more detailed features for testing. Through comparative experiments, we found that our proposed network with attention and multi-scale feature fusion achieved a 5% and 13% improvement in PSNR and SSIM, respectively, compared to the Pix2pixCGAN network, demonstrating that our improvements have effectively solved the problem of color leakage and semantic ambiguity in the original network. Our study focuses solely on infrared images, but future research could consider infrared videos, which require greater attention to the continuity between frames to avoid color jumping.

Author Contributions: Conceptualization, Y.A.; Methodology, Y.A.; Software, H.Z.; Validation, Y.A. and H.Z.; Formal analysis, X.L.; Investigation, J.L.; Resources, S.L.; Data curation, H.A.; Writing—original draft, X.L.; Writing—review & editing, Y.A.; Visualization, H.Z.; Supervision, W.Z.; Project administration, W.Z.; Funding acquisition, W.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the key science and technology project of HBIS Materials Institute (No. HG2022328), and the Innovation Group Project of Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai) (No. 311021013).

Data Availability Statement: The datasets generated during or analyzed during the current study are available from the corresponding author on reasonable request. The public dataset used in this paper is from [29] Brown M, S Süsstrunk. Multi-spectral SIFT for scene category recognition [C]//IEEE Conference on Computer Vision & Pattern Recognition. IEEE, http://dx.doi.org/10.1109/CVPR.2011.5995637 (accessed on 1 January 2023), 2011.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Shi, M.; Zhang, J.Q.; Chen, S.Y.; Gao, L.; Lai, Y.K.; Zhang, F.L. Deep Line Art Video Colorization with a Few References. *arXiv* **2020**, arXiv:2003.106852020. [CrossRef] [PubMed]
2. Wu, M.; Jin, X.; Jiang, Q.; Lee, S.J.; Liang, W.; Lin, G.; Yao, S. Remote sensing image colorization using symmetrical multi-scale DCGAN in YUV color space. *Vis. Comput.* **2020**, *37*, 1707–1729. [CrossRef]
3. Zheng, C.Y.; Fu, Y.; Zhao, Z.; Wang, C.; Nie, J. Imbalance Satellite Image Colorization with Semantic Salience Priors. In Proceedings of the Twelfth International Conference on Graphics and Image Processing (ICGIP 2020), Xi'an, China, 13–15 November 2020; Volume 11720.
4. Khan, M.; Gotoh, Y.; Nida, N. Medical image colorization for better visualization and segmentation. In Proceedings of the Annual Conference on Medical Image Understanding and Analysis, MIUA 2017, Edinburgh, UK, 11–13 July 2017; Springer: Cham, Switzerland, 2017.
5. Golyadkin, M.; Makarov, I. Semi-automatic Manga Colorization Using Conditional Adversarial Networks. In Proceedings of the Analysis of Images, Social Networks and Texts: 9th International Conference, AIST 2020, Moscow, Russia, 15–16 October 2020.
6. Suarez, P.L.; Sappa, A.D.; Vintimilla, B.X.; Hammoud, R.I. Near InfraRed Imagery Colorization. In Proceedings of the 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018.
7. Zhang, C.; Zhao, Y.Y.; Yan, T.Y.; Bai, X.; Xiao, Q.; Gao, P.; Li, M.; Huang, W.; Bao, Y.; He, Y.; et al. Application of near-infrared hyperspectral imaging for variety identification of coated maize kernels with deep learning. *Infrared Phys. Technol.* **2020**, *111*, 1350–4495. [CrossRef]
8. Zhang, L.; Hui, J.; Zhi, B. Application of medical infrared thermal imaging in the diagnosis of human internal focus-ScienceDirect. *Infrared Phys. Technol.* **2019**, *101*, 127–132. [CrossRef]
9. Xing, Z.; He, Y. Many-objective multilevel thresholding image segmentation for infrared images of power equipment with boost marine predators algorithm. *Appl. Soft Comput.* **2021**, *113*, 107905. [CrossRef]
