remote sensing

MDPI

Technical Note

An Unpaired Thermal Infrared Image Translation Method Using GMA-CycleGAN

Shihao Yang, Min Sun *, Xiayin Lou, Hanjun Yang and Hang Zhou

Institute of Remote Sensing and GIS, Peking University, Beijing 100871, China; yangshihao@stu.pku.edu.cn (S.Y.); xiayin_lou@stu.pku.edu.cn (X.L.); hanjuny@stu.pku.edu.cn (H.Y.); hang.zhou@pku.edu.cn (H.Z.)

* Correspondence: sunmin@pku.edu.cn

Abstract: Automatically translating chromaticity-free thermal infrared (TIR) images into realistic color visible (CV) images is of great significance for autonomous vehicles, emergency rescue, robot navigation, nighttime video surveillance, and many other fields. Most recent designs use end-to-end neural networks to translate TIR directly to CV; however, compared to these networks, TIR has low contrast and an unclear texture for CV translation. Thus, directly translating the TIR temperature value of only one channel to the RGB color value of three channels without adding additional constraints or semantic information does not handle the one-to-three mapping problem between different domains in a good way, causing the translated CV images not only to have blurred edges but also color confusion. As for the methodology of the work, considering that in the translation from TIR to CV the most important process is to map information from the temperature domain into the color domain, an improved CycleGAN (GMA-CycleGAN) is proposed in this work in order to translate TIR images to grayscale visible (GV) images. Although the two domains have different properties, the numerical mapping is one-to-one, which reduces the color confusion caused by one-to-three mapping when translating TIR to CV. Then, a GV-CV translation network is applied to obtain CV images. Since the process of decomposing GV images into CV images is carried out in the same domain, edge blurring can be avoided. To enhance the boundary gradient between the object (pedestrian and vehicle) and the background, a mask attention module based on the TIR temperature mask and the CV semantic mask is designed without increasing the network parameters, and it is added to the feature encoding and decoding convolution layers of the CycleGAN generator. Moreover, a perceptual loss term is applied to the original CycleGAN loss function to bring the translated images closer to the real images regarding the space feature. In order to verify the effectiveness of the proposed method, the FLIR dataset is used for experiments, and the obtained results show that, compared to the state-of-the-art model, the subjective quality of the translated CV images obtained by the proposed method is better, as the objective evaluation metric FID (Fréchet inception distance) is reduced by 2.42 and the PSNR (peak signal-to-noise ratio) is improved by 1.43.

Keywords: thermal infrared image; image translation; CycleGAN; temperature information; semantic mask

Citation: Yang, S.; Sun, M.; Lou, X.; Yang, H.; Zhou, H. An Unpaired Thermal Infrared Image Translation Method Using GMA-CycleGAN. *Remote Sens.* **2023**, *15*, 663. https://doi.org/10.3390/rs15030663

Academic Editors: Maha Driss, Anis Koubaa, Akrem Sellami, Imed Riadh Farah and Wadii Boulila

Received: 17 December 2022
Revised: 19 January 2023
Accepted: 20 January 2023
Published: 22 January 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

A thermal infrared (TIR) camera captures infrared radiation emitted by objects in scenes as the visible spectrum images do not have ideal color and texture, such as night or low-light working environments where TIR images have a strong advantage; thus, in recent years, TIR cameras have been widely used in industrial surveillance and drones. However, such images do not have color information, which makes it difficult for humans to distinguish objects in the scene and affects their use in some important contexts, such as emergency rescue environments. In order to improve human eye recognition and

computer intelligent processing, many researchers are studying how to translate TIR images into color visible (CV) images for computer vision tasks such as object tracking, crowd counting, panoramic segmentation, and image fusion.

Furthermore, TIR relies on converting temperature into an image, so there is only one active channel measuring mainly the temperature information; however, CV relies on the conversion of colors into images, so there are three channels, which we usually call RGB, and the information carried by TIR and CV is not in the same domain, such that TIR to CV translation is a one-to-three value mapping, including the translation of texture and color. Moreover, due to the limitations of imaging mechanisms and camera manufacturing processes, TIR images have limited resolution and a less prominent texture; these huge differences between image modalities develop challenges to the design of image translation models [1]. Most deep learning (DL)-based image translation algorithms use an end-to-end neural network that directly translates single-channel TIR images to three-channel CV images. As a result, when the CV content is simple and the TIR texture is relatively rich (such as face TIR images), the translation effect is better [2]; however, when the CV scene is complex (such as the street view or the natural landscape), the translated images tend to have large areas of color confusion and texture anomalies.

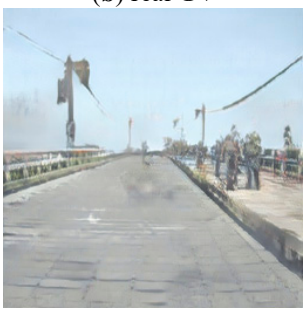
The translation from TIR to CV has a one-to-many mapping relationship between the different domain values, the most important process of which is to translate the information of the temperature domain into the color domain. In order to reduce the ambiguity of the translation process, the translation of different domain values is only considered a one-to-one mapping relationship learning, i.e., first translating the TIR images to grayscale visible (GV) images. Existing translation algorithms, such as QS-Attn [3], decompose one temperature value into three RGB values, causing the translated image to show blurred edges and color confusion, as shown in Figure 1e, which represents the ambiguity in the mapping from the temperature value to RGB due to a lack of sufficient constraints. This ambiguity can be reduced when the temperature is translated only to grayscale values, as shown in Figure 1c.

(**a**) real TIR

(**b**) real GV

(**c**) translated GV by TIR-GV

(**d**) real CV



(**e**) translated CV by QS-Attn[3]

(**f**) translated CV by GV-CV

Figure 1. Comparison between TIR-GV-CV and TIR-CV.

