electronics

MDPI

Article

Dual-Branch Colorization Network for Unpaired Infrared Images Based on High-Level Semantic Features and Multiscale Residual Attention

Tong Jiang [1], Junqi Bai [2], Lin Xiao [3], Tingting Liu [1], Xiaodong Kuang [4], Yuan Liu [1,*], Xiubao Sui [1] and Qian Chen [1,5]

[1] School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China; jiangtongff@njust.edu.cn (T.J.); tingtingliu@njust.edu.cn (T.L.); sxb@njust.edu.cn (X.S.); chenq@mail.njust.edu.cn (Q.C.)
[2] The 28th Institute of China Electronics Technology Group Corporation, Nanjing 210007, China; baijunqi@cetc.com.cn
[3] China Academy of Aerospace Science and Innovation, Beijing 100871, China; xiaolin_82@163.com
[4] Zhejiang Lab, Hangzhou 311112, China; kuangxiaodong@zhejianglab.com
[5] School of Instruments and Electronics, North University of China, Taiyuan 030051, China
* Correspondence: liu_yuan_eo@163.com

Abstract: The infrared image colorization technique overcomes the limitation of grayscale characteristics of infrared images and achieves cross-modal conversion between infrared and visible images. Aiming at the problem of lack of infrared-visible pairing data, existing studies usually adopt unsupervised learning methods based on contrastive loss. Due to significant differences between modalities, reliance on contrastive loss alone hampers the learning of accurate semantic features. In this paper, we propose DC-Net, which is a dual-branch contrastive learning network that combines perceptual features and multiscale residual attention for the unsupervised cross-modal transformation of infrared to visible images. The network comprises a patch-wise contrastive guidance branch (PwCGB) and a perceptual contrastive guidance branch (PCGB). PwCGB focuses on discerning feature similarities and variances across image patches, synergizing patch-wise contrastive loss with adversarial loss to adaptively learn local structure and texture. In addition, we design a multiscale residual attention generator to capture richer features and adaptively integrate multiscale information. PCGB introduces a novel perceptual contrastive loss that uses perceptual features from pre-trained VGG16 models as positive and negative samples. This helps the network align colorized infrared images with visible images in the high-level feature space, improving the semantic accuracy of the colorized infrared images. Our unsupervised infrared image colorization method achieves a PSNR of 16.833 and an SSIM of 0.584 on the thermal infrared dataset and a PSNR of 18.828 and an SSIM of 0.685 on the near-infrared dataset. Compared to existing algorithms, it demonstrates substantial improvements across all metrics, validating its effectiveness.

Keywords: infrared image colorization; cross-modal conversion; semantic features; multiscale residual attention

check for updates

Citation: Jiang, T.; Bai, J.; Xiao, L.; Liu, T.; Kuang, X.; Liu, Y.; Sui, X.; Chen, Q. Dual-Branch Colorization Network for Unpaired Infrared Images Based on High-Level Semantic Features and Multiscale Residual Attention. *Electronics* **2024**, *13*, 3784. https://doi.org/10.3390/electronics13183784

Academic Editor: Dimitris Kanellopoulos

Received: 5 September 2024
Revised: 18 September 2024
Accepted: 20 September 2024
Published: 23 September 2024

1. Introduction

Infrared detectors operate effectively both day and night, particularly excelling in low-light conditions by capturing the thermal emissions from objects. In contrast to full-color visible images, single-channel infrared imagery lacks color and textural detail, diverging from human visual perception. Translating infrared to the visible spectrum can imbue infrared images with color, enhancing human perception under all lighting conditions [1]. However, bridging the gap between these spectral domains to generate semantically rich color images remains challenging. Deep learning approaches, particularly convolutional neural networks (CNNs), have advanced the colorization of infrared images [2]. While supervised methods like conditional generative adversarial networks (cGANs) [3] show promise in this domain, they typically depend on extensive paired datasets, which are impractical to collect in real-world settings due to high costs and time constraints. Learning to map between unpaired infrared and visible images thus presents a valuable alternative.

Unsupervised infrared image colorization, which aims to map grayscale thermal images into the visible spectrum without matched references, is an emerging focus in imaging research. Traditional methods, relying on preset knowledge or multispectral fusion, often fall short due to rigid color translation rules and a lack of flexibility [4–7]. Recent developments in deep learning, with architectures like CycleGAN [8], CUT [9], and UNIT [10], have made strides in unpaired image-to-image translation, offering innovative strategies for cross-domain colorization [11–13]. Despite these advancements, significant challenges persist. Infrared and visible imagery differ fundamentally in color space and information content, complicating the task of inferring accurate semantic colors from monochromatic data without labeled guidance. Current unsupervised methods struggle with semantic interpretation, hindering their ability to establish complex mappings between domains. Consequently, the colorization results often lack naturalness and semantic precision. Furthermore, a distinct domain difference exists between infrared and visible images. Infrared and visible domains correspond to different wavelength ranges, directly impacting their imaging mechanisms and the information content they carry. While visible imaging usually relies on the reflection of an external light source, thermal infrared imaging is mainly based on the thermal radiation of the object itself. Consequently, the same object exhibits entirely different visual characteristics at different wavelengths. For example, in the visible domain, an object may exhibit vibrant colors, while in the infrared domain, it shows a different temperature distribution. While visible images can provide rich color and texture information, images in the thermal infrared domain mainly reflect temperature distributions. Single-channel infrared grayscale images not only lack semantic color information but also suffer from structural blurring and lack of texture information [14,15], which increases the difficulty of cross-domain feature learning. In summary, existing unsupervised image translation methods have certain limitations in processing infrared images. They are difficult to accurately predict the chromaticity information and structural texture information that matches the semantic information from the grayscale information of infrared images, resulting in the generation of colorized infrared images that do not meet the standards of human visual perception.

