CHAPTER 13

WGGAN: A wavelet-guided generative adversarial network for thermal image translation

Ran Zhang[a,*], Junchi Bin[a,*], Zheng Liu[a], and Erik Blasch[b]
[a]University of British Columbia, Kelowna, BC, Canada
[b]MOVEJ Analytics, Dayton, OH, United States

13.1 Introduction

Thermal or infrared (IR) images are widely used in different applications inducing night-vision navigation and surveillance [1], face recognition [2], and remote sensing [3]. IR images are produced by infrared cameras to record the thermal information of objects. These images are monochrome and are usually shown in gray scale [4]. They are different from RGB-converted gray scale images that maintain the texture information. Compared with RGB images, IR images are less affected by environmental factors such as illumination differences, fog, and smoke. However, they do not contain color and texture information, which is critical to understanding the objects in images. RGB and IR images have different characteristics and advantages for capturing the information of objects. IR images tend to catch more thermal structure, while the RGB images are more sensitive to colors. The VI and IR images can be fused [1, 5, 6] to generate more useful and comprehensive images that take advantage of both sources. However, the acquisition of RGB images in dark conditions is difficult due to the hardware limitation, which requires lights. RGB images are more easily understood by humans and play an important role in machine vision applications. Therefore, IR-to-RGB translation is needed [7].

Traditional methods of image translation require specifying color manually [8], reference image [4], or paired image datasets [9]. Manually specifying colors are able to produce vivid, colorful images according to the guide of human [8]. However, it requires a lot of labor and is a relatively time-consuming method compared with automatic methods. Reference-based methods colorize images automatically by establishing the feature connections from source to target images. They can be combined with manual intervention to improve performance. But they need the reference images which are selected manually. With the advances of convolutional neural networks (CNNs), image

[*] These authors contributed equally to this work.

Generative Adversarial Networks for Image-to-Image Translation
https://doi.org/10.1016/B978-0-12-823519-5.00015-4

translation can be fully automated. CNNs extract both low-level and semantic features. The object can be localized and colorized when CNNs learn semantic information [10]. CNN-based translation methods are trained on a large number of images that can cover various objects in different situations. During the testing process, these methods are fully automatic and do not need reference images. Although the CNN-based translation has high accuracy, all these methods need to have fully paired images to learn the direct mapping between IR and RGB images. The acquisition of paired images is challenging due to the difficulty of hardware calibration in industrial applications. A generative adversarial network (GAN) is proposed to address the unpaired image translation by including discriminators during training generative models. Recent state-of-the-art image translation methods are proposed based on GAN [11–13].

However, when transferring IR images to RGB images, contemporary GANs are unable to keep the structure of the object or produce clear texture information. In this study, we proposed a wavelet-guided generative adversarial network (WGGAN) to address the challenge. Similar to contemporary methods, the WGGAN is also comprised of an autoencoder for image translation and a discriminator for training. To deal with the spatial distortion problem, we combine discrete wavelet transformation and variational autoencoder to keep structural information in the early stages of the network. It brings clear synthetic RGB images, as shown in Fig. 13.1. Contrarily, both qualitative and quantitative analyses are implemented to evaluate the proposed method's performance. Compared with novel methods, the proposed method has more promising results in both results. To conclude, our contributions are as follows:

- Design combining the discrete wavelet transformation and variational autoencoder for IR-to-RGB image translation, which improves both qualitative and quantitative analyses. To the best of our knowledge, this is the first method to adopt discrete wavelet transformation in GAN-based methods of IR-to-RGB translation.
- Robust performance as the WGGAN does not require paired IR and RGB image datasets facilitating the thermal image translation when paired images are not available.

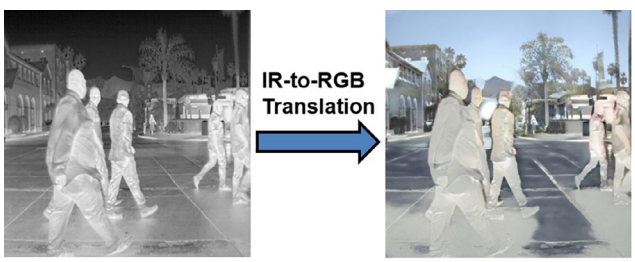

Fig. 13.1 An example of IR-to-RGB translation via the proposed wavelet-guided generative adversarial network (WGGAN).