Although further translation from GV images to CV images is required, there is no blurred edge and color confusion because both effects are translated in the same domain. Some existing research results ([4,5]) show that the only disadvantage of the translation from GV to CV is that it is not easy to restore the real scene color; however, the resulting image can fully meet the visual needs of the human eye, as shown in Figure 1f.

Since this study focuses on TIR to GV translation, an improved CycleGAN, called GMA-CycleGAN (Gray Mask Attention-CycleGAN), is proposed, i.e., a gray image cycle-consistent GAN with mask attention. The mask attention mechanism helps in improving the texture of salient objects (pedestrians and vehicles) to meet the needs of object detection and semantic segmentation in practical applications. Thus, a mask attention module has been proposed. Moreover, this latter does not increase network parameters based on TIR temperature masks and CV semantic masks that separate salient objects from the background. Therefore, in GMA-CycleGAN, the mask attention module is added to feature encoding and the feature decoding convolutional layers of the generator. In addition, a perceptual loss term is added to the original CycleGAN loss function to make the translated image closer to the real image in the feature space.

The subsequent parts of this paper are arranged as follows: first, the relevant research status is explained in Section 2, the improved algorithm that was proposed is explained in Section 3, the experiments and datasets are explained in Section 4, the obtained results are analyzed and discussed in Section 5, and a conclusion concludes this work in Section 6.

2. Related Work

Due to the different imaging principles of thermal and visible light sensors, the temperature domain, where TIR is located, is very different from the color domain, where CV is located, and it is difficult for traditional methods to directly identify the mapping relationship from TIR to CV. Early studies ([6,7]) used the fusion of near-infrared images and TIR to supplement texture information, to obtain a fusion image that approximates gray-scale visible light, and to color the fusion image according to the color distribution of a reference color image, so that the visual realism of the obtained image is poor.

In recent years, DL has been widely used in various computer vision tasks, and the powerful fitting ability of neural networks has engendered certain progress in TIR to CV image translation tasks. Moreover, scholars have proposed many network models based on DL, and these models were divided into two categories: convolutional neural networks (CNNs) models and generative adversarial networks (GANs) models according to whether adversarial training is used. In addition, models based on CNNs, such as the TIR2lab model proposed by Berg et al. [8], are the first end-to-end TIR translation models. They hypothesized that the CNN model, based on the autoencoder structure, could identify the luminance-to-chromaticity mapping relationship of paired TIR and CV, and, for the first time, they used the neural network to directly translate TIR to CV. In order to make the small objects of the translated images have more realistic and richer texture information, Wang et al. [9] proposed an attention-based hierarchical thermal infrared image colorization network (AHTIC-Net), which uses multi-scale network structures to extract the features of objects of different sizes in order to enhance the model's attention to small objects during the training process. In general, the translation model, based on CNNs, has an intuitive structure and a simple network training mode; however, due to the insufficient constraint of the loss function of CNNs on the image translation, the translated CV images have the disadvantages of local detail distortion, low image contrast, and blurred visual effect.

Thus, due to the adversarial training of the generative model and the discriminative model, GANs have better behavior than CNNs when applied to image generation, and they can use the unpaired TIR-CV dataset for training. For instance, Isola et al. [10] propose pix2pix to identify the mapping of the source image to a target image using paired datasets. In order to break through the dataset limitations, Zhu et al. [11] proposed the

CycleGAN, which uses two symmetric GANs to form a closed-loop network where one GAN translates images from the source domain to the target domain and the other GAN translates the target domain image back to the source domain and uses cycle-consistency loss to boost the image after two translations to be identical to the original image. Moreover, Pix2pix and CycleGAN have greatly improved the translation effect between CV images, so that image translation, based on the GAN model, has attracted the attention of many scholars. Subsequently, several have improved GAN, introduced methods such as contrastive learning and attention mechanisms, and proposed a variety of unpaired image translation models [3,12–14]. These GAN-based image translation models achieve good results in tasks such as semantic map to CV translation, super resolution, image inpainting, and style transfer; however, the visual effect is not ideal when they are directly applied to TIR-CV translation. Furthermore, Kuang et al. [15] improved the pix2pix method and proposed a new algorithm, the TIC-CGAN, which used GAN for TIR image translation in traffic scenes for the first time. To make the translated image richer in texture, TIC-CGAN applied a coarse-to-fine generator instead of the pix2pix generator, which led to finer texture features in the target images.

The methods mentioned above, based on CNNs and GANs, use paired TIR-CV datasets for neural network training. This latter translates daytime TIR into daytime CV and translates nighttime TIR into nighttime CV, respectively. Since the high beams of oncoming vehicles in nighttime traffic scenes interfere with RGB imaging, resulting in the visual effect of nighttime CV being inferior to daytime CV, Luo et al. improved the CycleGAN algorithm and proposed PearlGAN [16], which used an unpaired training mode to translate nighttime TIR to daytime CV. Although these improved GAN models improve the realism of the translated images, the generated CV images still generate defects related to unclear texture and color distortion. This is due mainly to the one-to-many correspondence in the TIR to CV translation, which is itself an ill-posed solution [17]. Thus, including the previous CNN models, which directly translate single-channel TIR to three-channel CV, these end-to-end translations cannot handle the one-to-many mapping between the temperature domain and the color domain well, resulting in different degrees of color confusion and edge blurring in the translated image.

In order to reduce the instability of the ill-posed problem-solving process, we propose the decomposition of the TIR to CV translation into a two-phase translation process: the first one consists of translating from TIR to GV, whereas the second one achieves the translation from GV to CV. In the proposed experiment, we use the original CycleGAN as the base model. First, we change the TIR to CV one-to-three channel translation to TIR to GV one-to-one channel translation. Although this process does not intrinsically solve the problem of temperature and color matching, it reduces the uncertainty of temperature-to-color translation, helping to improve the sharpness of image edges and reduce unwanted color noise. Inspired by the spatial attention module of AttentionGAN [18], and in order to better distinguish between salient objects (movable pedestrians and vehicles) in the generated image, we separate the object and background, then the semantic mask and the temperature mask are extracted in CV and TIR, respectively, making the salient objects in the image clearer without increasing the network parameters. Moreover, the use of adversarial loss in image translation tasks tends to produce distorted textures [17]. In order to mitigate this problem, a perceptual loss term is added to the original CycleGAN loss function to encourage the translated image to be more similar to the real image in the feature space, which makes the texture information of the translated image closer to the true GV image. Thus, considering that the grayscale image coloring task is only performed in the temperature domain and does not produce edge blurring and noise [4], the original CycleGAN is directly used for the GV to CV translation.