To solve the above problems, we propose a dual-branch structure colorization network for infrared images. Unsupervised infrared image to visible image conversion is realized based on perceptual features and multiple contrastive learning. The generator of the infrared image colorization network is improved by combining multiscale residual blocks with an attention mechanism. Our unsupervised infrared image colorization method aims to effectively capture image features at different scales, preserve and enhance image detail information, and generate colors that are consistent with human visual perception. We note that the patchNCE loss learns patch-level similarity features by maximizing the mutual information between image patches in the input and output images, while the generative adversarial loss digs the overall image-level similarity features by playing the mutual game between the generated image and the target domain image. In order to learn multi-level feature information, we combine contrastive learning with the generative adversarial network framework in the patch-wise contrastive guidance branch (PwCGB) and design a composite loss function. In addition, considering the importance of semantic information for the infrared image-visible image conversion task, we propose perceptual contrastive loss. Inspired by the perceptual loss, we extract the high-dimensional features of infrared images, generated colorized images, and visible images in the perceptual contrastive guidance branch (PCGB) by a pre-trained VGG16 network, respectively. By minimizing the representation distance of positive sample pairs and maximizing the representation distance of negative sample pairs, the similarity and difference between the infrared and visible domains in the feature space are optimized. The strategy of perceptual contrastive learning combined with high-dimensional feature extraction helps to improve the quality of colorization of infrared images and makes the semantic colors of the generated colorized images more accurate. In order to further improve the quality of colorized infrared images and enhance the feature extraction capability of the generator network in the colorization task, our designed multiscale residual block is added in the down-sampling stage of the generator. Parallel residual blocks with different convolutional kernel sizes are utilized to acquire feature information at different scales. In addition, we combine channel attention and residual connectivity by designing feature fusion attention residual blocks in the up-sampling stage of the generator. Through the attention mechanism, the multiscale information can be better integrated, which enables the low-level features and high-level features to be combined more efficiently during the up-sampling process and improves the accuracy of colorization.

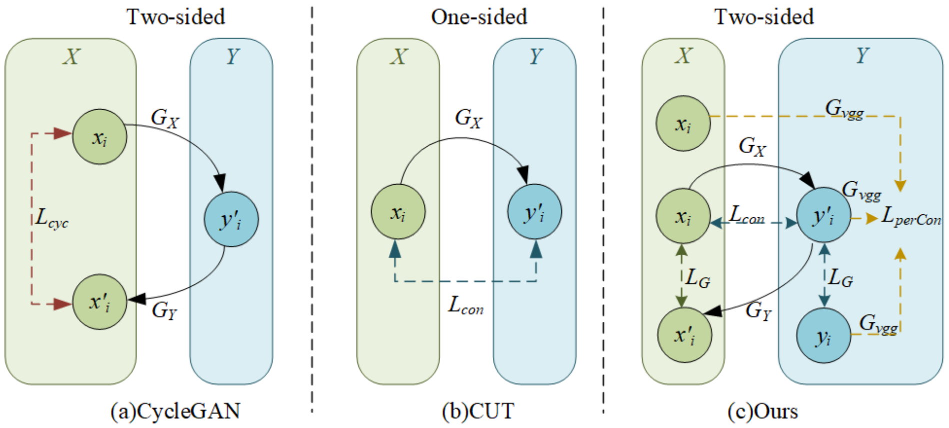
Therefore, the main contributions of this paper are as follows:

- We propose a dual-branch network for unsupervised infrared image colorization. The large differences between the infrared and visible domains are fitted by multiple contrastive learning.
- We propose perceptual contrastive loss to enhance the similarity between the generated image and the visible image in the high-dimensional feature space, making the colorized image more compatible with human visual perception.
- A multiscale residual module is designed to help the encoder process feature maps at different scales, enhancing the generator network's feature extraction capability.
- A feature fusion attention residual block is designed to integrate multiscale feature information, focusing on important features during up-sampling to produce higher-quality colorized images.

2. Related Work

2.1. Image Translation Task

Image translation techniques [16] aim to convert an image from one image domain X to another image domain Y, enabling cross-domain translation between images. This involves removing some attribute X from the original image and giving it a new attribute Y. The image-to-image translation task is a task of converting an input image into an output image, which usually involves translation between different domains such as semantic segmentation maps to real street maps, grayscale maps to color maps, and so on [17,18]. Common image-to-image translation methods include Generative Adversarial Networks (GANs) [19], Conditional GANs [2], pix2pix [16], StarGAN [20], Unsupervised Image-to-image Translation [12], etc. GANs use adversarial training to learn to generate realistic images, while cGANs introduce conditional information to generate images under specific conditions. StarGAN supports multi-domain image translation, allowing diverse transformations such as face translation into different styles of make-up, etc. Supervised learning approaches such as cGANs and pix2pix are constrained by the need for paired training data, which limits the feasibility of the algorithms. To overcome this limitation, unsupervised learning methods become the key means to solve the problem of unpaired data for image translation tasks [21–24]. Unsupervised image translation tasks have far-reaching implications for applications such as cross-domain image synthesis, style migration, and data enhancement. CycleGAN [10] and CUT [11] are representatives of unsupervised image translation methods. As shown in Figure 1, cycleGAN has a two-sided framework and learns the mapping between two domains unsupervised by introducing the cycle consistency loss. CUT adopts a one-sided framework and realizes unsupervised cross-domain translation by learning the shared latent space.



(a)CycleGAN (b)CUT (c)Ours

Figure 1. Comparison of our method with CycleGAN and CUT flowchart.

2.2. Unsupervised Image Translation Method Based on Contrastive Learning

The goal of contrastive learning is to learn an encoder such that similar data have similar representations in the coding space, while the representations of different classes of data are as different as possible [25–27]. Mutual information plays an important role in contrast learning, which measures the similarity information implied between two feature vectors. Typically, contrast learning divides the samples into two classes, positive and negative samples, and employs the Noise Contrastive Estimation (NCE) framework to maximize the similarity between related samples while minimizing the similarity between unrelated samples. In the image translation task, the input image and the target image have similarity information in the corresponding spatial locations. Based on this, CUT (Contrastive Unpaired Translation) introduces the idea of contrast learning in image translation and proposes a new loss function, the InfoNCE loss. CUT selects a patch at a random location in the generated image as the CUT selects a random patch in the generated image as an anchor point, and it considers the patch at the corresponding position in the input image as a positive sample and the patches at the remaining positions as negative samples. By skillfully constructing the patch-level contrast loss, CUT can effectively map the input image to the target domain and realize unsupervised image translation. This approach is suitable for many real-world scenes where a large amount of paired data are not available.

