sensors

MDPI

Article

V2T-GAN: Three-Level Refined Light-Weight GAN with Cascaded Guidance for Visible-to-Thermal Translation

Ruiming Jia, Xin Chen, Tong Li and Jiali Cui *

School of Information Science and Technology, North China University of Technology, Beijing 100144, China; jiaruiming@ncut.edu.cn (R.J.); xinchen@mail.ncut.edu.cn (X.C.); litong300@pingan.com.cn (T.L.)
* Correspondence: jialicui@ncut.edu.cn

Abstract: Infrared image simulation is challenging because it is complex to model. To estimate the corresponding infrared image directly from the visible light image, we propose a three-level refined light-weight generative adversarial network with cascaded guidance (V2T-GAN), which can improve the accuracy of the infrared simulation image. V2T-GAN is guided by cascading auxiliary tasks and auxiliary information: the first-level adversarial network uses semantic segmentation as an auxiliary task, focusing on the structural information of the infrared image; the second-level adversarial network uses the grayscale inverted visible image as the auxiliary task to supplement the texture details of the infrared image; the third-level network obtains a sharp and accurate edge by adding auxiliary information of the edge image and a displacement network. Experiments on the public dataset Multispectral Pedestrian Dataset demonstrate that the structure and texture features of the infrared simulation image obtained by V2T-GAN are correct, and outperform the state-of-the-art methods in objective metrics and subjective visualization effects.

Keywords: image domain translation; infrared image simulation; generative adversarial network

Citation: Jia, R.; Chen, X.; Li, T.; Cui, J. V2T-GAN: Three-Level Refined Light-Weight GAN with Cascaded Guidance for Visible-to-Thermal Translation. *Sensors* **2022**, *22*, 2119. https://doi.org/10.3390/s22062119

Academic Editor: Min Yong Jeon

Received: 16 February 2022
Accepted: 7 March 2022
Published: 9 March 2022

1. Introduction

Infrared images are widely used in military, medical, industrial and agricultural fields, and are generally obtained by shooting target scenes with an infrared thermal imager. However, in some special environments, the amount of image data that can be obtained by an infrared thermal imager is relatively insufficient, and the equipment is expensive. These problems limit the acquisition of infrared image data. Therefore, related research on infrared image simulation has been progressing.

The traditional infrared simulation approach can be divided into two types: infrared image simulation based on three-dimensional modeling and infrared image simulation based on visible light image. The first method uses three-dimensional modeling of the scene, and then simulates according to infrared radiation characteristics [1–3], without the need for real visible light images. The disadvantage is that the overall process is complicated, and the texture of the simulation result is unnatural. Furthermore, because it only targets a single scene, the generalization performance of the model is poor. The second method requires real visible light images, which is a simpler and more convenient method than the previous one, but it also has the disadvantages of low simulation accuracy and poor generalization ability. In view of the above problems, we aim to provide a more convenient, accurate and robust infrared simulation approach.

Infrared image simulation based on visible light image is a pixel-level image conversion task, which can predict and simulate corresponding infrared images through visible light images. Recently, the pixel-level image conversion task based on the deep learning method has achieved great success, and the algorithm is relatively simple and convenient. Common pixel-level image conversion tasks include monocular depth estimation [4,5], semantic segmentation [6,7], optical flow estimation [8], image style conversion [9], etc.

Among them, the first three tasks require high accuracy, and are generally implemented by convolutional neural networks (CNN). Due to the constraint of the objective function, although the methods based on CNN can obtain better results in the objective metrics, the predicted result map generally has the problem of blurred edges and texture loss. In order to solve this problem, some studies have used conditional generative adversarial network (cGAN) to achieve such tasks [9,10]. The generated image of the cGAN have a natural image texture, clear edges, and better visualization. Based on these observations, we use cGAN to achieve the conversion of a visible light image to the corresponding infrared image.

The visible light image and the infrared image have similar structural information, semantic information and edge information, and the grayscale inverted visible (GIV) im-age have a high degree of similarity with the infrared image in visualization and texture details. The GAN designed in this paper uses semantic segmentation images and GIV images as auxiliary tasks, and visible light edge images as auxiliary information, which can realize the conversion of a visible light image to infrared image end-to-end. To improve the efficiency of the algorithm, a variety of light-weight convolutions are used to reduce the amount of overall network parameters.

The contributions of this paper consist of three aspects. First, a three-level refined light-weight GAN with cascaded guidance (V2T-GAN) is proposed. It aims at converting a visible light image to infrared image end-to-end. Second, a three-level network framework for cascading guidance through auxiliary tasks and auxiliary information is proposed. The first-level network of V2T-GAN uses a semantic segmentation image as an auxiliary task to generate a coarse infrared image with a relatively correct structure; the second-level network uses the GIV image as an auxiliary task to supplement the detailed texture information of the infrared image; and the third-level network adds auxiliary information of the edge image and displacement field to obtain a clear and accurate edge. Third, extensive experiments were carried out on the public dataset Multispectral Pedestrian Dataset (MPD) [11], which clearly demonstrate the effectiveness of the proposed V2T-GAN.

The rest of this study is organized as follows. In Section 2, we discuss the related work on infrared image simulation. In Section 3, we introduce the proposed method in detail. In Section 4, we present the main experiments. In Section 5, we conclude with a brief summary and mention future work.

2. Related Work

2.1. Infrared Image Simulation

The traditional method of infrared simulation based on visible light image is generally divided into two stages: in the first stage, the visible light image needs to be segmented; and in the second stage, the gray-scale mapping relationship between the visible light image and infrared of different objects is established, and then the infrared image is simulated from the segmented visible light image. Zhou et al. [12] used the threshold method to segment the image, and combined the reflectivity of the ground object to establish the gray-scale mapping relationship, so as to obtain the simulated image. Li et al. [13] achieved image segmentation through a pulse-coupled neural network, after artificially calibrating the material to obtain simulation results through radiation calculation. Infrared imaging systems and visible light imaging systems are both complex and affected by multiple variables. This feature makes it difficult to express the mapping relationship between visible light images and infrared images with a unified formula. Therefore, the model generalization ability of this type of method is poor, and the simulation result lacks natural image texture information.

2.2. Pixel-Level Image Conversion Tasks

Image conversion tasks include image style conversion, image perspective conversion, depth estimation, optical flow estimation, semantic segmentation and so on. Generating a corresponding infrared image from a visible light image can be regarded as a mapping

from the visible light image domain to the infrared image domain, which is a kind of image domain conversion task. In recent years, deep learning methods have achieved good research results in image domain conversion tasks, such as monocular depth estimation and image style conversion. In [4], Eigen et al. first used an end-to-end CNN to predict the depth map. In further works, some people introduced concepts such as an attention mechanism and continuous conditional random field to improve the performance of the algorithm [14–16], whereas some achieved better results by optimizing the network structure [5]. In addition, there is also the use of multi-task [8,17–19] learning methods to obtain auxiliary information.