3. Methods

3.1. The Framework of GMA-CycleGAN

The flowchart of TIR-GV image translation, based on our improved CycleGAN (GMA-CycleGAN), is shown in Figure 2, where A and B present two data domains, namely the TIR and the GV, G_{AB} and G_{BA} are two mask attention-based CycleGAN generators, and D_A and D_B are two CycleGAN discriminators. The first row represents the TIR to GV translation, whereas the second row represents the GV to TIR translation. The unpaired training scheme is used, where the real images A and B are randomly selected from the TIR and GV datasets, respectively. Taking the first row of Figure 2 as an example (the same would have been performed on the second row), the input real TIR image A is translated by G_{AB} to GV, and then the discriminator D_B determines whether the generated GV and the real GV image B are real or fake and calculates the adversarial loss. Since the adversarial loss, calculated by CycleGAN discriminators, will cause some distorted textures in the generated images, we introduce the perceptual loss (pl), based on the VGG-16 feature extractor [19], to calculate the difference between the global features of the generated and the real images. This makes the overall visual effect of the generated images more realistic. In addition, CycleGAN will also input the generated image and the real image A into generator G_{BA}, and calculate the difference between the two translated images and the real image A, namely, determining the cycle consistency loss and identity mapping loss.

Thus, the remainder of this section consists of introducing the temperature mask and the semantic mask in Section 3.2, the improved mask attention-based generator in Section 3.3, and the improved loss function of CycleGAN in Section 3.4.

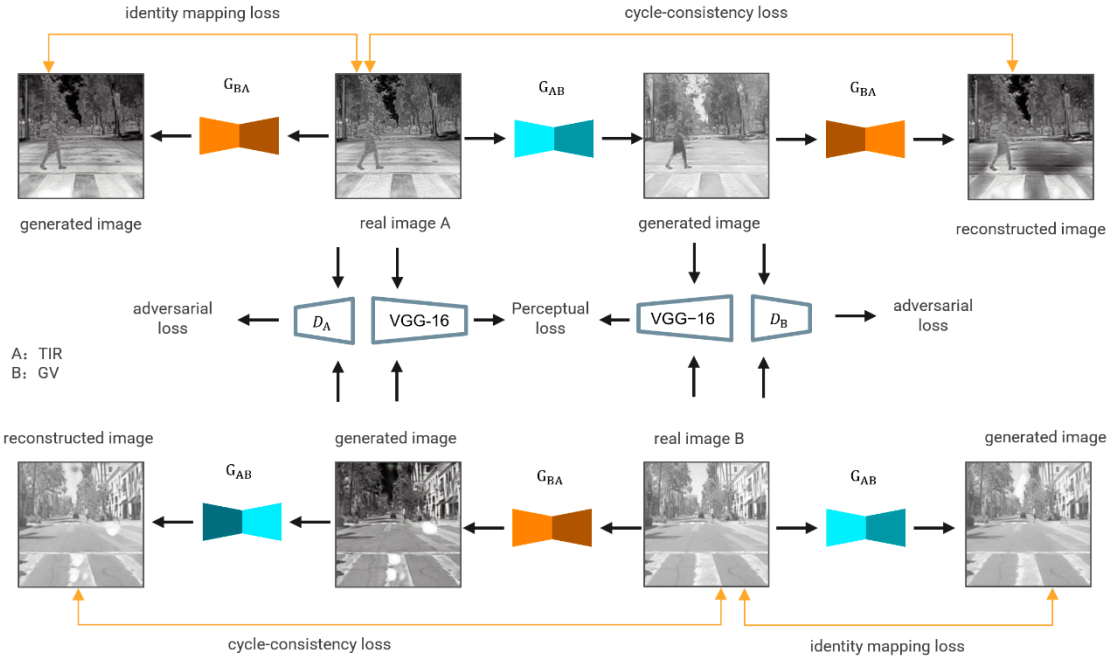


Figure 2. The schematic diagram of GMA-CycleGAN.