2.3. Deep Learning-Based Cross-Domain Colorization of Infrared Images

The infrared image cross-domain colorization task is an important research direction in the field of infrared image processing, which aims to map infrared grayscale images from the infrared domain to the visible domain [28,29]. This task is of great significance for expanding the availability of infrared image data and improving the adaptability of human eye observation. In recent years, fully automated deep learning-based colorization methods for infrared images across domains have made great progress in achieving remarkable success. Compared with traditional colorization methods for infrared images, these methods exhibit excellent robustness and generalization. Deep learning-based colorization methods were first applied to process NIR images due to their richer detail information, higher contrast, etc. Limmer et al. [30] implemented a cross-domain translation task for NIR images using convolutional neural networks. Influenced by thermal infrared sensors, atmosphere and other factors, thermal infrared images often suffer from image distortion and unclear targets, which increases the difficulty of the thermal infrared image colorization task. Berg et al. [1] proposed a fully automated thermal infrared image colorization method, which uses the U-Net architecture to convert thermal infrared images into visible color images with reasonable brightness and chromaticity. Kuang [31] and Neeraj Bhat [32] were trained using the KAIST-MS [33] dataset with a combination of content loss and adversarial loss. Existing deep learning-based colorization methods for infrared images are mainly fully supervised learning approaches, which require a large amount of paired infrared-color image data. However, limited by various reasons such as scene and camera, paired infrared image data is almost non-existent in real scenes. Therefore, fully supervised learning methods have certain limitations in practice.

3. Proposed Method

3.1. Architecture

Figure 2 depicts the architecture of DC-Net, featuring two principal branches: the patch-wise contrastive guidance branch (PwCGB) for local and global feature extraction and the perceptual contrastive guidance branch (PCGB), which focuses on semantic detail aligned with human visual perception. DC-Net operates via a dual-path framework illustrated in Figure 1c. In PwCGB, the input infrared image x_i is converted to the color output y_i' by the generator $G_{I2V}(G_X)$, which is then reduced to the infrared representation x_i' by the generator $G_{V2I}(G_Y)$. The DC-Net model employs a contrast loss rather than a cyclic consistency loss to compare the infrared input with the generated color image, using the generative adversarial loss to constrain x_i and x_i' as well as y_i and y_i'. In PCGB, the input infrared image x_i, the color output y_i', and the visible image y_i are processed through a pre-trained VGG16 network to obtain the feature maps used for contrast loss computation. With this multi-level feature learning approach, DC-Net facilitates the conversion of high-fidelity infrared to visible images.

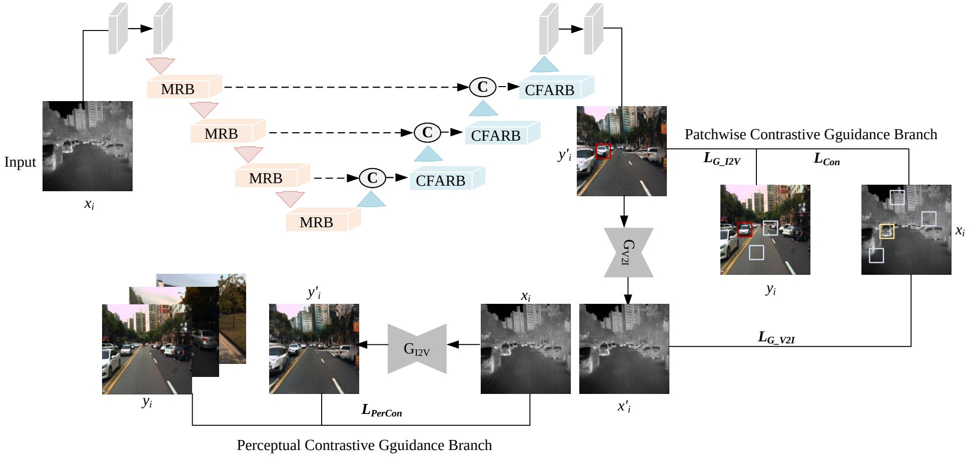


Figure 2. Overall architecture of DC-Net.

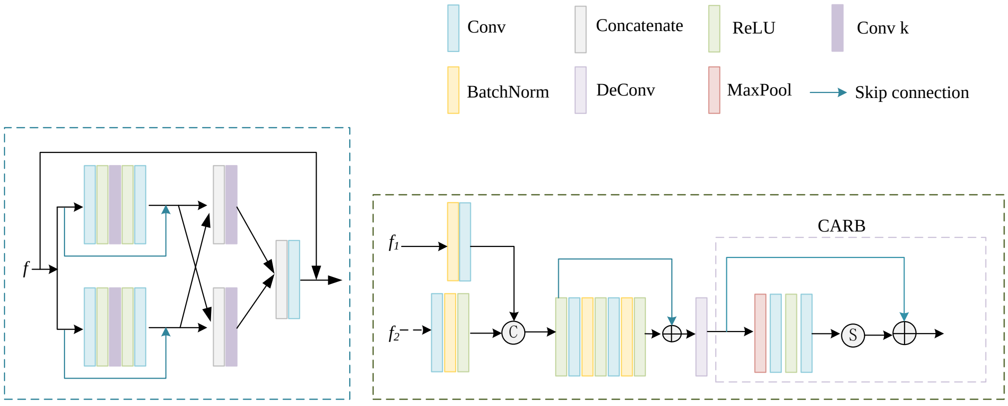
3.2. Patch-Wise Contrastive Guidance Branch

In the PwCGB, we apply a patch-based strategy to dissect the input image into small chunks, selecting positive and negative samples randomly. Positive samples correspond to areas with target similarities, while negatives exhibit differences. Through contrastive learning, we refine color and texture matching across similar patches, differentiating features where needed to bring the synthetic colorized infrared images closer to actual visible-domain images in terms of local qualities. For comprehensive feature extraction, we incorporate generative adversarial loss. This fosters a convergence of characteristics between the colorized infrared and visible images, aligning global colors and textures.

U-Net is commonly used as a generator in generative adversarial networks, featuring down-sampling paths (encoder) and up-sampling paths (decoder) to learn high-resolution feature mappings. While the traditional U-shaped network structure can fuse contextual information, U-Net relies solely on convolution and pooling operations, which are insufficient to capture multiscale information. Additionally, U-Net uses direct feature fusion and cannot selectively focus on important features.To improve the generator network's capacity to capture and represent the intricate features of input infrared images, we have developed a multiscale residual attention U-Net (MRA-UNet) as the generator in the PwCGB branch.

