remote sensing

MDPI

Article

DGC_GAN: An Unpaired Method for Cross-Spectral Image Translation from Visible to Thermal Infrared

Shun Yao [1,2,3], Xiaobing Sun [1,3], Bo Song [1,3], Yichen Wei [1,2,3], Yuyao Wang [1,2,3], Yiqi Li [1,2,3] and Xiao Liu [1,3,*]

[1] Anhui Institute of Optics and Fine Mechanics, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei 230031, China; yaoshuninhere@mail.ustc.edu.cn (S.Y.); xbsun@aiofm.ac.cn (X.S.); songb@aiofm.ac.cn (B.S.); weiyc@mail.ustc.edu.cn (Y.W.); wangyuyuao@mail.ustc.edu.cn (Y.W.); lyq9710@mail.ustc.edu.cn (Y.L.)
[2] Science Island Branch of Graduate School, University of Science and Technology of China, Hefei 230026, China
[3] Key Laboratory of Optical Calibration and Characterization, Anhui Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Hefei 230031, China
* Correspondence: liux@aiofm.ac.cn

Highlights

What are the main findings?

- A dual geometric cycle framework is proposed to jointly enforce geometric consistency and cycle consistency for unpaired visible-to-thermal infrared translation in drone scenes.
- Disentangled structure–style representations enable one-to-many synthesis and improve the realism–diversity trade-off on the AVIID and Drone Vehicle datasets.

What are the implications of the main findings?

- The proposed framework provides a practical way to expand drone thermal infrared datasets when real thermal infrared images are scarce or sensitive.
- It has the potential to benefit downstream drone thermal remote sensing tasks by providing diverse and realistic synthetic thermal infrared imagery.

Abstract

Thermal infrared imaging is widely used in applications such as disaster monitoring and target recognition because it remains stable under illumination changes and supports nighttime observation. However, thermal infrared data are expensive to acquire, and the related application scenarios are often sensitive, which leads to limited publicly available thermal infrared datasets and restricts the development of relevant research. Cross-spectral image translation from visible to thermal infrared provides a solution for expanding infrared datasets, but accurate mapping remains difficult because visible light reflection and thermal infrared emission follow different physical mechanisms. This paper proposes a Dual Geometric Cycle Generative Adversarial Network (DGC_GAN), for unpaired visible-to-thermal infrared translation. The proposed method improves cross-spectral mapping accuracy by combining geometric-consistency constraints with cycle-consistency constraints. In addition, disentangled representation learning is introduced to decompose cross-spectral images into a domain-invariant semantic structure space and a domain-specific imaging style space, enabling one-to-many synthesis through the cross-combination of structure and style. Experiments on public aerial datasets, including AVIID and Drone Vehicle, demonstrate that DGC_GAN significantly improves the realism and diversity of generated images compared with other popular unpaired translation methods. Specifically, DGC_GAN achieves FID and KID values of 63.727 and 0.008711 on the Day Road dataset (part of Drone Vehicle), 69.419 and 0.019352 on the Night Road dataset (part of Drone

Check for updates

Received: 4 January 2026
Revised: 3 February 2026
Accepted: 9 February 2026
Published: 11 February 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Vehicle). Moreover, it outperforms other methods on all four evaluation metrics on the AVIID dataset. Furthermore, real drone data collected using a dual-spectrum platform are used to validate the practical usefulness of the proposed method. We also collected real data using a dual-spectrum drone platform to verify the practical usefulness of the proposed method.

Keywords: visible image; thermal infrared image; image translation; GAN

1. Introduction

Thermal infrared imaging has been widely used in fields such as intelligence reconnaissance, disaster monitoring, target tracking, and recognition due to its strong resistance to interference and its capability for nighttime imaging, and it has demonstrated irreplaceable advantages in these applications [1–3]. Thermal infrared imaging is often deployed in highly sensitive application domains, which directly results in a scarcity of publicly available datasets. The existing datasets are frequently inadequate for research needs across diverse and complex task scenarios, thereby constraining downstream application studies based on thermal infrared images [4,5]. In contrast, visible images are less expensive to acquire, and the corresponding datasets are easier to obtain. Therefore, cross-spectral image translation methods that map visible images to thermal infrared images can effectively broaden the channels for acquiring thermal infrared data, which in turn helps promote the development of thermal infrared–related research and industry [6–8].

The spectral ranges of visible and thermal infrared are shown in Figure 1. Visible imaging mainly relies on reflected light signals from the surface of a target, whereas thermal infrared imaging primarily depends on the thermal radiation emitted by the object itself. The two modalities differ fundamentally in their physical imaging mechanisms, which makes cross-spectral image translation a technically challenging task [9,10]. Research on generating corresponding thermal infrared images from visible images has been conducted for a long time. Traditional thermal infrared image generation typically involves land-cover material classification, computing surface temperature based on first principles, estimating radiance, and then producing the thermal infrared image [11–13]. In the area of thermal infrared image generation using traditional methods, researchers have carried out a series of studies. Early work used a one-dimensional thermal model to compute surface temperatures for objects with simple geometric shapes [14], marking the starting point of this research line. Subsequent studies incorporated the effects of solar and sky radiation, mass-transfer processes, fluid transport, and multiple reflections, leading to the development of the thermal infrared scene image generation system GTVISIT [15]. Other work analyzed the thermal properties of a scene and combined them with texture characteristics from real images to simulate thermal infrared images under specific weather conditions [16]. In addition, studies on infrared radiation interactions between targets and backgrounds have noted that mutual influences should be considered when estimating target and background temperatures to improve the realism of generated thermal infrared images [17]. Although traditional methods are supported by solid physical principles, they often require substantial manual effort, and the texture details of land cover are difficult to reproduce in the final simulated results, which limits scalability in complex scenes [18–20].