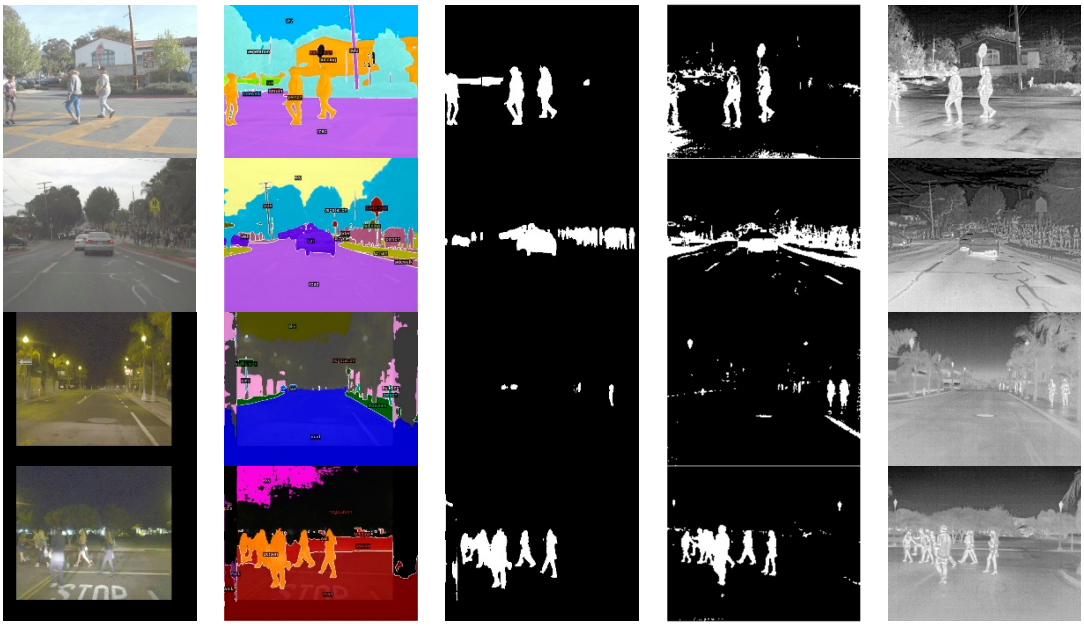
3.2. Temperature Mask and Semantic Mask

In order to better identify the salient objects (movable pedestrians and vehicles) in the translated images and perform downstream tasks, such as object recognition and object tracking, after the completion of the image translation task, the object will be separated from the background in CV and TIR, the semantic mask and the temperature mask will be extracted, and they will be added to the generator as prior knowledge.

Furthermore, we extracted the semantic images from the real CV using Mask2Former [20], a semantic segmentation model pre-trained on the CitySpace dataset, and then

assigned a value of zero to the background (e.g., sky, vegetation, and road) and a value of one to the objects to obtain binary semantic masks, as shown in Figure 3a–c. Among them, the semantic masks of the daytime scene have better effect than the semantic masks of the nighttime scene, as the pre-trained semantic segmentation model will make an incorrect judgment on the object category in the black border of the nighttime CV and it is difficult to distinguish pedestrians in the distance; thus, a more accurate temperature mask is used instead of the semantic mask in the nighttime scene.

Moreover, the raw TIR contains temperature information for the imaging area. Due to the different characteristics of various types of objects absorbing, emitting, and reflecting heat, the specific objects have different manifestations on TIR. Considering this particular property of TIR, the object is separated from the background by setting a pixel threshold. Among them, the human body temperature is relatively stable, and it is the easiest to be separated from the background; moreover, the temperature of the car is higher when the engine is working, and the metal and glass materials on the outer surface of the car are more reflective when the engine is turned off; thus, it is easier to distinguish it. Since the thermal infrared sensor receives both the object's own radiation and the environmental radiation, the influence of solar radiation at night on the thermal imaging is small, and the camera is almost only sensitive to the heat emitted by the object itself, so the segmentation of the human body and the vehicle in the nighttime scene is more accurate than during the day. In addition, as daytime lighting conditions vary, the threshold chosen for segmentation should also vary [21]. Therefore, we used the method proposed in [22] to divide the FLIR dataset into three scenarios: sunny day, cloudy day, and night, and then extract the corresponding salient object temperature threshold windows for the different scenarios.



(**a**) CV (**b**) semantic image (**c**) semantic mask (**d**) temperature mask (**e**) TIR

Figure 3. Temperature masks and semantic masks in different illumination scenarios.

We set the pixel value in the TIR image threshold window to one and the pixel value beyond the threshold window to zero in order to obtain binary temperature masks, as shown in Figure 3d,e. From this figure, pedestrians can be better divided in sunny days, but the car division noise in the distant parking lot is large, and the road surface also has a certain noise. As for the cloudy days, vehicles and pedestrians can also be separated, but there is a certain amount of noise. Finally, at night, people in the distance can be accurately divided with less noise. Since semantic masks are less noisy than temperature masks during the day and vice versa in nighttime scenes with less ambient radiation, we use

semantic masks during the day and temperature masks at night. In general, both temperature masking and semantic masking are a little noisy, but our network model does not rely entirely on masks as it also relies on other feature information of the original CV-TIR image pairs; thus, the mask noise has a small effect.

3.3. Generator Based on Mask Attention

Spatial attention focuses on local information within the spatial domain, that is, the identification of the areas on the feature map that deserve our attention, yielding better network outputs. General spatial attention is calculated using neural networks, which is posterior knowledge, while the temperature mask and the semantic mask, inferred by the threshold window and pre-training model in this study, can be regarded as a type of spatial attention (*i.e.*, mask attention) based on prior knowledge without increasing the amount of network parameters, and can focus on the mask region. The proposed mask attention multiplies the input feature map with the mask on a channel-wise pixel-wise basis, as expressed in Equation (1):

$$y_{c,h,w} = f^{multi}(x, tm) = x_{c,h,w} \times (tm_{1,h,w} \times (1 - \alpha) + \alpha) \quad (1)$$

Where $x_{c,h,w}$ and $y_{c,h,w}$ are the feature maps of the mask attention input and output, respectively, $tm_{1,h,w}$ represents the binary masks whose length and width are equal to the feature maps, and c, h and w represent the number of channels, the height, and the width of the feature maps, respectively. Added to that, α is a parameter for adjusting the attention strength of the mask. Afterwards, we translate the binary mask to weight mask by applying $(tm_{1,h,w} \times (1 - \alpha) + \alpha)$. In order to emphasize the salient object and suppress the background, we set $\alpha < 1$ as this parameter yields a weaker attention when it is close to one and a strong attention when it is close to zero. For feature maps of different spatial sizes, the original mask passes through a pooling layer of size 3×3, stride 2, and padding 1 to keep the length and the width consistent.