2.3. Conditional Generative Adversarial Network

Unlike the image depth estimation task, the conversion from visible light to infrared requires better objective evaluation results, as well as better visual effects. Constrained by the objective function, although the CNN method in the image depth estimation task has obtained good results in objective evaluation indicators, the output image is relatively blurry and loses many texture details. In the field of deep learning, the cGAN derived from the GAN [20] has excellent performance in the image domain conversion task [9,20–24], and the visual effect of the output image is better.

2.4. Lightweight Network

Various computer vision tasks implemented through deep learning have shortcomings, such as high network model redundancy and complex calculations. With the development of deep learning, the lightweight and high-efficiency of the network models has become more and more important. Currently, lightweight and efficient network structures are mostly used in tasks such as image classification, object detection and semantic segmentation. Moreover, most lightweight network models use lightweight convolution instead of standard convolution to build network structures. Zhang et al. [25] proposed ShuffleNet with efficient computing power to realize image classification. In [26], Krizhevsky et al. proposed group convolution (GConv) and constructed a network model that includes group convolution, which has fewer network model parameters and higher accuracy than ShuffleNet. MobileNetV2 [27] included Depthwise Separable Convolution (DSConv), which is a lightweight network model that can be applied to mobile architectures. Mehta et al. [28] proposed depthwise dilated separable convolution (DDSConv) and used it to construct the network ESPNetv2, which has high computational efficiency and has a large receptive field. Haase et al. [29] analyzed the depth separable convolution and improved it to obtain the blueprint separable convolution (BSConv). Han et al. [30] proposed the Ghost module, which can effectively reduce the redundancy of feature maps through simple linear operations, thereby greatly improving network computing efficiency.

3. V2T-GAN

In order to achieve the conversion from a visible light image to infrared image, we propose a three-level refined light-weight GAN with cascaded guidance. As shown in Figure 1, V2T-GAN is a three-level cascaded network. The first-level network uses semantic segmentation images as an auxiliary task to guide G_1 to learn infrared images with more accurate structural information; the second-level network uses GIV images as an auxiliary task to guide G_2 to learn more accurate infrared images with detailed textures; and the third-level network uses visible light edge images as auxiliary information to further optimize the predicted infrared images. G_d predicts the displacement offset map of the second-level network's output image T_2 in the x and y directions, and then resamples T_2 according to the displacement offset information to obtain the final infrared image T_3.

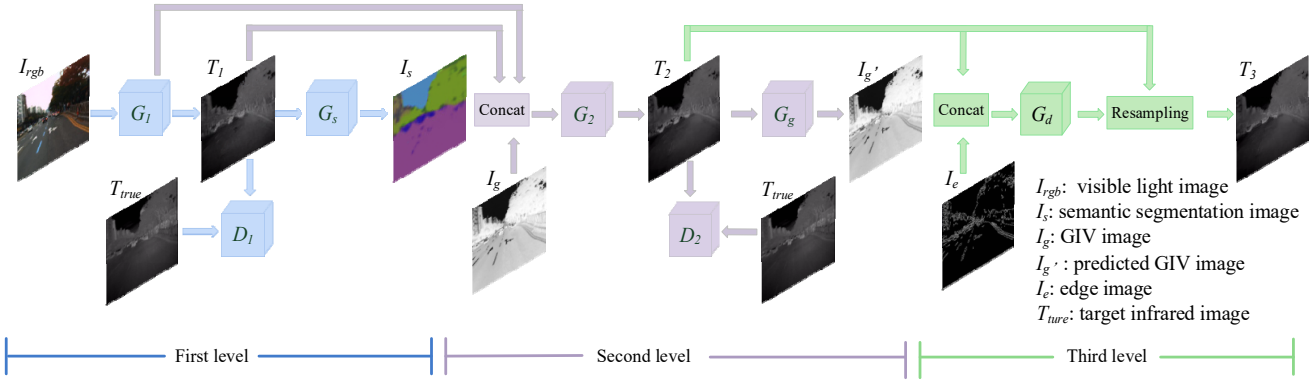

Figure 1. V2T-GAN network structure.

3.1. First-Level Network

The first-level network uses semantic segmentation images as auxiliary tasks to guide the first-level target task generative network to predict infrared images with more correct structure information. As shown in Figure 1, the blue part is the first-level network, including a target task generator G_1, a discriminator D_1 and an auxiliary task network G_s. G_1 estimates the corresponding infrared image T_1 from the visible light image, and D_1 is responsible for identifying the authenticity of the predicted infrared image T_1 and the target infrared image T_{true}. Then, G_s estimates the semantic segmentation image from T_1, and guides G_1 to pay more attention to the structure information by predicting the semantic segmentation image, thereby predicting the infrared image with more correct structure information.

The network structure of G_1 and G_s is the generator U-net [31] in pix2pix [9], and the network structure of D_1 is the discriminator in pix2pix. To reduce the overall parameter amount of the network, we adjust the initial output channel number of G_s to 4; that is, the number of all the feature map channels in the network is 1/16 of the original U-net. In order to improve the calculation efficiency of the overall algorithm, this paper generally uses lightweight convolutions in the network, such as GConv, DSConv, BSConv, Ghost module, etc. In V2T-GAN, G_1 has the largest overall network parameters, and its lightweight operation has the greatest impact on the overall network. Therefore, we analyzed and compared the different lightweight methods of G_1, and finally adopted GConv, and the standard convolution of G_1, G_s and D_1 are all replaced by GConv with a group number of 4.

There have been many research studies on lightweight convolutions, and the methods applied in this paper will be introduced below.

The specific implementation of GConv is divided into three steps:

- GConv divides the input channels into even and non-overlapping groups according to the grouping number g;
- Perform standard convolution independently on each group that has been divided;
- Concat the results of the standard convolution in the dimension of the channel.

Depthwise Convolution (DConv) [32] is a special type of GConv. The number of groups and output channels are the same as the number of input channels. The DSConv consists of two steps:

- The first step is to perform DConv;
- Use standard convolution with a 1×1 convolution kernel to adjust the number of output channels.

The BSConv is also divided into two steps:

- First perform standard convolution with a 1×1 convolution kernel to adjust the number of output channels;
- Use DConv.

The GhostModule is divided into three steps:

- Perform standard convolution with a 1×1 convolution kernel. The number of output channels in this step is: $C_1 = [C_{in}/r]$, where C_1 is the number of output channels in the first step, C_{in} is the number of input channels, and r represents the manually set rate;
- Use GConv on the output result of the first step, the number of groups is the number of output channels in the first step (that is, equal to C_1), the number of output channels in this step is: $C_1 \times (r-1)$;
- Concatenate the output result of the first step and the second step to get the final result.

3.2. Second-Level Network

The second-level network further optimizes the infrared image output by the first-level network, and uses the GIV images as an auxiliary task to guide the second-level target task generative network to predict infrared images with more accurate details and textures. The network structure is shown in the purple part of Figure 1, including a target task generator G_2, a discriminator D_2 and an auxiliary task network G_g. The input of G_2 is concatenated with the predicted infrared image T_1 of the first-level network, the output feature map of the penultimate layer of G_1 and the GIV image I_g, and finally the predicted infrared image T_2 is output. D_2 has the same structure and function as D_1, which used to discriminate the predicted infrared image T_2 and the target infrared image T_{true}. G_g estimates the GIV image from T_2, and guides G_2 to further optimize the predicted infrared image.