The rest of the chapter is organized as follows. Section 13.2 introduces the progress in infrared image translation and relevant GANs' applications in image translation. Section 13.3 introduces the proposed WGGAN from the overall architecture to details of implements. Section 13.4 presents the experimental setup and results compared with contemporary methods. Finally, Section 13.5 concludes the experiments and WGGAN.

13.2 Related work

13.2.1 Infrared image translation

Infrared image translation aims to transfer single-channel gray scale images to multichannel RGB images that contain color and texture information. It can be divided into scribble-based [10, 14–16], reference-based [17–19], and fully automatic [20, 21] methods. Scribble-based methods assume that adjacent areas have a similar color. The scribbles can be added by human intervention or edge detection algorithm. Reference-based methods rely on reference images that have a structure similar to the source image. The reference images can be selected automatically by feature matching. Then the IR images can be transferred to color images by image analogy [18]. Fully automatic methods usually utilize the CNNs [22] or GANs [22, 23] to extract features and automatically establish the pixel-wise mapping from source images to target images. The IR images can be transferred directly to RGB images without manual intervention or reference images. However, they are usually supervised methods and require that the training dataset be paired, that is, the IR images should have corresponding calibrated RGB images. In most scenarios, it is hard to obtain paired datasets. Our proposed WGGAN only needs the unpaired IR and RGB images for training, and gaining significant practical values comparing with other fully automatic methods.

13.2.2 GANs in image translation

Transferring IR images to RGB images can be considered as a specific application of image translation. Much research has been conducted in this field. Image translation focuses on transferring the style of the image from one domain to another domain. Depending on whether the dataset is paired or unpaired, the image translation can be divided into paired or unpaired ways. Paired data and unpaired data can also be utilized to train a model [24], thus, obtaining the advantages of both paired and unpaired methods. The training image translation model with paired data can lead to better performance despite that the paired data are not easy to collect and calibrate. Conditional GAN [25] is used in the pixel-to-pixel level paired image-to-image translation. It showed great performance on the paired datasets. Unpaired image translation methods are more widely exploited as they have fewer limitations on the datasets. These methods usually contain more than one autoencoder, which generates target images first and reconstructs

images from the target domain to the source domain. The reconstructed images should be similar or consistent with the source image in this process. CycleGAN [11] introduces cycle consistency loss to keep the reconstructed images similar to the source image. Resembling cycle consistency loss, reconstructed loss is designed in DiscoGAN [26] to keep the similarities between original and reconstructed images. DualGAN [27] adopts dual learning to GANs and learns to transfer images between two domains. UNIT [28] makes a shared-latent space assumption for transferring images. UGATIT [12] produces attention masks and fuses the generation output with the attention mechanism to generate higher quality target images. Apart from the above unimodal image translation problem, which is limited between two domains, multimodal image translation using a single model in unpaired datasets is more challenging. StarGAN [29] performs image-to-image translation for multiple domains. StarGAN v2 [30] enhances the performance by introducing the mapping network and style encoder. MUNIT [13] is another multimodal translation model which assumes that the image composition can be decomposed into a domain-invariant content code and a domain-specific style code. However, an empirical study reveals that these methods may have strong spatial distortion during paired thermal image translation [7]. Moreover, the quantitative results also indicate the unsatisfactory of translated images of both CycleGAN and UNIT. To address the problem of deformed translation, our proposed WGGAN aims to preserve structural information by adopting discrete wavelet transformation to variational autoencoder for unpaired thermal image translation.

13.3 Wavelet-guided generative adversarial network

13.3.1 Overall architecture

The proposed wavelet-guided generative adversarial network (WGGAN) is designed for converting images from the IR domain to the RGB domain. Fig. 13.2 shows the overall

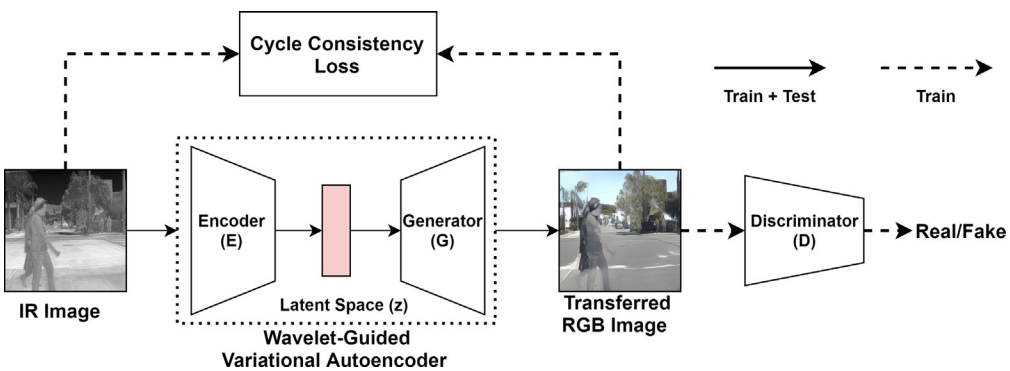

Fig. 13.2 The architecture of the proposed WGGAN.