The GMA-CycleGAN generator adds mask attention between the convolutional layers, the instance normalization (IN), and the ReLU activation layers of the original CycleGAN encoder and decoder. As shown in Figure 4, among them, the encoder uses three convolutional layers to extract the feature maps of $64 \times 64 \times 256$ from the source image of $256 \times 256 \times 3$, and then the translator translates the $64 \times 64 \times 256$ dimension feature maps extracted in the previous step into the $64 \times 64 \times 256$ dimensional feature maps of the target image through nine residual blocks. Finally, the decoder uses three deconvolution layers to restore the low-level features from the feature maps so as to output a $256 \times 256 \times 3$ image. To maintain the symmetry of the CycleGAN generator, in the encoder, we insert the mask attention module after the initialization convolutional layer (kernel size equal to 7×7 and stride equal to one), and after two down-sample convolutional layers (kernel size is equal to 3×3 and stride is equal to two), other structures are the same as the original CycleGAN, and then the corresponding improvements are added symmetrically to the decoder. Since the CycleGAN model is mirror-symmetric, the mask attention module is placed after the convolution operation when encoding and before the deconvolution operation when the feature is decoded.

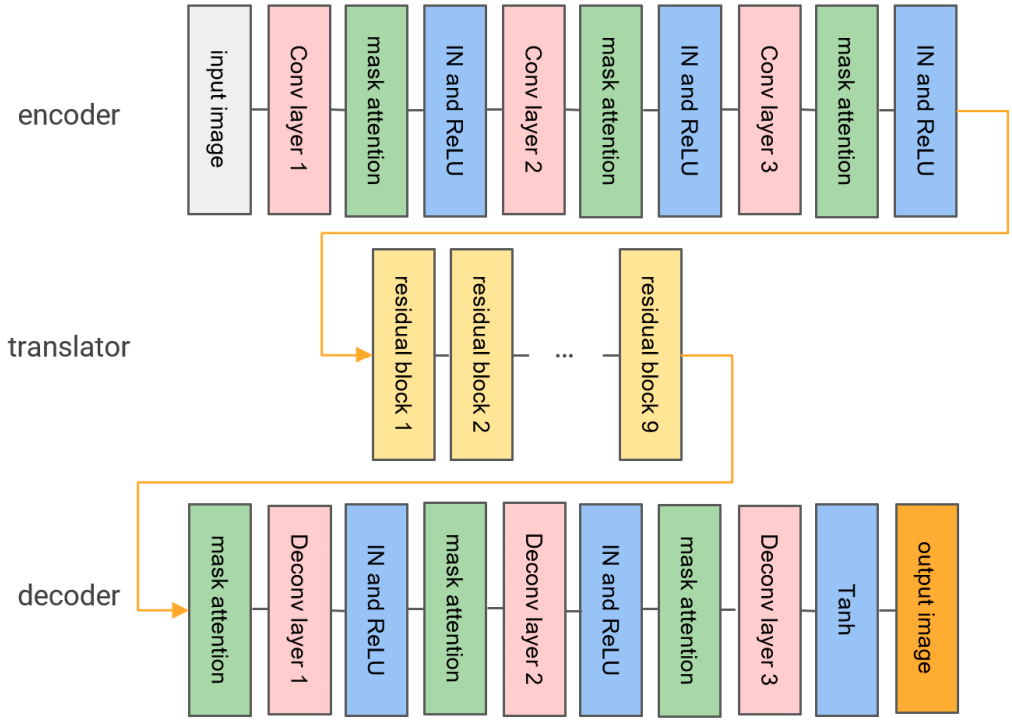


Figure 4. Schematic diagram of the model framework of the GMA-CycleGAN generator based on mask attention.

3.4. Loss Function of GMA-CycleGAN

Let A and B represent the source image domain (TIR domain) and the target image domain (GV domain), respectively, and a and b represent the source image and target image, respectively. There are three loss functions for the original CycleGAN: GAN loss, identity mapping loss, and cycle consistency loss, where GAN loss includes two GAN loss functions, *i.e.*, $\mathcal{L}_{GAN}(G_{AB}, D_B, A, B)$ and $\mathcal{L}_{GAN}(G_{BA}, D_A, B, A)$. Moreover, GAN loss guarantees that the generated sample is distributed the same way as the real sample. The cycle-consistency loss $\mathcal{L}_{cyc}(G_{AB}, G_{BA})$ encourages the sample to remain unchanged after passing through two generators, *i.e.*,$G_{BA}(G_{AB}(a)) \approx a$ and $G_{AB}(G_{BA}(b)) \approx b$. As for the identity mapping loss $\mathcal{L}_{ide}(G_{AB}, G_{BA})$, it guarantees hue associativity between the generated image and the original image.

Since adversarial loss can cause texture distortion in the generated image, we added the perceptual loss item to make the generated image texture more realistic. Thus, perceptual loss calculates the ℓ_2 distance between the feature maps obtained by the convolution of the generated image and the feature maps obtained by the convolution of the real image, so that their high-level semantic information is closer. Referring to [23], we use the pre-trained VGG-16 network on the ImageNet dataset [24] as a feature extractor, and the perceptual loss is expressed in Equation (2):

$$\begin{aligned}\mathcal{L}_{per}(G_{AB}, G_{BA}) = E_{a\sim p_{data}(a)}[\|vgg16(G_{BA}(b)) - vgg16(a)\|_2] \\ +E_{b\sim p_{data}(b)}[\|vgg16(G_{AB}(a)) - vgg16(b)\|_2]\end{aligned} \quad (2)$$

Since the VGG-16 network is pre-trained and its input must be a three-channel image, while TIR and GV are single-channel images, we copy the single channel into a three-channel image and use it as the input of VGG-16.

The final loss function of the model is a weighted combination of each loss, and the formula is expressed as follows:

$$\mathcal{L}(G, F, D_X, D_Y) = \mathcal{L}_{GAN}(G_{AB}, D_B, A, B) + \mathcal{L}_{GAN}(G_{BA}, D_A, B, A) + \lambda_{cyc}\mathcal{L}_{cyc}(G_{AB}, G_{BA}) + \quad (3)$$

$$\mathcal{L}_{ide}(G_{AB}, G_{BA}) + \lambda_{per}\mathcal{L}_{per}(G_{AB}, G_{BA})$$

where λ_{cyc}, λ_{ide}, λ_{per}, λ_{pyr} are the weight coefficients used to adjust the ratio of the cycle-consistency loss, identity mapping loss, and perceptual loss, respectively.