In the case of good lighting, visible light images have more detailed texture information than infrared images. GIV images also have rich detailed texture information, and compared with visible light images, it has a high similarity with infrared images in terms of human visual effects and some objective metrics, such as FID [33], LIPIS [34], etc. Therefore, we adopt the GIV image as the auxiliary task of the second-level network to guide G_2 to further optimize the detailed texture information of the predicted infrared image.

3.2.1. Target Task Generator G_2

An illustration of the second-level target task generator G_2 is depicted in Figure 2, consisting of an MFM module [28], four L-FMR modules that add skip connection and a Ghost module. The MFM module is shown in Figure 3. In order to obtain the information of multiple receptive fields, the input is respectively passed through four dilated convolutions with a convolution kernel size of 3×3 and a dilation rate of 1, 2, 3 and 4. The output of the dilated convolution with different dilation rates are added and fused, and then the added results are concatenated. The input and output of G_2 are similar. To increase the direct mapping between input and output, the input is added to the final concatenated result after a pointwise convolution. The L-FMR module is improved from the FMRB [35], which is a network module for image deblurring tasks. It has been verified that FRMB can learn and restore the detailed texture information of the image. In order to reduce the amount of overall network parameters, we replace all the standard convolutions in FMRB with a Ghost module with a rate of 4, which is the L-FMR module in Figure 2.

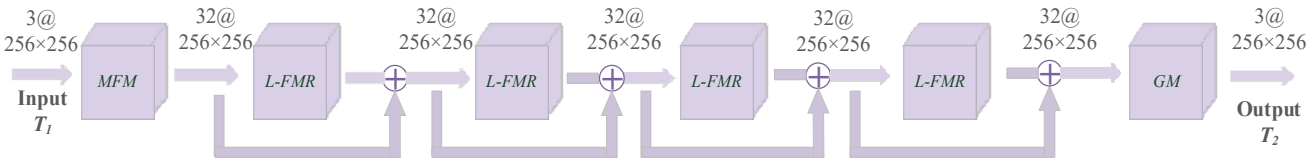

Figure 2. The proposed network of G_2.

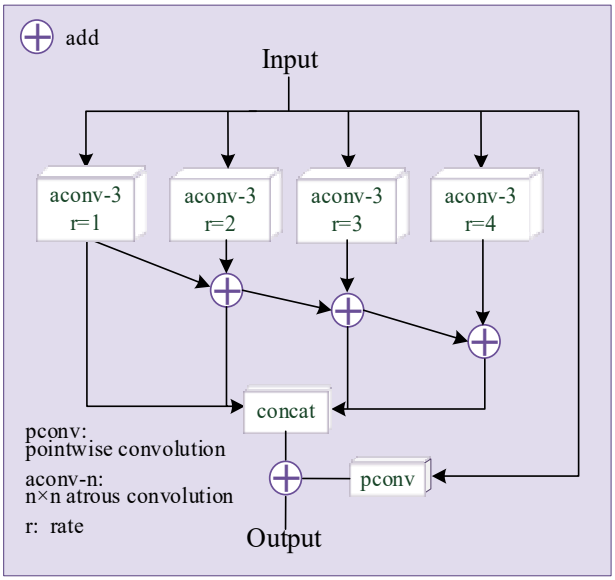

Figure 3. The structure of MFM.

3.2.2. Second-Level Auxiliary Task Network G_g

In order to better guide G_2 to learn the detailed texture information and obtain the predicted infrared image T_2 with rich detailed texture, G_g only needs to pay attention to the details of T_2. Therefore, the receptive field of G_g should be smaller. The network structure of G_g is shown in Figure 4. The upper part is the overall network structure of G_g, which contains four blocks, and the number in the middle represents the number of channels. The lower part represents the specific network structure of each block: it contains three cascaded Ghost modules, the number is the size of the convolution kernel and the convolution step length is 1. This kind of network structure makes the overall network receptive field size of G_g only 5×5, and the parameter quantity is extremely small.

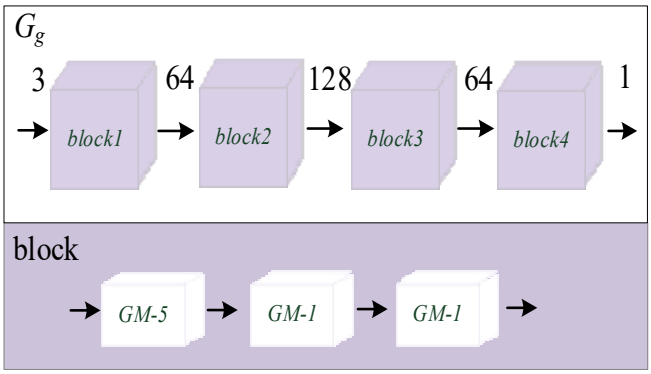

Figure 4. The proposed network of G_g.

3.3. Thrid-Level Network

To further optimize the predicted infrared image and obtain a clear edge, the third-level network adds the edge image of visible light as auxiliary information. At the same time, inspired by [36], we learn the position offset information to further obtain infrared images with sharper and more accurate edges. As shown in the green part of Figure 1, the third level has just one displacement network, G_d. The input of G_d is concatenated with the predicted infrared image T_2 of the second level network and the edge image of the visible light image. The output is the positional offset map of the input image in the row direction and the column direction. Then, the input image T_2 is resampled from the two position offset maps to obtain the final predicted infrared image T_3.

The overall network structure of G_d is shown in Figure 5a, using a codec network structure, and the encoding end includes four down-sampling residual blocks (D-Res). The specific network structure of D-Res is shown in Figure 5b. The input goes through two standard convolutions with 3 × 3 convolution kernels, and then through a bilinear interpolation down-sampling to compress the resolution of the feature map twice. Finally, the skip connection of the convolution with a convolution kernel of 4 × 4 and step size of 2 is added to the down-sampling result. The network structure of the decoding end is symmetrical to the encoding end, including four up-sampling residual blocks (U-Res). The specific network structure of U-Res is shown in Figure 5c. The input goes through two standard convolutions with 3 × 3 convolution kernels, and then through a bilinear interpolation up-sampling to double the resolution of the feature map. Finally, the deconvolution skip connection with a convolution kernel of 4 × 4 and step size of 2 is added to the up-sampling result.