architecture of the proposed method. The overall process consists of training and testing stages. The training stage includes a proposed wavelet-guided variational autoencoder (WGVA), a discriminator, and a cycle consistency loss. The proposed WGVA aims to generate RGB images from real IR images, while a discriminator aims to recognize the transferred RGB images from real RGB images. Moreover, the cycle consistency loss [11] is implemented to train the WGVA due to the unaligned IR-RGB image pairs. After the training stage, the WGVA can be launched to translate IR images as a standalone model without the other components during the testing stage.

13.3.2 Wavelet-guided variational autoencoder

The proposed wavelet-guided variational autoencoder (WGVA) is developed based on deep variational autoencoder and discrete wavelet transformation (DWT) [31]. As shown in Fig. 13.3, the WGVA consists of two subnetworks: an encoder E for converting IR image to latent space z, and a generator G to reconstruct the RGB image from z. Like MUNIT, StarGAN, and CycleGAN [13, 29], the design of WGVA follows standard residual autoencoder architecture with two convolutional layers, two pooling layers, and four residual blocks at the encoder. Moreover, the architecture of the generator is symmetrically reversed to the encoder. The details of residual autoencoder can be found in Ref. [28]. Unlike standard residual autoencoder, the latent space z is reparameterized to variational distribution, which enables smooth and accurate generation [13]. On the other hand, inspired by DWT, wavelet pooling, and wavelet unpooling are designed to substitute conventional pooling layers in standard residual autoencoder. Then, the high-frequency components skip bridge the corresponding pooling and unpooling layers for improving generative resolution, as illustrated in Fig. 13.3. The details of reparameterization and wavelet pooling are introduced in the following sections.

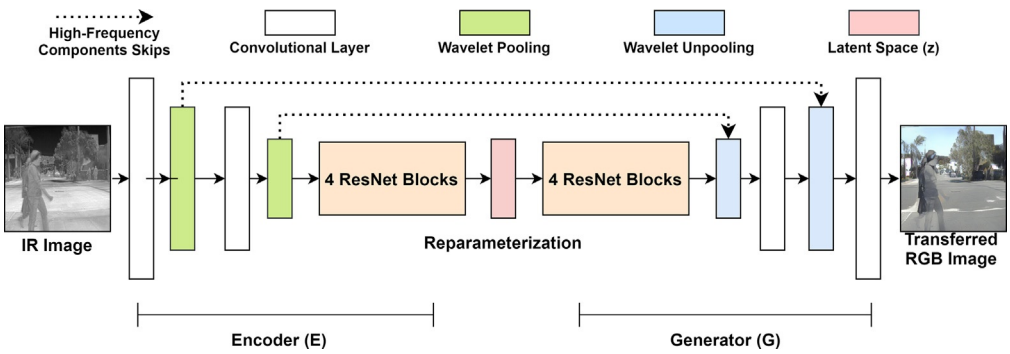

Fig. 13.3 The illustration of wavelet-guided variational autoencoder (WGVA).

13.3.2.1 Reparameterization in latent space

The encoder-generator pair $\{E, G\}$ constitutes the WGVA for IR-to-RGB translation as shown in Fig. 13.3. The latent space z is the representation of an input IR image x. According to the theoretical study, the z should represent the variational distribution of data to have smooth generation [13]. For achieving this purpose, the reparameterization is implemented as follows:

$$z \sim q(z|\, x) = \mathcal{N}\left(z; \mu, \sigma^2 I\right) \tag{13.1}$$

$$z = \mu + \sigma \odot \epsilon, \quad \text{where } \epsilon \sim \mathcal{N}(0, \sigma^2 I) \tag{13.2}$$

where $\odot$ refers to element-wise product; $\mathcal{N}(.)$ represents the normal distribution; μ is the mean of z; σ is the standard deviation of z; I is the identity matrix; and ϵ is the normal noise. Both σ and μ are learnable variables in WGVA. In other words, the σ and μ can be regarded as approximated mean and standard deviation of the entire dataset after training. Through the above equations, the latent space z is standardized to the intended distribution with the σ and μ. On the other hand, the normal noise ϵ is added to smooth the z for stochastic optimization [13]. Therefore, the latent space z can be regarded to the variational distribution of input image after reparameterization with learned σ and μ.