4. Experiments

4.1. Experimental Platform and Dataset

All experiments in this paper were performed using a Dell PowerEdge T640 tower server with 1080Ti GPUs, and all neural networks were trained and validated using the Pytorch framework.

Most thermal infrared visible datasets are used primarily for object detection and tracking. One of the most well-known datasets is KAIST [25], which does not store TIR TIFF images with the original temperature radiation value because its TIR JPEG or PNG format image is obtained by preprocessing, such as affine transformation of TIFF files, and the original temperature radiation value cannot be retrieved. Published in 2018, the FLIR dataset, taken on streets and highways, contains TIRs in 14-bit TIFF format images and is sharper than KAIST; however, 9620 pairs of infrared-visible images from the original FLIR dataset are not aligned, which results in the semantic mask extracted from the dataset being misaligned with the temperature mask. Therefore, we experimented using the aligned FLIR dataset published by Zhang et al. [26], which was selected and aligned from the original FLIR dataset, where 4129 pairs were deployed for training data and 1013 pairs for testing the system. Moreover, TIR and CV, having a resolution of 640×512, and taken by the FLIR Tau2 camera, contain approximately 80% (4130) daytime images and 20% (1012) nighttime images.

4.2. Evaluating Metrics and Parameter Configuration

We evaluate the quality of the translated image from both faithfulness and realism points of view. For faithfulness, we calculate the two most commonly used image distance measurements, including the average of the structural similarity (SSIM) and the peak signal-to-noise ratio (PSNR) metrics for each pair of generated and real images. Higher SSIM and PSNR indicate that the translated image is more similar to the real image. As for realism, we quantitatively calculate the widely used FID (Frechet inception distance) metric to measure the distance between two distributions of real and generated images, and its expression is as follows:

$$FID(g, r) = ||\mu_r - \mu_g||^2 + Tr(\sum_r + \sum_g - 2\sqrt{\sum_r \sum_g}) \quad (4)$$

where μ_r and μ_g represent the mean of the 2048-dimensional feature vectors obtained by importing the real images and the translated images into the Inception v3 model [27], respectively. Moreover, $\sum_r$ and $\sum_g$ represent the covariance matrices of real images and generated images, respectively, and Tr indicates the trace of a matrix. The smaller the FID is, the closer the feature vectors are, and the more realistic the translated image will be.

In temperature mask extraction, since the FLIR dataset did not label the weather conditions, we manually classified the aligned FLIR dataset into three categories: sunny day, cloudy day, and night. For these three types of illumination conditions, according to the parameter settings in [22], we set the corresponding threshold windows as 7500–7700, 7300–7500 and 7200–8000 to extract the temperature masks of the salient objects from TIR TIFF files.

During the training, we scale the input images of resolution 640×512 to 256×256. Using Adam optimizer for backward propagation [28] and setting the Adam momentum parameter $\beta_1 = 0.5$ and $\beta_2 = 0.999$, as well as the batch size to one, the initial learning rate to 0.0002, and the training epochs to 200 (where the learning rate of the

first 100 epochs is unchanged and the learning rate of the last 100 epochs is linearly decayed to 0). For the loss function weights, we follow the experimental parameters presented in [11], and we set the cycle-consistency factor λ_{cyc} to 10 and the identity mapping factor λ_{cyc} to 1. Referring to the experimental parameters proposed in [23], the perceptual loss uses the feature maps of the 3rd, 8th, 15th, and 22nd layers from VGG-16 pre-trained on ImageNet and sets the factor λ_{per} to 5.

5. Results and Discussion

First, when the mask attention is used in different positions of the generator encoder, the FID of the corresponding six models of different mask attention parameters w is calculated as shown in Table 1. Referring to this table, when the mask attention is added to the first two convolutional layers, the FID is the smallest, i.e., the image translation effect of the GMA-CycleGAN_4 model is the best. We believe that the first two convolutional layers extract low-level features of the TIR image, such as edge information, and then the mask attention can guide the network to distinguish between the boundaries of an object and those of a background. As for the third convolutional layer, it extracts more complicated image global features, and using mask attention may destroy the extracted global features. As can be seen from the table, it is better to set the parameter w to 0.6, a smaller w will overly suppress the characteristics of the background area, and a larger w does not distinguish between the background and the object enough.

Table 1. Results of the FID after that the mask attention was inserted into different layers.

Name	Mask Attention			w = 0.2	w = 0.4	w = 0.6	w = 0.8
	Layer 1	Layer 2	Layer 3	FID	FID	FID	FID
GMA-CycleGAN_1	✓			71.68	71.42	71.19	71.67
GMA-CycleGAN_2		✓		71.82	71.58	71.36	71.71
GMA-CycleGAN_3			✓	72.61	72.50	72.03	72.29
GMA-CycleGAN_4	✓	✓		71.05	70.77	**70.62**	70.85
GMA-CycleGAN_5		✓	✓	72.99	72.54	72.06	73.15
GMA-CycleGAN_6	✓		✓	72.46	72.21	71.88	72.13
GMA-CycleGAN_7	✓	✓	✓	72.25	71.76	71.65	72.03

Since few studies have used FLIR datasets for TIR translation, we selected five typical and popular unpaired image translation models to compare with our model, including CycleGAN [11], U-GAT-IT [12], NICE-GAN [13], CUT [14], and QS-attn [3], and their open-source code was implemented for model training and testing. Table 2 displays the quantitative evaluation metrics of each model on the testing set, where our model outperforms the other models across the board. In the typical model, QS-attn has the highest realism indicator and NICE-GAN has the highest faithfulness indicator. Compared to these two state-of-the-art (SOTA) models, our model's FID is reduced by 2.42, the PSNR is increased by 1.43, and the SSIM is basically unchanged. Moreover, our model is twice as fast as CycleGAN in terms of training time but faster than the other models, where QS-attn takes the longest training time of about 9 days.