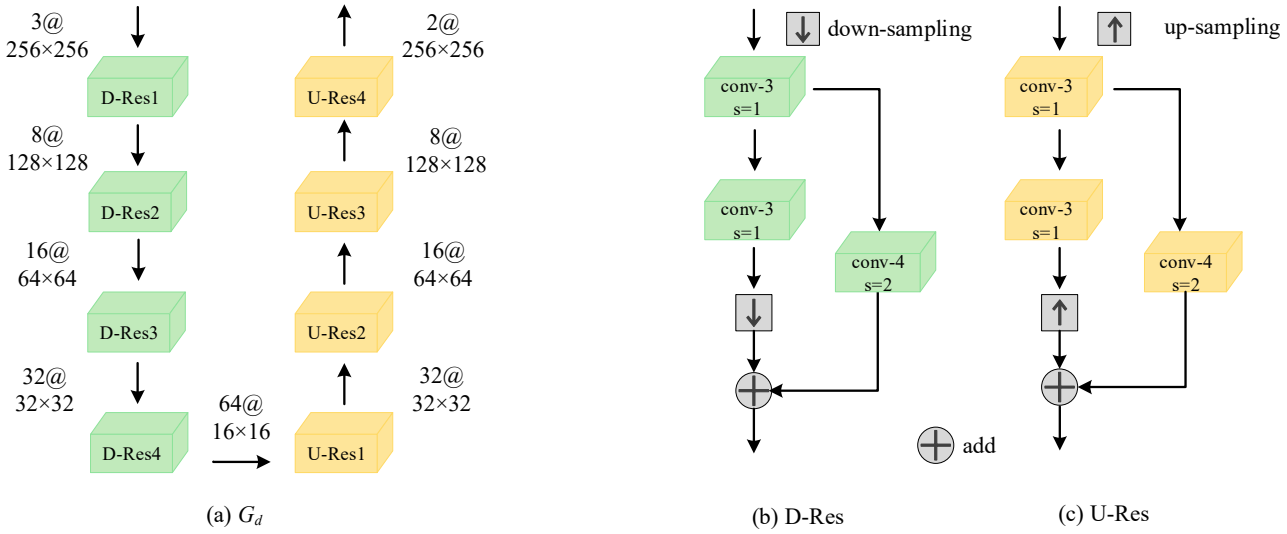

Figure 5. The proposed network of G_d.

In this paper, according to the row direction position offset map, I_{Row}, and the column direction position offset map, I_{Col}, predicted by G_d, the second-level output result T_2 is resampled to obtain the final predicted infrared image, T_3. The resampling process is defined as Equation (1). T_3 (x, y) represents the gray value of the third-level network output image at the position (x, y); and T_2 (x, y) represents the gray value of the second-level network output image at position (x, y). Row (x, y) and Col (x, y) denote the position offset in the row and column direction.

$$T_3(x, y) = T_2(x + \text{Row}(x, y), y + \text{Col}(x, y)), \quad (1)$$

3.4. Loss Function

The three-level network in this paper is jointly trained in an end-to-end manner. The gradient descent of the discriminator and the generator is performed alternately; we first fix the parameters of D_1 and D_2, train G_1, G_s, G_2, G_g and G_d, and then fix G_1, G_s, G_2, G_g and G_d, and train D_1 and D_2. The overall loss function L_{final} uses a minimum–maximum training strategy, and the expression is as follows:

$$\min_{\{G_1,G_s,G_2,G_g\}} \max_{\{D_1,D_2\}} L_{final} = L_{GAN} + L_{pixel}, \quad (2)$$

L_{GAN} is the sum of adversarial loss functions, and L_{pixel} is the sum of pixel-level loss functions. L_{GAN} includes the first-level adversarial loss, L_{GAN1}, and the second-level adversarial loss, L_{GAN2}. The expression is as follows:

$$L_{GAN} = L_{GAN1} + 10 \times L_{GAN2}. \tag{3}$$

The first-level discriminator D_1 is used to distinguish the synthetic image pair [I_{rgb}, T_1] and the real image pair [I_{rgb}, T_{true}]. The loss function adopts the combination of cross entropy, which is expressed as

$$\begin{aligned} L_{GAN1} = E_{I_{rgb},T_{true}}\left[\log D(I_{rgb}, T_{true})\right] \\ +E_{I_{rgb},T_1}\left[\log(1 - D(I_{rgb}, T_1))\right] \end{aligned} \tag{4}$$

The second-level discriminator D_2 is used to distinguish the synthetic image pair [I_{rgb}, T_2] and the real image pair [I_{rgb}, T_{true}], expressed as

$$\begin{aligned} L_{GAN2} = E_{I_{rgb},T_{true}}\left[\log D(I_{rgb}, T_{true})\right] \\ +E_{I_{rgb},T_2}\left[\log(1 - D(I_{rgb}, T_2))\right] \end{aligned} \tag{5}$$

The total pixel-level loss function L_{pixel} includes the L_1 loss function L_{G1} and L_{Gs} of the first-level generative network G_1 and Gs; the L_1 loss function, L_{G2} and LG_g, of the second-level generative network, G_2 and G_g; the gradient loss function L_{g_G2}, which is more sensitive to texture; and the L_1 loss function L_{Gd} after resampling. The expression is defined as follows:

$$\begin{aligned} L_{pixel} = \lambda_1 L_{G1} + \lambda_2 L_{Gs} + \lambda_3 L_{G2} \\ +\lambda_4 L_{Gg} + \lambda_5 L_{g\_G2} + \lambda_6 L_{Gd} \end{aligned} \tag{6}$$

λ is a hyperparameter, which represents the weight of each loss function. G_1, G_2 and G_d are the target task networks with the highest weights; the networks G_s and G_g are responsible for auxiliary tasks and have lower weights; the gradient loss function is used to increase the network's ability to perceive edges, with the smallest weights. After experiments, we finally set λ from 1 to 6 as 100, 5, 200, 10, 0.5 and 100, respectively. The L_1 loss function represents the average absolute error, expressed as

$$L_1 = \frac{1}{N}\sum_{i=1}^{N}|y_i - y_i^*|, \tag{7}$$

where i is the pixel index, N is the total number of all pixels in an image, and y_i and y_i^*, respectively, represent the real and predicted gray value at pixel i. The expression of the gradient loss function L_{g_G2} is as follows:

$$L_{g\_G2} = \frac{1}{2N}\sum_{i=1}^{2N}\left(\left|\nabla_h y_i - \nabla_h \hat{y}_i\right| + \left|\nabla_v y_i - \nabla_v \hat{y}_i\right|\right), \tag{8}$$

$\nabla_h \hat{y}_i$ and $\nabla_h y_i$ represent the gradient value in the horizontal direction at pixel i of the target infrared image T_{true} and the infrared simulation image respectively.

4. Experiments

4.1. Experimental Details and Evaluation Metrics

4.1.1. Dataset

We performed the experiments on MPD [11], which consists of image pairs for visible light images and corresponding infrared images with a resolution of 640 × 512. The training set and test set contain 50,187 and 45,141 image pairs, respectively. Both the training set and the test set involve three scenes—campus, street and suburbs—and each

scene contains images taken during the day and night. We select image pairs in the daytime as the training set of the network, and the training set size consisted of 33,399 image pairs. Correspondingly, we randomly select 565 image pairs from the daytime image pairs in the MPD test set as the test set of the network. We resize the image resolution to 256 × 256 through bilinear interpolation down-sampling.