10. Cheng, L.; Tong, Z.; Xie, S.; Kersemans, M. IRT-GAN: A generative adversarial network with a multi-headed fusion strategy for automated defect detection in composites using infrared thermography. *Compos. Struct.* **2022**, *290*, 115543. [CrossRef]
11. Cao, Z.; Kong, X.; Zhu, Q.; Cao, S.; Peng, Z. Infrared dim target detection via mode-k1k2 extension tensor tubal rank under complex ocean environment. *ISPRS J. Photogramm. Remote Sens.* **2021**, *181*, 167–190. [CrossRef]
12. Gui, T.; Gu, X.; Shi, Y. Gray-scale Image Colorization based on Conditional Deep Convolution Generation Adversarial Network. *Int. Core J. Eng.* **2021**, *7*, 116–122.
13. Zhuo, L.; Tan, S.Q.; Li, B.; Huang, J.W. ISP-GAN: Inception sub-pixel deconvolution-based lightweight GANs for colorization. *Multimed. Tools Appl.* **2022**, *81*, 24977–24994. [CrossRef]
14. Huang, S.; Jin, X.; Jiang, Q.; Li, J.; Lee, S.J.; Wang, P.; Yao, S. A fully-automatic image colorization scheme using improved CycleGAN with skip connections. *Multimed. Tools Appl.* **2021**, *80*, 1–28. [CrossRef]
15. Xiao, Y.; Jiang, A.; Liu, C.; Wang, M. Semantic-aware automatic image colorization via unpaired cycle-consistent self-supervised network. *Int. J. Intell. Syst.* **2021**, *37*, 1222–1238. [CrossRef]
16. Berg, A.; Ahlberg, J.; Felsberg, M. Generating Visible Spectrum Images from Thermal Infrared. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018.
17. Nyberg, A.; Eldesokey, A.; Bergstrom, D.; Gustafsson, D. Unpaired Thermal to Visible Spectrum Transfer using Adversarial Training. In Proceedings of the Multimodal Learning and Applications Workshop (MULA)-ECCV Workshop, Munich, Germany, 9 September 2018.
18. Zhong, X.; Lu, T.; Huang, W.; Ye, M.; Jia, X.; Lin, C.W. Grayscale Enhancement Colorization Network for Visible-infrared Person Re-identification. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 1418–1430. [CrossRef]
19. Liang, W.; Ding, D.; Wei, G. An improved DualGAN for near-infrared image colorization. *Infrared Phys. Technol.* **2021**, *116*, 103764. [CrossRef]
20. Xu, J.; Lu, K.; Shi, X.; Qin, S.; Wang, H.; Ma, J. A DenseUnet Generative Adversarial Network for Near-Infrared Face Image Colorization. *Signal Process.* **2021**, *183*, 108007. [CrossRef]
21. Luo, F.; Li, Y.; Zeng, G.; Peng, P.; Wang, G.; Li, Y. Thermal Infrared Image Colorization for Nighttime Driving Scenes with Top-Down Guided Attention. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 15808–15823. [CrossRef]
22. Li, J.; Huo, H.; Li, C.; Wang, R.; Feng, Q. AttentionFGAN: Infrared and Visible Image Fusion Using Attention-Based Generative Adversarial Networks. *IEEE Trans. Multimed.* **2021**, *23*, 1383–1396. [CrossRef]

23. Liu, J.; Fan, X.; Jiang, J.; Liu, R.; Luo, Z. Learning a Deep Multi-Scale Feature Ensemble and an Edge-Attention Guidance for Image Fusion. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 105–119. [CrossRef]
24. Mirza, M.; Osindero, S. Conditional Generative Adversarial Nets. *Comput. Sci.* **2014**, 2672–2680. [CrossRef]
25. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *arXiv* **2014**, arXiv:1406.2661. [CrossRef]
26. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015.
27. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
28. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; Springer: Cham, Switzerland, 2018.
29. Brown, M.; Süsstrunk, S. Multi-spectral SIFT for scene category recognition. In Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition, Colorado Springs, CO, USA, 21–23 June 2011. [CrossRef]
30. An, Z.; Yu, M.; Zhang, C. An improved Adam Algorithm using look-ahead. In Proceedings of the 2017 International Conference, Venice, Italy, 22–29 October 2017.
31. Wang, J.; Yu, L.; Tian, S. MsRAN: A multi-scale residual attention network for multi-model image fusion. *Med. Biol. Engl. Comput.* **2022**, *60*, 3615–3634. [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.