Table 2. Results of different models.

Name	FID	PSNR	SSIM	Training Time (Day)
cyclegan	84.78	38.44	0.9258	**1**
UGATIT	83.44	38.59	0.9239	6
NICE-GAN	80.23	39.33	0.9309	2
CUT	77.73	39.01	0.9270	3
QSGAN	73.04	38.99	0.9253	9
GMA-CycleGAN_4	**70.62**	**40.76**	**0.9417**	2

To sum up, we provide three nighttime TIR images, four daytime TIR images, and their corresponding translated CV images to subjectively evaluate the different models, as shown in Figure 5. Since all models in Figure 5 were trained with the unpaired dataset, in which daytime CV images accounted for the majority (about 80%), these models can easily translate the night images into the day images; this does mean these models are overfitted for the night image translation; however, the loss curves did not show overfitting during the model training. In fact, in order to obtain better RGB visual effects, we expect to translate all these TIR images into day images regardless they were captured at day or night. In the future work, we consider to supplement extra information to augment the translation effect from night TIR images to day CV images.

Figure 5. Translation results of different models.

Whether it is a daytime scene or a nighttime scene, the translated image of the typical models has different degrees of color chaos, and the proposed model does not have this feature, which shows the superiority that TIR is translated to GV first and then translated from GV to CV. Moreover, while computing the translated images of the typical models, the edges of the sky and the roads are clear, but the edges of the ground scene are more blurred (such as pedestrians and vegetation), while the translated images of our model have less blurred edges. Overall, the proposed model's translated images have a better overall effect, with more realistic textures and better distinction between salient objects (people and vehicles) and backgrounds.

We also explore the impact of each component in ablation studies. Comparison results are shown in Figure 6. As can be seen, removing the perceptual loss leads to distorted details, such as the white circles with blue edges (marked in green rectangles) in the two

rows of Figure 6b and the abnormal red color around the two pedestrians in the second row of Figure 6b. Additionally, removing the mask attention leads to semantic confusion between salient objects and the background, such as the two pedestrians and the pick-up truck in the second row of Figure 6c. Therefore, each component is indispensable for generating high-quality CV images.

(**a**) TIR (**b**) w/o perceptual loss (**c**) w/o mask attention (**d**) GMA-CycleGAN_4 (**e**) CV

Figure 6. Comparisons of different ablation studies for GMA-CycleGAN_4.

Table 3 displays the quantitative evaluation metrics of different ablation studies for GMA-CycleGAN_4. The translated TIR images produced by GMA-CycleGAN_4 with full structures achieve the best quantitative metrics. The lack of perceptual loss and mask attention causes performance degradation, in which omitting perceptual loss reduces the FID, PSNR, and SSIM by 2.66, 1.07, and 0.0086 and omitting mask attention reduces the FID, PSNR, and SSIM by 2.02, 0.92, and 0.0062.

Table 3. Results of different ablation studies for GMA-CycleGAN_4.

Name	FID	PSNR	SSIM
w/o perceptual loss	73.28	39.69	0.9331
w/o mask attention	72.64	39.84	0.9355
GMA-CycleGAN_4 (full)	70.62	40.76	0.9417

6. Conclusions

To solve the problem of color distortion and edge blurring in the images generated by the existing end-to-end TIR to CV translation model, we propose an improved CycleGAN (GMA-CycleGAN) consisting of the translation from TIR to GV first, then using the original CycleGAN to translate from GV to CV. Thus, for temperature domain to color domain translation, the one-to-one mapping relationship is only considered, that is, the TIR to GV translation, which reduces the color ambiguity caused by the different domain translation. We also take the temperature mask of TIR and the semantic mask of CV as prior knowledge to add the edge information of salient objects. In addition, to mitigate texture distortion caused by adversarial loss, perceptual loss is added to the CycleGAN loss function. In terms of objective evaluation, compared to the existing SOTA methods, our model training time is shorter, the FID is reduced by 2.42, and the PSNR is increased by 1.43. In terms of subjective evaluation, experimental results show that the texture and color of the translated image, obtained by our method, are more realistic, and the salient object edge information is richer. The results also validate the effectiveness of the proposed method and indicate its importance for many fields, such as autonomous vehicles, emergency rescue, robot navigation, and nighttime video surveillance.

Further work is threefold. First, in order to apply our method to versatile datasets, extraction of the temperature mask from JPG or PNG format image files is needed. Second,

semantic segmentation of TIR and GV images, which can be used to maintain semantic consistency between real and generated images, is needed to further improve the translated image quality. Third, when using test datasets that are quite different from the aligned FLIR dataset, the generalization ability of our model is not ideal, and we consider solving this problem in further research works.

Author Contributions: Conceptualization, S.Y. and M.S.; methodology, S.Y.; validation, S.Y., M.S., and H.Z.; formal analysis, S.Y.; investigation, S.Y., X.L., and H.Z.; writing—original draft preparation, S.Y.; writing—review and editing, S.Y., M.S., H.Z., X.L., and H.Y.; visualization, S.Y. and H.Y.; supervision, M.S. and X.L.; project administration, M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable

Data Availability Statement: The original FLIR dataset is openly available, access through the link: https://www.flir.com/oem/adas/adas-dataset-form (accessed on 22 Jan 2023). The aligned FLIR dataset was published by Zhang et al. [26] (accessed on 22 Jan 2023).

Acknowledgments: We would like to thank Zhang et al. [26] for the publicly available well-aligned FLIR dataset, which laid the foundation for our experiments.