13.3.2.2 Discrete wavelet transformation for pooling

According to the empirical studies [13, 31], the standard residual blocks may generate distorted and blurred images. The major reason is that the generator is lack of structural information from the encoder. To address the issue, we adopt the discrete wavelet transformation (DWT) to extract structural information at pooling layers of the proposed method [31].

Discrete wavelet transformation (DWT) has four kernels, $\{\mathrm{LL}^\top, \mathrm{LH}^\top, \mathrm{HL}^\top, \mathrm{HH}^\top\}$, where the low (L) and high (H) pass filters are

$$\mathrm{L}^\top = \frac{1}{\sqrt{2}}[1, 1], \quad \mathrm{H}^\top = \frac{1}{\sqrt{2}}[-1, 1] \tag{13.3}$$

Thus, the DWT can generate four types of output denoted as LL, LH, HL, HH, respectively. Fig. 13.4 shows the examples after DWT. The output of LL has the smooth texture of the images, while the rest of the outputs capture the vertical, horizontal, and diagonal edges [31]. For simplicity, we denote the output of LL as low-frequency components and outputs of LH, HL, and HH as high-frequency components. The DWT enables the proposed model to control the IR-to-RGB conversion by different components separately. Specifically, the low-frequency component can affect the overall generative texture, while the high-frequency components affect the generative structure. Without processing the generative network's high-frequency components, the structural information can be well maintained in these components. From this point of view, the

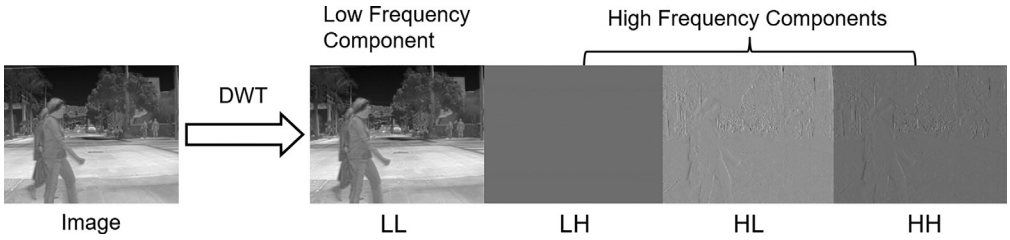

Fig. 13.4 The illustration of discrete wavelet transformation (DWT).

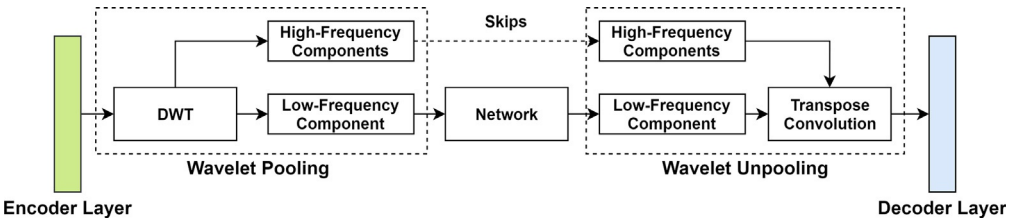

Fig. 13.5 The illustration of wavelet pooling and wavelet unpooling.

wavelet pooling and wavelet unpooling are proposed to use these components in auto-encoder for better IR-to-RGB translation, as shown in Fig. 13.5 [31].

The wavelet pooling applies DWT to the encoder layer to have low frequency and high-frequency components. The kernels of the convolutional layer are changed to DWT kernels to apply the DWT in the deep neural layer. On the other hand, the wavelet pooling layer is locked during the optimization. Moreover, the stride of the layer is 2 to have downsampling features as same as conventional pooling layers [31]. The low-frequency component will be further processed by the network. Meanwhile, the high-frequency components skipped to the symmetrical wavelet unpooling layer in the generator. In the wavelet unpooling layer of the generator, both high-frequency and low-frequency components are concatenated. Then, the concatenated components are processed to have upsampling features by transpose convolution [31].