3.2.1. Multiscale Residual Block

We introduce a multiscale residual block (MRB) in the encoder, as shown in Figure 3a. Residual blocks from ResNet are combined with convolution kernels of different sizes to construct a parallel two-branch structure. This structure helps the generator better extract multiscale low-frequency texture information through cross-layer residual connections, strengthening the network's feature learning and information transfer capabilities. The multiscale residual module extracts features at different scales in the down-sampling stage, capturing richer image details and improving feature extraction capability. MRB directly passes important information to deeper layers, reducing the loss of feature information.



(a) MRB

(b) CFARB

Figure 3. The structure of MRB and CFARB.

3.2.2. Channel Attention Residual Block

As shown in Figure 3b, we combine the channel attention module with residual connections to design the Channel Attention Residual Block (CARB) and introduce the Feature Fusion Attention Residual Block (CFARB) in the decoder. The attention mechanism guides the network to learn effective features faster by modeling the interdependence between feature channels, adaptively fusing the features of each channel. This allows for the better integration of multiscale information, enabling a more effective combination of low-level and high-level features during up-sampling, thus improving colorization accuracy. It also facilitates the colorization generator to capture more useful channel feature information.

3.3. Perceptual Contrastive Guidance Branch

The perceptual loss aims to measure the perceptual similarity between the model's generated output and real data. Leveraging a pre-trained CNN, specifically VGG16, it assesses high-level feature correspondence. This measure has proven effective in visual applications like image synthesis and style transfer, prompting its use in our PCGB branch for infrared-to-visible image translation, emphasizing semantic content to yield colorized outputs closely matching the visible spectrum.

As shown in Figure 4, we extract features from visible, colorized infrared, and infrared images using VGG16 and then implement perceptual contrastive learning with sampled blocks from identical spatial locations across these features. This enables the model to integrate semantic and color attributes pertinent to both spectrums. To broaden diversity, negative samples are drawn not just from infrared but also from visible feature maps. Specifically, grayscale and non-corresponding blocks from visible images serve as negatives, enhancing structural and content fidelity within the colorized products, thus ensuring their alignment with human visual expectations.

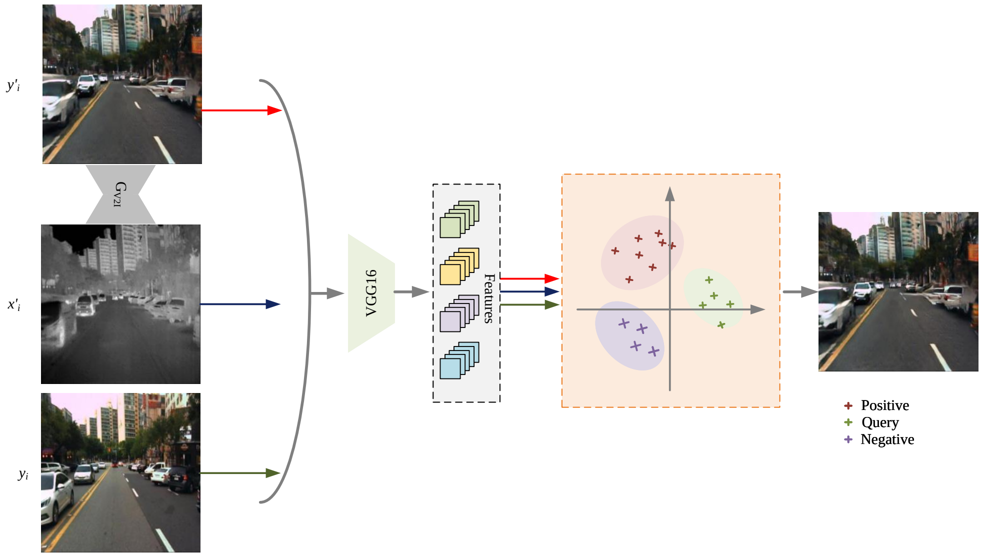


Figure 4. Structure of the perceptual contrastive guidance branch.

3.4. Loss Function

In this section, we will discuss the loss functions employed in both the patch-wise contrastive guidance branch (PwCGB) and the perceptual contrastive guidance branch (PCGB). The composite loss function in the PwCGB branch is a combination of the contrastive loss (L_{Con}) and the generative adversarial loss (L_{PerCon}).

3.4.1. Contrastive Loss

The initial step involves selecting an anchor vector Q as the query vector. Subsequently, one positive sample K^+ and $N-1$ negative samples K^- are chosen. The contrastive loss is then computed to compare the similarity features in the cross-domain coloring task between the infrared and visible domains. L_{Con} serves as a constraint for the colorized infrared image generator G, eliminating the need for paired visible images as label information. The representation of the contrastive loss is as follows:

$$L_{Con} = -\log[\frac{\exp(Q \cdot K^+/\tau)}{\exp(Q \cdot K^+/\tau) + \sum_{i=1}^{N-1} \exp(Q \cdot K^-/\tau)}] \tag{1}$$

In the equation, the parameter τ is a fixed value set to 0.07.

3.4.2. Generative Adversarial Loss

To better restore the chromatic and luminance information of the overall infrared image, a generative adversarial loss is introduced in the PwCGB branch. For any input infrared image I_x, the generator G and discriminator D collaborate to encourage the generated colorized infrared image $G(I_x)$ to compete with the target domain visible image I_y, aiming for more accurate color information. The representation of the generative adversarial loss L_G is as follows:

$$L_G = \mathbb{E}_{I_y \sim Y} \log D(I_y) + \mathbb{E}_{I_x \sim X} \log(1 - D(G(I_x))) \tag{2}$$

3.4.3. Perceptual Contrastive Loss in PCGB Branch

Through a fixed pre-trained VGG16 network, the infrared image, colorized infrared image, and visible image are individually employed as inputs to the pre-trained VGG16 network, yielding corresponding output features: $feature\_i$, $feature\_per$, and $feature\_rgb$. Then, the query vector q is selected in $feature\_per$, and the contrastive loss is computed with positive samples k^+ and negative samples k^-, respectively. Make the generated color information and structural detail information more in accordance with their semantic features. The representation of the perceptual contrastive loss L_{PerCon} is as follows:

$$L_{PerCon} = \sum_{0}^{i} \lambda \frac{\|q,k^{+}\|_1}{\|q,k^{-}\|_1 + t} \tag{3}$$

where i is the number of feature layers extracted by VGG, λ is the weight, and t is a fixed parameter with a fixed value of 1×10^{-7}. Based on the two aforementioned loss functions, the overall loss function can be defined as

$$L = \lambda_1 L_{Con} + \lambda_2 L_G + \lambda_3 L_{PerCon} \tag{4}$$

where λ_1, λ_2, and λ_3 represent the weights for contrastive loss, global feature loss, and perceptual loss, respectively.