Predicting the semantic segmentation image and gray-scale inversion image of visible light is the auxiliary task of this network. The gray-scale inversion image is obtained by converting the visible light image from a color image to a gray-scale image and then performing the gray-scale value inversion operation. Semantic segmentation images can be predicted by feeding visible light images into a model trained by Refinenet [37] on Cityscapes. Cityscapes is a large dataset mainly used for semantic segmentation. The main scene is outdoor streets, similar to MPD. The edge image of the visible light is the auxiliary information in the third-level network, which is extracted by the Canny operator with the upper and lower thresholds set to 60 and 120, respectively.

4.1.2. Evaluation Metrics

In the previous work of the image domain conversion task, there are some recognized evaluation metrics to evaluate the similarity between the network predicted image and the real target image. We used the mean absolute relative error (*Rel*), mean log10 error (*Log*10), root mean squared error (*Rms*) and accuracy index ($\delta < 1.25^i$, $i = 1, 2, 3$). The calculation expressions of each metrics are as follows:

$$Rel = \frac{1}{|N|}\sum_{i=1}^{N} |y_i - y_i^*| / y_i^*, \tag{9}$$

$$Log10 = \frac{1}{|N|}\sum_{i=1}^{N} |\lg y_i - \lg y_i^*|, \tag{10}$$

$$Rms = \sqrt{\frac{1}{|N|}\sum_{i=1}^{N} \|y_i - y_i^*\|^2}, \tag{11}$$

$$\delta = \max\left(\frac{y_i}{y_i^*}, \frac{y_i^*}{y_i}\right) < thr, \tag{12}$$

where i is the pixel index, and N is the total number of pixels in an infrared image. y_i and y_i^* respectively, represent the gray value of the target image and the gray value of the predicted image at pixel i. We also employed pixel-level similarity metrics to evaluate our method, i.e., Structural-Similarity (SSIM) and Peak Signal-to-Noise Ratio (PSNR). PSNR and SSIM, as evaluation metrics for image deblurring and super-resolution, can better reflect the similarity of the two images.

4.1.3. Training Setup

Our method was implemented with Pytorch using one NVIDIA GeForce RTX 2080 Ti GPU with 16 GB memory. We used a Gaussian distribution with a mean of 0 and a standard deviation of 0.2 for weight initialization. We minimized the loss function using the Adam optimizer with a momentum of 0.5 and initial learning rate of 0.0001. We set the batch size to 4.

4.2. Results

This section compares our method with other state-of-the-art image domain conversion methods based on generative adversarial networks. The comparison results are shown in Table 1. Pix2pix [9] is a popular cGAN that can realize image-to-image conversion and is suitable for all image domain conversion tasks. The input of the network is conditional information. In this paper, the input of pix2pix is set as a visible light image, and the output is a corresponding infrared image, without other auxiliary tasks or auxiliary information. X-Fork [38] is a GAN that realizes cross-view image translation and requires auxiliary tasks

of semantic segmentation. Selection-GAN [24] is also a GAN that realizes cross-view image translation, and its network structure is a two-level GAN, where each level of the network is guided by an auxiliary task of semantic segmentation. SEAN [39] can achieve image fusion and conversion. The style image needs to be added as auxiliary information in the process of converting the input image to the target image. In this paper, the semantic segmentation image is used as input, the GIV image is used as the style image, and the output is the predicted infrared image. LG-GAN [40] explores the generation of scenes in the local environment, and considers the global and local context at the same time, which can effectively deal with the generation of small objects and scene details.

Table 1. Comparison of the algorithms in objective metrics.

Methods	The Lower, The Better			The Higher, The Better		
	Abs Rel	Avg log10	RMS	$\delta < 1.25$	PSNR	SSIM
Pix2pix [9]	0.248	0.107	0.906	0.571	22.431	0.985
X-Fork [38]	0.314	0.130	1.074	0.480	20.692	0.984
Selection-GAN [24]	0.284	0.112	0.958	0.554	21.976	0.982
SEAN [39]	0.293	0.114	0.966	0.564	21.804	0.983
LG-GAN [40]	0.262	0.102	0.886	0.616	22.601	0.989
Ours	0.247	0.099	0.850	0.623	22.908	0.990

As can be seen from Table 1, compared with other advanced generative adversarial networks for image domain transformation, our algorithm achieves the best results on various objective evaluation metrics.

The visual comparison between our method and other advanced algorithms is shown in Figure 6. Our proposed V2T-GAN has the smallest network parameters, only 15.24 M, and the lowest error RMS. The overall parameters of the network are about 73.35%, 73.68%, 73.83%, 94.29% and 14.03% lower than the Pix2pix, X-Fork, Selection-GAN, SEAN and LG-GAN algorithms, respectively.

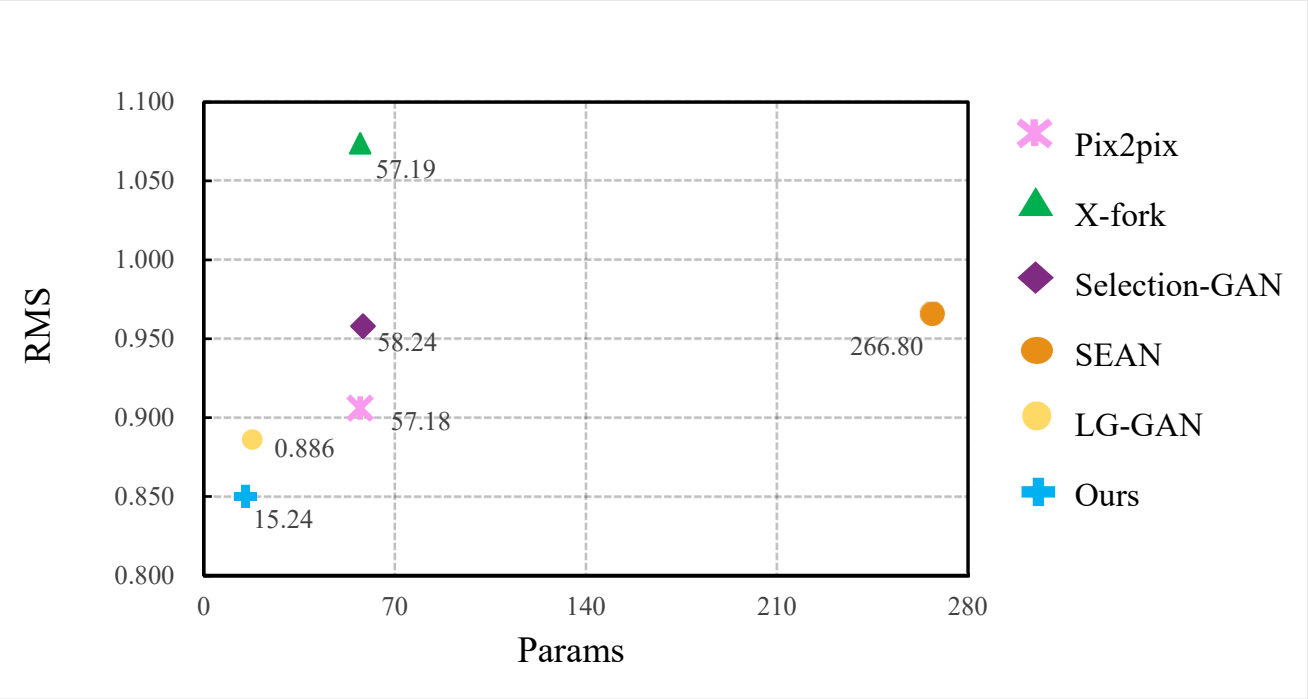

Figure 6. Compared with the computational efficiency of advanced algorithms.