13.3.3 Objective functions in adversarial training

The full objective of the WGGAN comprises four loss functions: cycle-consistency loss, ELBO loss, perceptual loss, and GAN loss [11, 13, 32].

Cycle-consistency loss. To train the proposed method with unpaired RGB and IR images, we adopt the cycle consistency loss which is similar to MUNIT and CycleGAN [11]. The basic idea of cycle consistency loss aims to include two generative networks for constraining the generative images. Two generative adversarial networks: $\text{GAN}_1 = \{E_1, G_1, D_1\}$

for IR-to-RGB translation and $GAN_2 = \{E_2, G_2, D_2\}$ for RGB-to-IR translation are used in training, where E, G, and D denote encoder, generator, and discriminator, respectively. For simplicity, $E(x) = z$ indicates the latent space z is generated by encoder E. The theory of the loss is that the image translation cycle should be capable of bringing converted images back to original images, i.e., $x \rightarrow E(x) \rightarrow G_1(z) \rightarrow G_2G_1(z) \approx x$. The cycle-consistency loss is shown as below:

$$\mathcal{L}_{CC}(E_1, G_1, E_2, G_2) = \mathbb{E}_{x_1 \sim p(x_1)}[\| G_2(G_1(z_1)) - x_1 \|] + \mathbb{E}_{x_2 \sim p(x_2)}[\| G_1(G_2(z_2)) - x_2 \|] \tag{13.4}$$

where $\| \cdot \|$ represents the ℓ_1 distance.

ELBO loss. ELBO aims to minimize the variational upper bound of latent space z. The objective function is

$$\mathcal{L}_E(E, G) = \lambda_1 \mathrm{KL}\big(q(z|x) \| p_\eta(z)\big) - \lambda_2 \mathbb{E}_{z \sim q(z|x)} \log p_G(x|z) \tag{13.5}$$

$$\mathrm{KL}(P||Q) = \sum P(x) \log \frac{P(x)}{Q(x)} \tag{13.6}$$

where the hyperparameters λ_1 and λ_2 control the weights of the objective terms and the KL divergence terms that penalize deviation of the distribution of latent space from the prior distribution $p_\eta(z)$. Here $q(.)$ represents the reparameterization mentioned in the previous section, while $p_\eta(z)$ represents zero-mean Gaussian distribution. $p_G(.)$ is the Laplacian distribution based on generator according to empirical studies [13].

Perceptual loss. Perceptual loss is a conventional loss function of neural style transfer with the assistant of pretrained VGG-16 [33] as shown in the following equation. The perceptual loss consists of two parts: the first term loss is content loss and the second term is style loss.

$$\mathcal{L}_P(E, G, x_c, x_s) = \frac{1}{C_j H_j W_j} \| \phi_j(G(z)) - \phi_j(x_c) \|_2^2 + \| \frac{1}{C_j H_j W_j} \left[\mathrm{Gr}\Big(\phi_j(G(z))\Big) - \mathrm{Gr}\Big(\phi_j(x_s)\Big) \right] \|_2^2 \tag{13.7}$$

where $\phi_j(x)$ represents the feature map of jth convolutional layers of shape $C_j \times H_j \times W_j$ in pretrained VGG-16; x_c denotes content images while x_s denotes style images; Gr is the Gram matrix which is used for representing image style. From this point of view, the perceptual loss aims to transfer the style of images while maintaining image structure. The details of the perceptual loss can be found in Ref. [32].

GAN loss. GAN loss aims to ensure the translated images resembling the images in the target domains, respectively [14]. For example, if the discriminator regards the synthetic IR images as real IR images, the synthetic IR images are successful.

$$\mathcal{L}_{\text{GAN}}(E, G, D) = \mathbb{E}_{x\sim p(x)} \log D(x) + \mathbb{E}_{z\sim p(z|x)}[\log 1 - D(G(z))] \tag{13.8}$$

Full loss. Finally, the complete loss function can be written as:

$$\begin{aligned}\mathcal{L}_{\text{total}} = {} & \lambda_E(\mathcal{L}_E(E, G)) + \lambda_P(\mathcal{L}_P(E_1, G_1, x_1, x_2)) + \lambda_{\text{GAN}}(\mathcal{L}_{\text{GAN}}(E, G, D)) \\ & + \lambda_{\text{CC}}(\mathcal{L}_{\text{CC}}(E_1, G_1, E_2, G_2))\end{aligned} \tag{13.9}$$

where $\lambda_E = 0.1$, $\lambda_P = 0.1$, $\lambda_{\text{GAN}} = 1$, and $\lambda_{\text{CC}} = 10$ represent the weights of ELBO loss, perceptual loss, GAN loss, and cycle consistency loss, respectively.