4. Experiments

4.1. Experiments Settings

4.1.1. Datasets

Our models are trained and evaluated on different datasets for different infrared bands. In the NIR band, we used the NIRScene [34] dataset. This dataset consists of 477 images in 9 categories captured in both RGB and NIR modalities. The scene categories of the NIRScene dataset contain a countryside, field, forest, indoor, mountain, old building, street, city, and lake. In the thermal infrared band, we randomly selected several images in the multispectral pedestrian detection dataset KAIST [33] as a training set. The KAIST pedestrian dataset consists of a total of 95,328 images, each of which contains both RGB color images and infrared image versions. The KAIST dataset captures a variety of regular traffic scenes including schoolyards, streets, and the countryside.

4.1.2. Quantitative Evaluation Metrics

Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Mean Squared Error (MSE) were employed in this study as quantitative evaluation metrics to measure the quality of colorized infrared images. The quantitative experimental results are reported as the average scores across all images in the test set.

4.1.3. Training Details

For training, we randomly select 1000 unpaired images in the KAIST dataset and crop the dataset's image size to 256 × 256 using a sliding window. Our model is trained on a single NVIDIA 2080Ti GPU with a batch size of 1. During training, we optimize using the Adam optimizer with $\beta_1 = 0.5$, $\beta_2 = 0.999$. The model is trained with an initial learning rate of 0.0001 for 200 epochs with a total training time of 24.38 h. To compute the contrast loss and the perceived contrast loss, we select layers 4, 8, 12, and 16 for the computation. For each layer, we select 256 patches. In the training process, how to assign appropriate weights for adversarial loss and contrast loss is an important issue. Too large or too small weights may lead to undesirable results. For example, if the weight of adversarial loss is too high, it may lead to the loss of details in the generated image; while if the weight of contrast loss is too high, it may inhibit the ability to generate diversity. After experiments, we set the loss weights of each component to 0.5, 0.5, and 1.

4.2. Quantitative Testing

In order to quantitatively analyze our method and other unsupervised image translation algorithms, tests are performed on the test sets of the NIR dataset and KAIST dataset, respectively.

Tables 1 and 2 show the quantitative test results of CycleGAN, CUT, FastCUT, IR-colorization [35], TIC-CGAN [31], and the method proposed in this paper on the KAIST and NIR datasets, respectively. The results show that the best performance is achieved in each quantitative metric using our method. Moreover, for the KAIST thermal infrared dataset, our method has a significant advantage. On the KAIST dataset, our method shows a significant improvement over the CycleGAN method. Our method improves 28.9% in PSNR, 39.0% in SSIM, and 48.4% in MSE. Compared to the CUT method, our method also achieved better results. It improved 23.8% in PSNR, 35.5% in SSIM, and 43.1% in MSE. Compared with the FastCUT method, our method shows improvements of 14.3% in PSNR, 32.4% in SSIM, and 44.1% in MSE. Compared with the IR-colorization method, our method shows improvements of 12.5% in PSNR, 17.7% in SSIM, and 50% in MSE. Compared with the TIC-CGAN method, our method shows improvements of 11.0% in terms of PSNR, 7.3% in terms of SSIM, and 40% in terms of MSE. Our method also achieves the best measurement results on the NIR dataset.

Table 1. Average quantitative results of different methods on the KAIST dataset.

Dataset	Method	Side Type	PSNR	SSIM	MSE
KAIST	CycleGAN	Two-sided	13.064	0.420	0.064
	CUT	One-sided	13.597	0.431	0.058
	FastCUT	One-sided	14.730	0.441	0.059
	IR-colorization	One-sided	14.961	0.496	0.066
	TIC-CGAN	Paired	15.155	0.544	0.055
	Ours	Two-sided	16.833	0. 584	0.033

Table 2. Average quantitative results of different methods on the NIR dataset.

Dataset	Method	PSNR	SSIM	MSE
NIR	CycleGAN	17.794	0.589	0.036
	CUT	17.892	0.621	0.034
	FastCUT	18.600	0.676	0.032
	IR-colorization	17.807	0.615	0.041
	TIC-CGAN	17.940	0.639	0.030
	Ours	18.828	0.685	0.031

To compare the performance of our method with other unsupervised image translation methods more intuitively, as shown in Figure 5, we use a line graph to represent the average measurement results on the KAIST dataset. As shown in Figure 6, we use a line graph to represent the average measurement results on the NIR dataset.

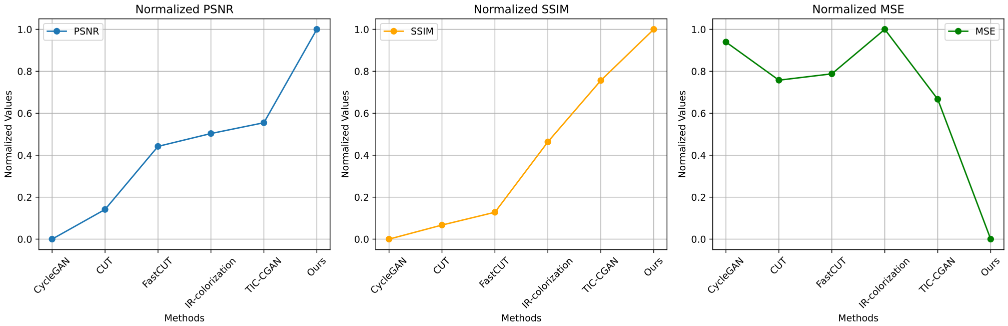


Figure 5. Average quantitative results of different methods on the KAIST dataset.