4.3. Ablation Study

To further analyze the details of the proposed approach, ablation experiments were conducted by investigating different configurations of the components of V2T-Net.

4.3.1. Three-Level Network Structure

To verify the effectiveness of the three-level network structure, this section compares the experimental results of the first-level network, the two-level network and the third-level network. The comparison results are show in Table 2. We can observe the improvement in the three-level network structure in this table, which outperforms other structures in all the metrics.

Table 2. Comparison of the different network structures.

Network Structure	The Lower, The Better			The Higher, The Better		
	Rel	**Avg log10**	**RMS**	$\delta < 1.25$	**PSNR**	**SSIM**
One-level	0.254	0.100	0.859	0.617	22.838	0.988
Two-level	0.254	0.099	0.853	0.619	22.872	0.990
Three-level	0.247	0.099	0.850	0.623	22.908	0.990

The predicted infrared images of the one-level and two-level networks are shown in Figure 7. It can be seen that the results of the one-level network are relatively rough, while the results of the two-level network are more accurate in detail and more similar to the target image. For example, for the road signs selected in the first image, part of the structure is missing in the result of the first-level network, and the outline of the two-level network is more complete. The framed parts in the second image include road signs and branches. Comparing the two results, we can observe that the detailed texture information of the two-level network results is relatively more accurate.

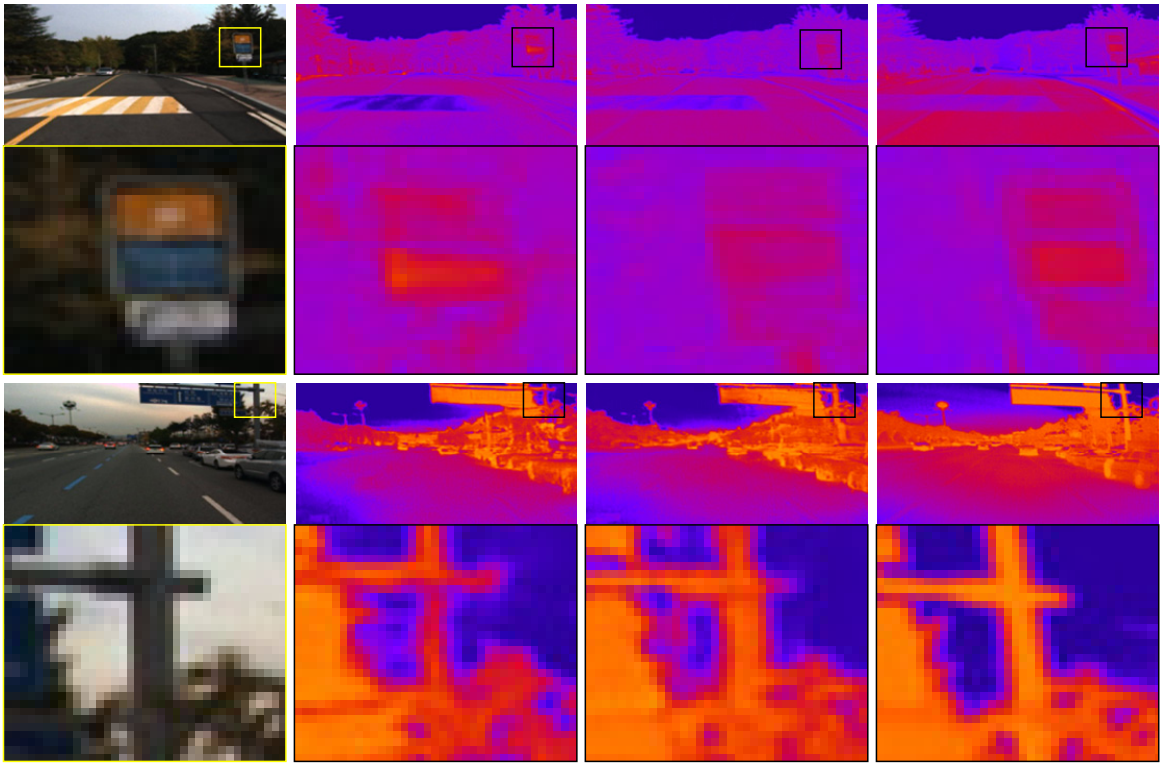

Figure 7. Comparison of the one-level and two-level networks' visualization results.

Figure 8 shows the infrared simulation images output by the two-level network and three-level network. We can see that the visualization results of the selected area in the yellow box after the position offset optimization are poor, mainly reflected in the blurred image edges, unclear textures and many errors at the edges. This is because the position offset network adopts the CNN training method; that is, it learns to convert the image directly through the pixel-level loss function. Although the converted result performs better on the pixel-level objective indicators, the subjective perception of the human eye is poor. From the perspective of the local image, it can be found from the cars selected in the second and fourth columns that the contour of the two-level network conversion result is easier to identify and more similar to the target infrared image.