13.4 Experiments

This section presents the details of the experiments. First, the section describes the implemented dataset and evaluation methods in this experiment. Then, the baselines and the relevant experimental setup are also introduced. Finally, both qualitative and quantitative analyses of translation results are presented.

13.4.1 Data description

The implemented dataset is FLIR ADAS [34], which is an open dataset for autonomous driving. The dataset contains RGB and IR images from the same driving car. However, the recorded RGB and IR images are unpaired due to the cameras' different properties [34]. For all experiments, the training and testing splits follow the dataset benchmark. The training dataset contains 8862 IR images and RGB images, while there are 1363 IR images and 1257 RGB images in the testing dataset. The statistics of the FLIR ADAS are presented in Table 13.1.

13.4.2 Evaluation methods

Qualitative analysis. In researches on generative models, the human perceptual study is a direct way to compare the quality of translation among models. In this study, several graduate students and computer vision engineers are invited to subjectively evaluate the translated results from the proposed method and baseline methods with source IR image. They are then required to select which output has the best quality with short comments.

Table 13.1 Statistics in FLIR ADAS.

Dataset	Image type	# of frames	Image size
Training	IR	8862	640×512
	RGB	8363	1800×1600
Testing	IR	1363	640×512
	RGB	1257	1800×1600

Quantitative analysis. Numerical evaluating the proposed WGGAN and other comparative methods is challenging since there are no paired RGB images. To measure the translated quality, we include four Inception-based metrics: 1-Nearest Neighbor classifier (1-NN), kernel maximum mean discrepancy (KMMD), Fréchet inception distance (FID), and Wasserstein distance (WD) [35]. These methods compute the distance between features in the target and generated images from the Inception network [35]. If an IR image is well translated, these metrics will have small values, which indicates the generated RGB image is similar to the distribution of target RGB images. Besides, two no-reference image quality assessment (NR-IQA) methods, blind/referenceless image spatial quality evaluator (BRISQUE) [36], and natural image quality evaluator (NIQE) [37], are also used to independently evaluate the generated images without any pair or unpaired images. The small values of them indicate the high quality of the translated images. Moreover, a multicriteria decision analysis, TOPSIS [38], is included to summarize all the quantitative evaluation metrics.

13.4.3 Baselines

CycleGAN. CycleGAN consists of two standard residual autoencoders for training with GAN loss and cycle-consistency loss [11].

MUNIT. MUNIT is similar to CycleGAN, which also consists of two autoencoders. For having diverse generative images, the encoder of MUNIT has one content branch and a style branch. Inspired by neural style transfer [13], where the two branches are combined by adaptive instance normalization in the generator for image reconstruction.

StarGAN. StarGAN is a state-of-the-art generative method in facial attribute transfer and facial expression synthesis. It includes mask vector and domain classification to generate diverse output [29].

UGATIT. The UGATIT adopts an attention mechanism to residual autoencoder with auxiliary classifier inspired by weakly supervised learning. Moreover, it also introduces an adaptive instance normalization to the residual generator [12]. It achieves novel performance in tasks of anime translation and style transfer.

13.4.4 Experimental setup

The adaptive moment optimization (ADAM) [12] is used as an optimizer for training the proposed method where the learning rate is set to 0.00001, and momentums are set to 0.5 and 0.99. For improving the model's robustness, the batch size was set as 1 with instance normalization after each neural layer. The discriminator is adopted from PatchGAN [11]. Moreover, all the activation functions of neurons are set to the rectified linear unit (ReLU), while the activation function of the output layer is Tanh to generate synthetic images.

To make a fair comparison, both WGGAN and baseline models are trained in 27 epochs with batch size 1. On the other hand, all images are resized to 512×512 before feeding to the network. A desktop with an NVIDIA TITAN RTX, and Intel Core i7 and 64 GB memory is used throughout the experiments.

13.4.5 Translation results

This subsection aims to present both qualitative and quantitative analyses of translation results compared with baselines. In qualitative analysis, the examples of translation are illustrated with subjective comments. On the other hand, the quantitative analysis presents numerical results to compare the proposed WGGAN with baselines.