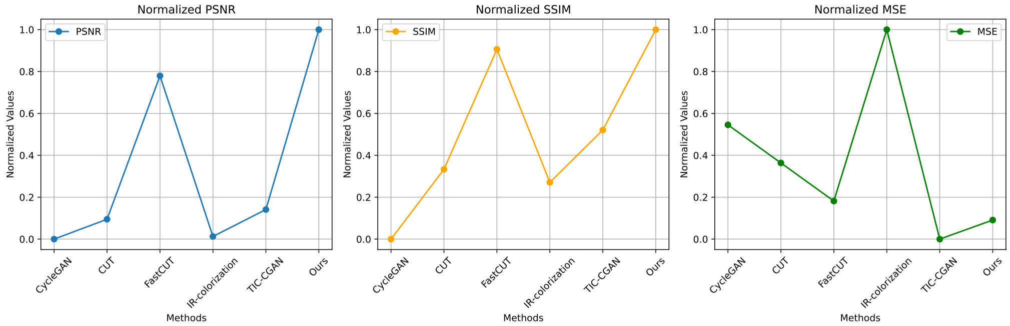


Figure 6. Average quantitative results of different methods on the NIR dataset.

4.3. Qualitative Testing

Figure 7 shows the colorization results of different unsupervised image translation methods on the KAIST dataset. For the street scene, each method recovers the approximate color. However, CUT, CycleGAN, and FastCUT fail to accurately recover the structure and details of the "vehicle" target on the street, and the taillights on the "vehicle" are not correctly colored. For the campus scene, the first three methods only recover the colors of the sky and the ground but seriously lose the color and structural information of the targets such as "buildings" and "lane lines". IR-colorization successfully recovers the colors of taillights but loses most of the feature colors of targets such as "crosswalks" and "buildings". TIC-CGAN performs well in recovering "lane lines" and "crosswalks", but the results are structurally ambiguous and lack texture information. Our method preserves the structural information of targets such as trees, buildings, vehicles, and lane lines while also restoring their colors based on human eye perception.

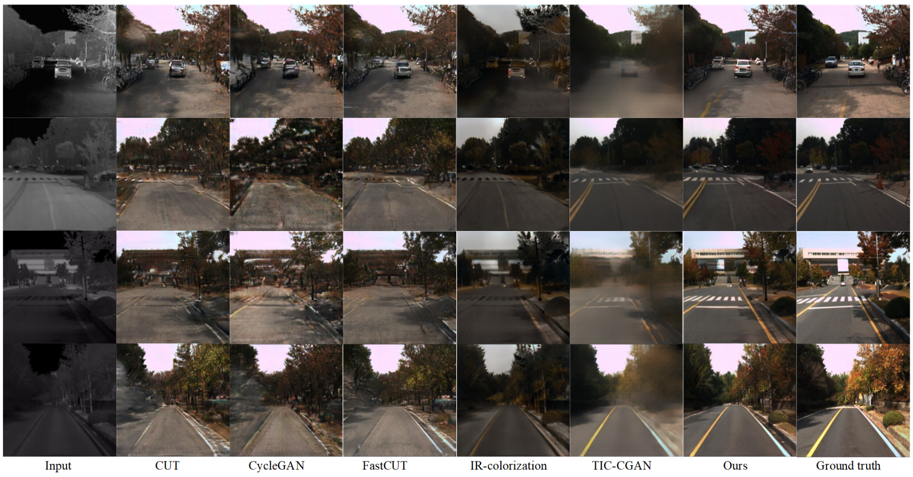


Figure 7. Colorization results of different methods on the KAIST dataset.

Figure 8 shows the colorization results of different unsupervised image translation methods on the test set NIR-Scene. For the mountain scene, CUT, CycleGAN, and FastCUT all exhibit a coloring disorder, incorrectly assigning roads the color of mountains. In the indoor scenes, CUT and FastCUT show different degrees of texture mosaicing, while CycleGAN results in blurred "chandeliers". On the contrary, IR-colorization and TIC-CGAN perform well for mountain scenes, but the "chandelier" appears blurred when dealing with indoor scenes. The colorization of our method matches the semantic information, and the texture is clear. For the lake scene, both CycleGAN and our method are able to color accurately, while CUT, FastCUT, IR-colorization, and TIC-CGAN have coloring errors. In addition, only our method succeeded in recovering the color of the mountains around the lake.

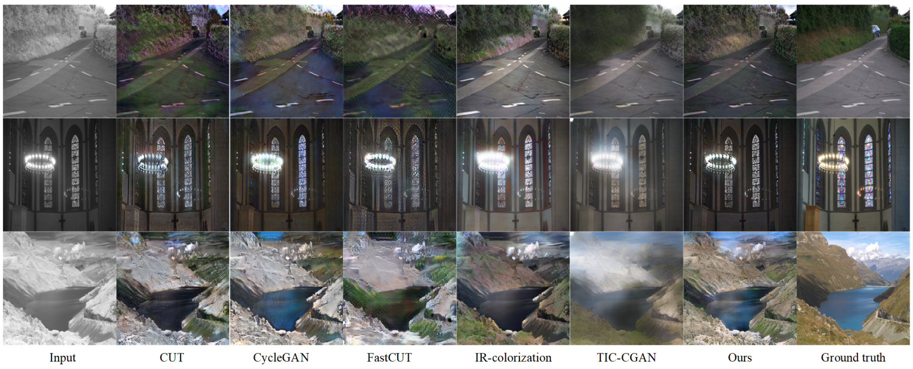


Figure 8. Colorization results of different methods on the NIR dataset.

To verify the effectiveness of DC-Net in detail, texture and coloring, as shown in Figure 9, on the thermal infrared dataset we show the effect of each method to colorize the target using a car as an example. Among them, CUT, CycleGAN edge distortion is the most obvious, TIC-CGAN detail information is lost the most, and only our method accurately colors the body and lights. On the NIR dataset, we take the chandelier as an example to show the effect of each method to colorize the target. Among them, CUT and FastCUT all show serious mosaic phenomena, while CycleGAN, IR-colorization and TIC-CGAN show blurred details.

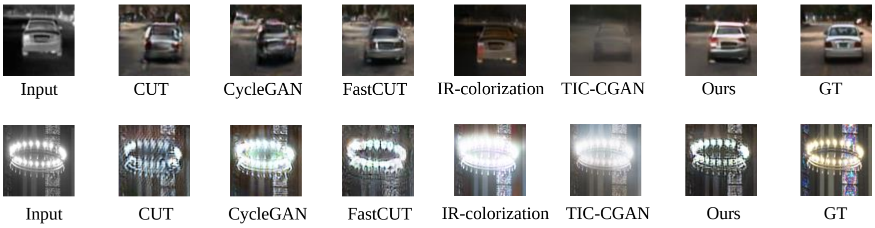


Figure 9. Colorization results each method for targets in different wavelength.