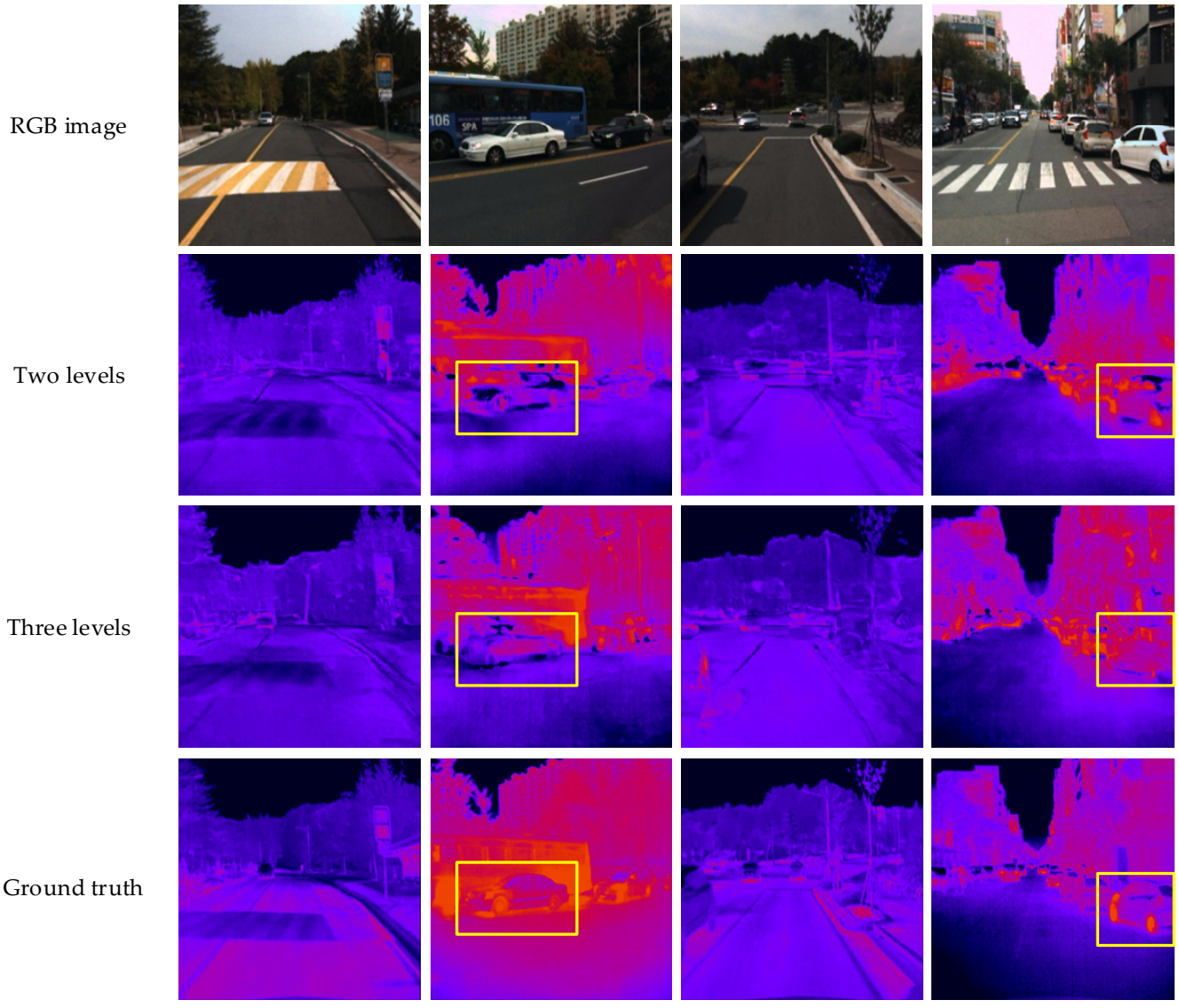

Figure 8. Comparison of two-level and three-level networks' visualization results.

4.3.2. Auxiliary Task

Auxiliary tasks are added to the method in this paper to improve network performance. In this section, we compare the effects of auxiliary tasks. The auxiliary task of the first-level network is semantic segmentation of images, and the auxiliary task of the second-level network is the GIV images. The results of specific ablation experiments are shown in Table 3: Structure 1, removing the two auxiliary tasks of semantic segmentation and GIV images at the same time; that is, removing the G_s and G_g networks. Structure 2, only

remove the GIV image, which means there is no G_g network. Structure 3, only remove the semantic segmentation image; that is, no G_s network. Row 4 represents the complete V2T-GAN.

Table 3. Comparison of the different auxiliary tasks.

Setup	The Lower, The Better			The Higher, The Better		
	Rel	**Avg log10**	**RMS**	**$\delta < 1.25$**	**PSNR**	**SSIM**
$-G_s, G_g$	0.257	0.103	0.876	0.609	22.674	0.989
$-G_g$	0.255	0.102	0.870	0.611	22.678	0.989
$-G_s$	0.249	0.101	0.855	0.615	22.811	0.990
Ours	0.247	0.099	0.850	0.623	22.908	0.990

It can be seen from Table 3 that the our complete V2T-GAN, including semantic segmentation and GIV image auxiliary tasks, obtains the best experimental results. The accuracy rate $\delta < 1.25$ is 2.30%, 1.96% and 1.30% higher than Structures 1–3, respectively. The structure one has no auxiliary tasks, and the performance is the worst. Structure 3 is better than the Structure 2 network in various objective metrics, indicating the GIV auxiliary task has a greater effect than semantic segmentation.

Although the auxiliary task of semantic segmentation in Structure 2 enables the network to learn more correct structure information, the calculation process of objective metrics cannot add weight to the structure information. The auxiliary task of GIV image in Structure 3 enables the network to obtain more detailed image information. Even if there are some differences in structure, it can still ensure better metrics. This is also a limitation of objective metrics.

4.3.3. Edge Auxiliary Information

In order to guide the third-level network to learn more clear and accurate edge information, the input of the third-level network adds the edge image of visible light as auxiliary information. We conducted an experimental analysis on the effectiveness of the edge image auxiliary information, and the results are shown in Table 4. We found that adding visible light edge images as auxiliary information can improve the objective metrics of predicting infrared images, which means that V2T-GAN has indeed learned a sharp edge from this auxiliary task.

Table 4. Effectiveness of the edge auxiliary information.

Methods	The Lower, The Better			The Higher, The Better		
	Abs Rel	**Avg log10**	**RMS**	**$\delta < 1.25$**	**PSNR**	**SSIM**
$-I_e$	0.248	0.099	0.850	0.616	22.922	0.989
Ours	0.247	0.099	0.850	0.623	22.908	0.990

4.3.4. Lightweight Convolution

To reduce the amount of overall network parameters, we generally use lightweight convolution in the proposed network. The two sub-networks with the largest amounts of parameters in V2T-GAN are G_1 and G_s. Therefore, this section compares the different lighweight methods of G_1 and G_s. The experimental results are shown in Table 5. BSConv, DSConv, GhostModule and GConv, respectively, represent the replacement of the standard convolutions in G_1 and G_s with blueprint separable convolution, depthwise separable convolution, GhostModule and group convolution. In the experiment, the grouping number of group convolution and GhostModule were both set to 4.

Table 5. Comparison of the different lightweight convolutions.

Methods	The Lower, The Better			The Higher, The Better			Params
	Abs Rel	Avg log10	RMS	$\delta < 1.25$	PSNR	SSIM	
BSConv	0.256	0.105	0.898	0.601	22.442	0.989	5.081M
DSConv	0.259	0.108	0.916	0.593	22.257	0.989	5.126M
GhostModule	0.260	0.105	0.891	0.604	22.565	0.987	15.263M
GConv	0.247	0.099	0.850	0.623	22.908	0.990	15.235M

It can be found from Table 5 that the overall network parameters using BSConv are the smallest, but the overall network using GConv performs best in various objective metrics. The goal of a lightweight network for V2T-GAN is to reduce the amount of overall network parameters, improve calculation efficiency, alleviate the problem of network overfitting and improve conversion accuracy. In order to trade off accuracy and efficiency, we finally use GConv in our network.