13.4.5.1 Qualitative analysis

Fig. 13.6 illustrates the translation results in the test set of FLIR ADAS. StarGAN has the worst translation performance, which has unsuitable colors and noisy black spots on the images. The rest of the translated images can generate clear edges of solid objects like vehicles. However, the UGATIT is not capable of clearly translating objects such as trees and houses. Compared with WGGAN, CycleGAN, and MUNIT, participants also point out that the road texture is not well translated by UGATIT, as shown in Fig. 13.6. The texture of the road is too smooth to present the details, such as the curb on the road. On the other hand, CycleGAN can accurately translate the objects from IR images with sharp edges and textures. Several participants also mention that there are some incorrect mapping objects on the generated RGB images from CycleGAN. For example, trees should not appear in the sky in Fig. 13.6B. On the contrary, both the proposed WGGAN and MUNIT are able to translate the IR images with clear texture information of objects. However, some parts of the images are not correctly translated, such as the sky and people, as shown in Fig. 13.6C. In qualitative evaluation, participants indicate that the proposed WGGAN can generate the best quality of images with clear texture and correct mapping objects. Compared with other state-of-the-art methods, the generated images are less scattered noises. To conclude, participants believe that the proposed WGGAN has the best performance in IR-to-RGB translation.

13.4.5.2 Quantitative analysis

Table 13.2 illustrates the quantitative results of the IR-to-RGB translation. The best result of each evaluation method is highlighted. It is difficult to identify the best method within contemporary methods. CycleGAN achieves excellent performance in NR-IQA evaluation, while MUNIT has better performance in 1-NN and KMMD. Unlike contemporary methods, the proposed WGGAN outperforms all the contemporary models with the smallest values in 1-NN, KMMD, FID, and NIQE. For inception-based metrics, WGGAN has 26.1% and 53.1% improvement in KMMD and FID, which means that the generative RGB images are similar to the target RGB domain. On the other

Fig. 13.6 Examples of (A) source IR images, (B) proposed WGGAN, (C) CycleGAN, (D) MUNIT, (E) StarGAN, and (F) UGATIT.

Table 13.2 Quantitative results of contemporary methods.

Models	1-NN	KMMD	FID	WD	BRISQUE	NIQE
CycleGAN	0.961	0.318	0.222	**61.60**	**15.17**	2.730
MUNIT	0.927	0.237	0.157	67.50	27.99	2.750
StarGAN	0.992	0.397	0.121	75.73	36.55	6.392
UGATIT	0.959	0.283	0.098	65.88	36.81	3.663
WGGAN	**0.924**	**0.175**	**0.046**	65.97	28.89	**2.477**

Table 13.3 Ranking results by TOPSIS based on quantitative result.

	CycleGAN	MUNIT	StarGAN	UGATIT	WGGAN
TOPSIS	0.479	0.569	0.313	0.559	**0.796**

hand, the WGGAN also achieves the best performance in NIQE, which indicates the image is similar to the natural images in terms of statistical regularities. To better conclude the best model with these evaluating methods, we use a common multicriteria decision model, TOPSIS, to rank the image translation models based on evaluating methods. Table 13.3 shows that WGGAN has the highest values of TOPSIS. The ranking results also demonstrate that the proposed WGGAN can generate more high-quality RGB images in comparison with other novel image translation methods.

13.5 Conclusion

In this chapter, an IR-to-RGB image translation method, wavelet-guided generative adversarial network (WGGAN), is proposed for context enhancement. A wavelet-guided variational autoencoder (WGVA) is proposed for generating smooth and clear RGB images from the IR domain, which combines variational inference and discrete wavelet transformation. In addition, more objective functions are introduced to improve generative quality, such as ELBO loss and perceptual loss. Both qualitative and quantitative results demonstrate the effectiveness of the proposed WGGAN to enable better context enhancement for IR-to-RGB translation. Many industrial applications can benefit from the proposed method, such as object detection at night for applications of semi-autonomous driving, unmanned aerial vehicle (UAV) surveillance, and urban security.

Although the proposed WGGAN has promising results in thermal image translation, there is still room for improvement. For example, the objects' colors should be more discriminative from the background. Therefore, we first aim to add numerous IR and RGB images for fully training the proposed WGGAN. Then, more advanced modules, such as adaptive instance normalization, will be included to enhance the translation.