Overall, the qualitative test results show that our method achieves good performance in the task of converting infrared images to visible images, and the generated visible images have accurate colors and rich details. All the unsupervised image translation methods mentioned in the figure can recover the low-frequency background information such as "sky" and "trees" very well. However, for the fast-changing details, textures, edges, and other parts of the image such as "vehicle", "crosswalk", "lane line", etc., the existing unsupervised image translation methods can not generate a high-quality colorized infrared image structure. High-quality colorized infrared image structure detail information. Our method adeptly manages scenes with significant temperature fluctuations, like those at object edges or in localized areas with notable temperature variations. In these conditions, our approach produces high-frequency data, effectively capturing intricate details and local features within the image. Additionally, it generates color information akin to that found in visible images within the target domain.

4.4. Ablation Study

To evaluate the impact of different parts within our dual-branch structure, consisting of PwCGB and PCGB, we conducted ablation studies. In this part, we train our network by reducing the weights of each part of the loss function to 0 to test their effects. In the PwCGB branch, we try to remove the contrastive loss L_{Con} and the multiscale residual attention generator. To test the effectiveness of PCGB, we try to remove L_{PerCon}. We performed ablation experiments on three scenes from the KAIST dataset, and the colorization results are shown in Figures 10–12. In the school scene, removing both the perceptual contrast loss and the MRA-UNet results in an overall dark color and the loss of most building details. In the street scene, removing contrast loss, perceptual contrast loss, and the MRA-UNet results in blurred building structures and loss of small targets in the blue box. In the traffic scene, targets like "lane lines" and "vehicles" show severe artifacts and are not colored correctly. The average quantitative results after removing different parts are shown in Table 3.

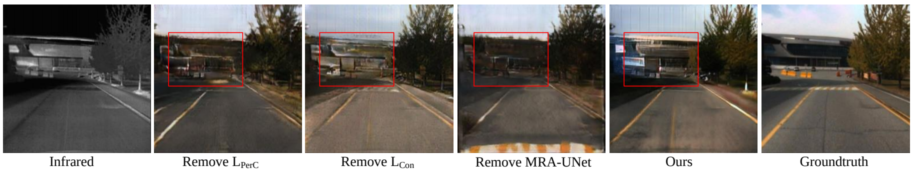


Figure 10. The infrared image colorization results under school scene.

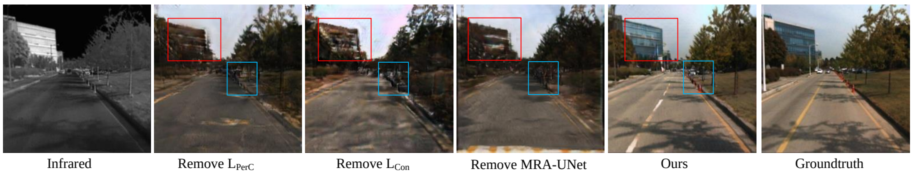


Figure 11. The infrared image colorization results under street scene.

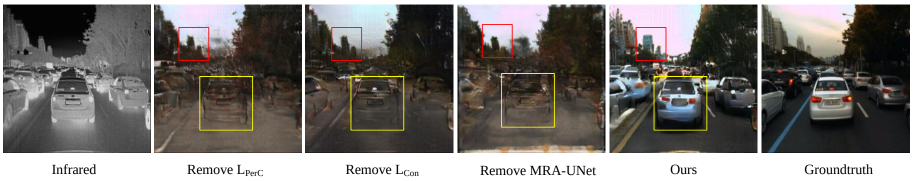


Figure 12. The infrared image colorization results under traffic scene.

Table 3. Average quantitative results with different parts removed.

Dataset	Method	PSNR	SSIM	MSE
KAIST	Without LCon	14.740	0.334	0.069
	Without LPerCon	14.562	0.332	0.061
	Without MRA-UNet	15.028	0.332	0.056
	Ours	16.833	0.584	0.033

5. Conclusions

This paper presents a novel unsupervised method for translating unpaired infrared images to visible images using a semantic-aware dual-branch contrastive learning network. The proposed network structure leverages contrastive learning for the transformation process and enriches it with high-level perceptual features extracted by pre-trained deep learning models. These features guide the colorization process, resulting in images that more closely align with human visual perception. In addition, the multiscale residual attention generator in PwCGB efficiently learns both local and global features of the image through multi-layer residual blocks. The residual connectivity enables the model to better capture the detailed information in the image, reduces information loss, and helps generate more realistic color images. The feature fusion attention residual block introduced by the generator enables a finer tuning of feature responses on different channels, improving detail retention and color accuracy during colorization. Experimental results indicate that our method effectively infers RGB values from infrared grayscale information, yielding colorized infrared images of high quality. The generated images exhibit accurate colors and detailed textures, performing well in translation tasks from infrared to visible imagery. Nevertheless, real-world scenarios where high-temperature objects diffuse thermal radiation can cause indistinct boundaries in colorized results due to similar grayscale values in surrounding areas. Future work will focus on addressing this issue to enhance boundary clarity. Additionally, the use of colorized infrared image models in real-time surveillance and defense systems is considered. In future research, we will investigate model pruning and quantization techniques to reduce model size and computational demands, enhancing inference speed.

Author Contributions: Conceptualization, T.J. and J.B.; methodology, T.J.; software, T.J.; validation, T.J., J.B. and L.X.; formal analysis, T.L.; investigation, L.X.; resources, Y.L.; data curation, T.J. and T.L.; writing—original draft preparation, T.J.; writing—review and editing, Y.L., J.B. and X.S.; visualization, T.J. and X.K.; supervision, Y.L. and X.K.; project administration, X.S. and Q.C.; funding acquisition, Y.L., X.S. and Q.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Science and Technology Major Project of China under Grant 2019-V-0008-0102; in part by the National Natural Science Foundation of China under Grant 62301253 and Grant 62305163; in part by the National Key Laboratory of Infrared Detection Technologies under Grant IRDT-23; and in part by the National Key Laboratory of Science and Technology on Multispectral Information Processing under Grant 6142113210205.

Data Availability Statement: Data underlying the results presented in this paper are available in Refs. [33,34].