5. Conclusions

We propose a three-level refined lightweight GAN with cascaded guidance (V2T-GAN) to address image domain conversion task on a visible light image to the corresponding infrared simulation image. In the three-level network, semantic segmentation images, GIV images and visible light edge images were used as input information for auxiliary tasks. The experimental results on the MPD show that our method obtains much better results than the state-of-the-art on the task of feature conversion from visible light to infrared images.

In the future, we would like to be able to convert from visible light to infrared images without having to create a one-to-one mapping between the training data, as well as apply the idea of the algorithm in this paper to other fields, such as the fusion of visible light and infrared images and the detailed enhancement of infrared images.

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

Funding: This research is supported by National Natural Science Fund (No.61371143), National Key Research and Development Program Project (2020YFC0811004), Beijing Science and Technology Innovation Service capacity-basic scientific research project (No.110052971921/002), the Science and Technology Development Center for the Ministry of Education "Tiancheng Huizhi" Innovation and Education Promotion Fund (No.2018A03029), Cooperative Education Project of Higher Education Department of the Ministry of Education (No.201902083001), Science and Technology Project of Beijing Education Commission (No.KM202110009002), Hangzhou Innovation Institute of Beihang University (No. 2020-Y3-A-014).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: The authors thank the assistance from other people of the School of Information Science and Technology, North China University of Technology.

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

References

1. Kim, S. Sea-based infrared scene interpretation by background type classification and coastal region detection for small target detection. *Sensors* **2015**, *15*, 24487–24513. [CrossRef] [PubMed]
2. Mu, C.P.; Peng, M.S.; Dong, Q.X.; Gao, X.; Zhang, R.H. Infrared Image Simulation of Ground Maneuver Target and Scene Based on OGRE. *Appl. Mech. Mater.* **2014**, *716–717*, 932–935. [CrossRef]

3. Yang, M.; Li, M.; Yi, Y.; Yang, Y.; Wang, Y.; Lu, Y. Infrared simulation of ship target on the sea based on OGRE. *Laser Infrared* **2017**, *47*, 53–57.
4. Eigen, D.; Puhrsch, C.; Fergus, R. Depth map prediction from a single image using a multi-scale deep network. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; Volume 3, pp. 2366–2374.
5. Laina, I.; Rupprecht, C.; Belagiannis, V.; Tombari, F.; Navab, N. Deeper depth prediction with fully convolutional residual networks. In Proceedings of the 2016 4th International Conference on 3D Vision, 3DV 2016, Stanford, CA, USA, 25–28 October 2016; pp. 239–248.
6. Noh, H.; Hong, S.; Han, B. Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015; pp. 1520–1528.
7. Badrinarayanan, V.; Handa, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Robust Semantic Pixel-Wise Labelling. *arXiv* **2015**, arXiv:1505.07293.
8. Yin, Z.; Shi, J. GeoNet: Unsupervised Learning of Dense Depth, Optical Flow and Camera Pose. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1983–1992.
9. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 5967–5976.
10. 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. 2242–2251.
11. Hwang, S.; Park, J.; Kim, N.; Choi, Y.; Kweon, I.S. Multispectral pedestrian detection: Benchmark dataset and baseline. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1037–1045.
12. Zhou, Q.; Bai, T.; Liu, M.; Qiu, C. Near Infrared Scene Simulation Based on Visual Image. *Infrared Technol.* **2015**, *37*, 11–15.
13. Li, M.; Xu, Z.; Xie, H.; Xing, Y. Infrared Image Generation Method and Detail Modulation Based on Visible Light Images. *Infrared Technol.* **2018**, *40*, 34–38.
14. Wang, P.; Shen, X.; Lin, Z.; Cohen, S.; Price, B.; Yuille, A. Towards unified depth and semantic prediction from a single image. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2800–2809.
15. Xu, D.; Ricci, E.; Ouyang, W.; Wang, X.; Sebe, N. Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 161–169.
16. Xu, D.; Wang, W.; Tang, H.; Liu, H.; Sebe, N.; Ricci, E. Structured Attention Guided Convolutional Neural Fields for Monocular Depth Estimation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3917–3925.
17. Qi, X.; Liao, R.; Liu, Z.; Urtasun, R.; Jia, J. GeoNet: Geometric Neural Network for Joint Depth and Surface Normal Estimation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 283–291.
18. Ranjan, A.; Jampani, V.; Balles, L.; Kim, K.; Sun, D.; Wulff, J.; Black, M.J. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 12232–12241.
19. Jiao, J.; Cao, Y.; Song, Y.; Lau, R. Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; Volume 11219 LNCS, pp. 55–71.
20. Mirza, M.; Osindero, S. Conditional Generative Adversarial Nets. *arXiv* **2014**, arXiv:1411.1784.
21. Hu, L.; Zhang, Y. Facial Image Translation in Short-Wavelength Infrared and Visible Light Based on Generative Adversarial Network. *Guangxue Xuebao/Acta Opt. Sin.* **2020**, *40*, 0510001. [CrossRef]
22. Ma, S.; Fu, J.; Chen, C.W.; Mei, T. DA-GAN: Instance-Level Image Translation by Deep Attention Generative Adversarial Networks. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5657–5666.
23. Mejjati, Y.A.; Richardt, C.; Cosker, D.; Tompkin, J.; Kim, K.I. Unsupervised Attention-guided Image-to-Image Translation. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; pp. 3693–3703.
24. Tang, H.; Xu, D.; Sebe, N.; Wang, Y.; Corso, J.J.; Yan, Y. Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 2412–2421.
25. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6848–6856.
26. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.

27. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.
28. Mehta, S.; Rastegari, M.; Shapiro, L.; Hajishirzi, H. ESPNetv2: A light-weight, power efficient, and general purpose convolutional neural network. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 9182–9192.
29. Haase, D.; Amthor, M. Rethinking depthwise separable convolutions: How intra-kernel correlations lead to improved mobilenets. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 14588–14597.
30. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. GhostNet: More features from cheap operations. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 1577–1586.
31. Ronneberger, O.; Fischer, P.; Brox, T. UNet: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
32. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.
33. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6627–6638.
34. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 586–595.
35. Jia, R.; QIU, Z.; Cui, J.; Wang, Y. Deep multi-scale encoder-decoder convolutional network for blind deblurring. *J. Comput. Appl.* **2019**, *9081*, 2552–2557.
36. Ramamonjisoa, M.; Du, Y.; Lepetit, V. Predicting sharp and accurate occlusion boundaries in monocular depth estimation using displacement fields. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 14636–14645.
37. Lin, G.; Milan, A.; Shen, C.; Reid, I. RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 5168–5177.
38. Regmi, K.; Borji, A. Cross-View Image Synthesis Using Conditional GANs. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3501–3510.
39. Zhu, P.; Abdal, R.; Qin, Y.; Wonka, P. SEAN: Image synthesis with semantic region-adaptive normalization. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 5103–5112.
40. Tang, H.; Xu, D.; Yan, Y.; Torr, P.H.S.; Sebe, N. Local Class-Specific and Global Image-Level Generative Adversarial Networks for Semantic-Guided Scene Generation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; Volume 1, pp. 7867–7876.