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

References

1. Hou, F.; Zhang, Y.; Zhou, Y.; Zhang, M.; Lv, B.; Wu, J. Review on Infrared Imaging Technology. *Sustainability* **2022**, *14*, 11161. https://doi.org/10.3390/su141811161.
2. Luo, Y.; Pi, D.; Pan, Y.; Xie, L.; Yu, W.; Liu, Y. ClawGAN: Claw connection-based generative adversarial networks for facial image translation in thermal to RGB visible light. *Expert Syst. Appl.* **2022**, *191*, 116269. https://doi.org/10.1016/j.eswa.2021.116269.
3. Hu, X.; Zhou, X.; Huang, Q.; Shi, Z.; Sun, L.; Li, Q. QS-Attn: Query-Selected Attention for Contrastive Learning in I2I Translation. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–20 June 2022; pp. 18270–18279. https://doi.org/10.1109/CVPR52688.2022.01775.
4. Huang, S.; Jin, X.; Jiang, Q.; Liu, L. Deep learning for image colorization: Current and future prospects. *Eng. Appl. Artif. Intell.* **2022**, *114*, 105006. https://doi.org/10.1016/j.engappai.2022.105006.
5. Liang, W.; Ding, D.; Wei, G. An improved DualGAN for near-infrared image colorization. *Infrared Phys. Technol.* **2021**, *116*, 103764. https://doi.org/10.1016/j.infrared.2021.103764.
6. Toet, A.; Hogervorst, M.A. Portable real-time color night vision. In Proceedings of the SPIE Defense and Security Symposium, Orlando, FL, USA, 17–20 March 2008; p. 697402. https://doi.org/10.1117/12.775405.
7. Hogervorst, M.A.; Toet, A. Fast natural color mapping for night-time imagery. *Inf. Fusion* **2010**, *11*, 69–77, https://doi.org/10.1016/j.inffus.2009.06.005.
8. Berg, A.; Ahlberg, J.; Felsberg, M. Generating Visible Spectrum Images from Thermal Infrared. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018, pp. 1224–122409. https://doi.org/10.1109/CVPRW.2018.00159.
9. Wang, H.; Cheng, C.; Zhang, X.; Sun, H. Towards high-quality thermal infrared image colorization via attention-based hierarchical network. *Neurocomputing* **2022**, *501*, 318–327. https://doi.org/10.1016/j.neucom.2022.06.021.
10. Sola, P.; Zhu, J.-Y.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5967–5976. https://doi.org/10.1109/CVPR.2017.632.
11. Zhu, J.-Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
12. Kim, J.; Kim, M.; Kang, H.; Lee, K. U-GAT-IT: Unsupervised Generative Attentional Networks with Adaptive Layer-Instance Normalization for Image-to-Image Translation. *arXiv* **2020**. Available online: http://arxiv.org/abs/1907.10830 (accessed on 29 November 2022).
13. Chen, R.; Huang, W.; Huang, B.; Sun, F.; Fang, B. Reusing Discriminators for Encoding: Towards Unsupervised Image-to-Image Translation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 8165–8174. https://doi.org/10.1109/CVPR42600.2020.00819.

14. Park, T.; Efros, A.A.; Zhang, R.; Zhu, J.-Y. Contrastive Learning for Unpaired Image-to-Image Translation. In *Computer Vision—ECCV 2020*; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; Volume 12354, pp. 319–345. https://doi.org/10.1007/978-3-030-58545-7_19.
15. Kuang, X.; Zhu, J.; Sui, X.; Liu, Y.; Liu, C.; Chen, Q.; Gu, G. Thermal infrared colorization via conditional generative adversarial network. *Infrared Phys. Technol.* **2020**, *107*, 103338, https://doi.org/10.1016/j.infrared.2020.103338.
16. 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. https://doi.org/10.1109/tits.2022.3145476.
17. Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Tao, A.; Kautz, J.; Catanzaro, B. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8798–8807.
18. Tang, H.; Liu, H.; Xu, D.; Torr, P.H.S.; Sebe, N. AttentionGAN: Unpaired Image-to-Image Translation Using Attention-Guided Generative Adversarial Networks. *IEEE Trans. Neural. Networks Learn. Syst.* **2021**, *11*, 1–16. https://doi.org/10.1109/tnnls.2021.3105725.
19. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**. Available online: http://arxiv.org/abs/1409.155 (accessed on 4 December 2022).
20. Cheng, B.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-attention Mask Transformer for Universal Image Segmentation. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 1280–1289. https://doi.org/10.1109/CVPR52688.2022.00135.
21. Nikolov, I.A.; Philipsen, M.P.; Liu, J.; Dueholm, J.V.; Johansen, A.S.; Nasrollahi, K.; Moeslund, T.B. Seasons in Drift: A Long Term Thermal Imaging Dataset for Studying Concept Drift. In Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, Montreal, Canada 2021. Available online: https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/c45147dee729311ef5b5c3003946c48f-Paper-round2.pdf (accessed on 22 January 2023).
22. Zhou, H.; Sun, M.; Ren, X.; Wang, X. Visible-Thermal Image Object Detection via the Combination of Illumination Conditions and Temperature Information. *Remote. Sens.* **2021**, *13*, 3656. https://doi.org/10.3390/rs13183656.
23. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *Computer Vision—ECCV 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; Volume 9906, pp. 694–711. https://doi.org/10.1007/978-3-319-46475-6_43.
24. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. https://doi.org/10.1007/s11263-015-0816-y.
25. Hwang, S.; Park, J.; Kim, N.; Choi, Y.; So Kweon, I. Multispectral pedestrian detection: Benchmark dataset and baseline. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1037–1045.
26. Zhang, H.; Fromont, E.; Lefevre, S.; Avignon, B. Multispectral Fusion for Object Detection with Cyclic Fuse-and-Refine Blocks. In Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 25–28 October 2020; pp. 276–280. https://doi.org/10.1109/ICIP40778.2020.9191080..
27. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826. https://doi.org/10.1109/cvpr.2016.308.
28. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization, in *ICLR (Poster*). 2015. Available online: http://arxiv.org/abs/1412.6980 (accessed on 22 January 2023).

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.