Conflicts of Interest: Author Junqi Bai was employed by the 28th Institute of China Electronics Technology Group Corporation. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Kuang, X.; Sui, X.; Liu, Y.; Chen, Q.; Gu, G. Single Infrared Image Enhancement Using a Deep Convolutional Neural Network. *Neurocomputing* **2019**, *332*, 119–128. [CrossRef]
2. Berg, A.; Ahlberg, J.; Felsberg, M. Generating Visible Spectrum Images From Thermal Infrared. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1143–1152.
3. Suárez, P.L.; Sappa, A.D.; Vintimilla, B.X. Infrared image colorization based on a triplet dcgan architecture. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 18–23.
4. Ji, J.; Zhang, Y.; Lin, Z.; Li, Y.; Wang, C.; Hu, Y.; Huang, F.; Yao, J. Fusion of Infrared and Visible Images Based on Optimized Low-Rank Matrix Factorization with Guided Filtering. *Electronics* **2022**, *11*, 2003. [CrossRef]
5. Jin, X.; Jiang, Q.; Yao, S.; Zhou, D.; Nie, R.; Hai, J.; He, K. A survey of infrared and visual image fusion methods. *Infrared Phys. Technol.* **2017**, *85*, 478–501. [CrossRef]
6. Xu, H.; Ma, J.; Jiang, J.; Guo, X.; Ling, H. U2Fusion: A unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 502–518. [CrossRef] [PubMed]
7. Fu, Q.; Fu, H.; Wu, Y. Infrared and Visible Image Fusion Based on Mask and Cross-Dynamic Fusion. *Electronics* **2023**, *12*, 4342. [CrossRef]
8. 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.
9. 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; pp. 319–345._19. [CrossRef]
10. Liu, M.-Y.; Breuel, T.; Kautz, J. Unsupervised Image-to-Image Translation Networks. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017.
11. Sigillo, L.; Grassucci, E.; Comminiello, D. StawGAN: Structural-Aware Generative Adversarial Networks for Infrared Image Translation. In Proceedings of the 2023 IEEE International Symposium on Circuits and Systems (ISCAS), Monterey, CA, USA, 21–25 May 2023; pp. 1–5. [CrossRef]
12. Wang, G.; Shi, H.; Chen, Y.; Wu, B. Unsupervised image-to-image translation via long-short cycle-consistent adversarial networks. *Appl. Intell.* **2023**, *53*, 17243–17259. [CrossRef]
13. Ye, J.; Guo, J. Dual-level interactive multimodal-mixup encoder for multi-modal neural machine translation. *Appl. Intell.* **2022**, *52*, 14194–14203. [CrossRef]
14. Liu, T.; Liu, Y.; Zhang, C.; Yuan, L.; Sui, X.; Chen, Q. Hyperspectral Image Super-Resolution via Dual-Domain Network Based on Hybrid Convolution. *IEEE Trans. Geosci. Remote. Sens.* **2024**, *62*, 1–18. [CrossRef]
15. Liu, Y.; Qiu, B.; Tian, Y.; Cai, J.; Sui, X.; Chen, Q. Scene-Based Dual Domain Non-Uniformity Correction Algorithm for Stripe and Optics-Caused Fixed Pattern Noise Removal. *Opt. Express* **2024**, *32*, 16591–16610. [CrossRef] [PubMed]
16. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.

17. Eskandar, G.; Abdelsamad, M.; Armanious, K.; Yang, B. USIS: Unsupervised Semantic Image Synthesis. *Comput. Graph.* **2023**, *111*, 14–23. [CrossRef]
18. Ma, Z.; Li, J.; Wang, N.; Gao, X. Semantic-related image style transfer with dual-consistency loss. *Neurocomputing* **2020**, *406*, 135–149. [CrossRef]
19. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2014.
20. Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; Choo, J. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8789–8797.
21. 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* **2019**, arXiv:1907.10830.
22. Lin, J.; Xia, Y.; Liu, S.; Zhao, S.; Chen, Z. ZstGAN: An adversarial approach for Unsupervised Zero-Shot Image-to-image Translation. *Neurocomputing* **2021**, *461*, 327–335. [CrossRef]
23. Tang, H.; Xu, D.; Sebe, N.; Yan, Y. Attention-Guided Generative Adversarial Networks for Unsupervised Image-to-Image Translation. In Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 14–19 July 2019; pp. 1–8. [CrossRef]
24. Han, J.; Shoeiby, M.; Petersson, L.; Armin, M.A. Dual Contrastive Learning for Unsupervised Image-to-Image Translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 746–755.
25. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. In Proceedings of the 37th International Conference on Machine Learning, PMLR, Virtual, 13–18 July 2020; pp. 1597–1607.
26. Gao, T.; Yao, X.; Chen, D. SimCSE: Simple Contrastive Learning of Sentence Embeddings. *arXiv* **2021**, arXiv:2104.08821.
27. Wu, H.; Qu, Y.; Lin, S.; Zhou, J.; Qiao, R.; Zhang, Z.; Xie, Y.; Ma, L. Contrastive Learning for Compact Single Image Dehazing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 10551–10560.
28. 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]
29. Luo, F.-Y.; Cao, Y.-J.; Yang, K.-F.; Li, Y.-J. Memory-Guided Collaborative Attention for Nighttime Thermal Infrared Image Colorization. *arXiv* **2022**, arXiv:2208.02960.
30. Limmer, M.; Lensch, H.P.A. Infrared Colorization Using Deep Convolutional Neural Networks. In Proceedings of the 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, USA, 18–20 December 2016; pp. 61–68. [CrossRef]
31. 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. [CrossRef]
32. Bhat, N.; Saggu, N.; Pragati; Kumar, S. Generating Visible Spectrum Images from Thermal Infrared using Conditional Generative Adversarial Networks. In Proceedings of the 2020 5th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 10–12 June 2020; pp. 1390–1394. [CrossRef]
33. 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, Boston, MA, USA, 7–12 June 2015; pp. 1037–1045.
34. Brown, M.; Süsstrunk, S. Multi-spectral SIFT for scene category recognition. In Proceedings of the VPR 2011, Colorado Springs, CO, USA, 20–25 June 2011; pp. 177–184. [CrossRef]
35. Chen, L.; Liu, Y.; He, Y.; Xie, Z.; Sui, X. Colorization of infrared images based on feature fusion and contrastive learning. *Opt. Lasers Eng.* **2023**, *162*, 107395. [CrossRef]

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