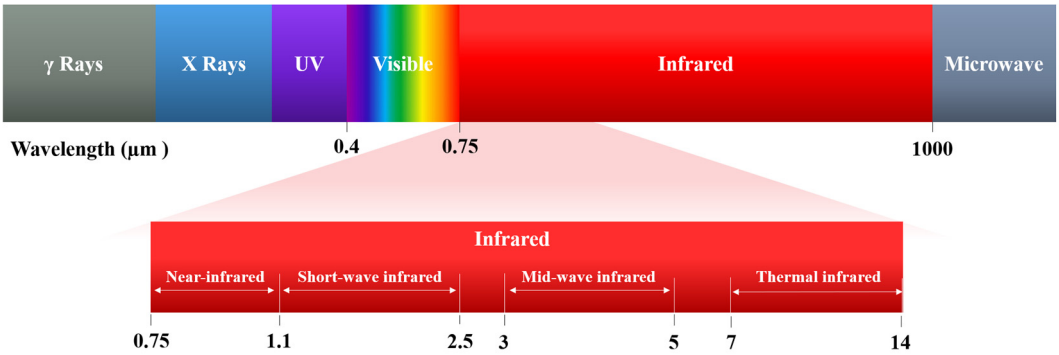

Figure 1. Spectral band ranges.

With the advancement of deep learning, it has also been applied to visible-to-thermal infrared cross-spectral image translation. On the one hand, building on traditional pipelines, researchers use deep learning to perform land-cover material classification from visible images, which greatly improves efficiency and reduces the workload of generating thermal infrared images using traditional methods. For example, some studies use an improved DeepLab V3 segmentation model to classify materials in visible images and then combine the classification results with first principles to simulate temperature and finally generate thermal infrared images [21]. On the other hand, deep learning has achieved major progress in image translation, image generation, and cross-modal learning [22–24]. In particular, deep learning-based image translation has shown a strong capability in information extraction and feature coupling. By developing deep learning-based translation methods between visible and thermal infrared images, researchers can generate realistic infrared images quickly and at a low cost [25–27]. Given the unique characteristics of remote sensing scenarios, related studies have focused on targeted optimization for cross-modal generation. For example, in visible-to-infrared generation for remote sensing, reference [3] emphasizes using thermal features to guide the translation process, so that the generated infrared images better match the true distribution of thermal information. Reference [28] proposes the PID model, namely Physics Informed Diffusion, and introduces a strong physics-constrained loss based on infrared physical priors into the training of diffusion models. This design effectively improves the consistency between generated infrared images and the real infrared domain, further strengthening the methodological framework for visible-to-thermal infrared spectral generation. In addition, the MappingFormer model proposed in [20] decomposes a visible image into low-frequency and high-frequency components and uses a two-branch architecture to perform low-frequency feature mapping, which preserves image structure and coarse outlines, and high-frequency feature refinement, which improves edge and texture details. As a result, the generated infrared images better follow the patterns of thermal features.

From this perspective, deep learning-based visible-to-infrared translation is essentially a mapping between two different image domains [29,30]. Early studies in computer vision have shown that generative adversarial networks, or GANs, are effective tools for learning such cross-domain mappings. Many GAN-based cross-spectral translation methods have been proposed and applied to various image translation tasks [31–35]. Methods such as ThermalGAN [35], IR GAN [36], and Pix2Pix MRFFF [37] generate infrared images from visible images using GAN-based frameworks. The above GAN-based cross-spectral image translation methods can be categorized into paired and unpaired approaches, depending on their reliance on paired images. Paired approaches require a training dataset with paired data, and this constraint greatly limits their application in visible-to-thermal infrared image translation [38]. Compared with paired approaches, unpaired approaches can overcome the requirement for paired training data and thus offer broader applicability [39]. Therefore,

this paper focuses on unpaired cross-spectral image translation from visible images to thermal infrared images.

The core of unpaired image translation is to learn a mapping from the source domain, namely the visible domain, to the target domain, namely the thermal infrared domain, under an unpaired data setting. In essence, it infers the joint distribution between domains from the marginal distributions of images in each domain. Since a marginal distribution can correspond to many possible joint distributions, additional constraints are required to support GAN training and to rule out unreasonable joint distribution estimates, thereby improving the reliability of translation results [40,41]. Pioneering works such as CycleGAN [42], DiscoGAN [43], and DualGAN [44] adopt cycle consistency as a key training principle and aim to obtain an optimal inference solution under this constraint. The main idea is to build a bidirectional translation chain of the input image, the target domain image, and the reconstructed source domain image, so that the mappings between the two domains are learned simultaneously, and the reconstructed image is required to remain consistent with the original input. However, this constraint involves a critical limitation, namely that translation between domains is assumed to be strictly reversible. This assumption conflicts with the needs of many real application scenarios [45] and has become a major bottleneck that limits the practicality and generalization of such methods, as well as a central challenge in this research area. To address this inherent limitation of cycle consistency, researchers have proposed various new constraints to improve domain mapping. Among them, ACLGAN [46] introduces an adversarial consistency constraint to relax the strict requirement implied by cycle consistency. CUT [47] adopts a contrastive learning-based constraint, using mutual information maximization to avoid reliance on a bijection assumption. GCGAN [48] proposes a geometric-consistency constraint and learns an unsupervised domain mapping by training only a one-way generator, which indirectly achieves two-way mapping and thus avoids the bijection requirement. DR AVIT [49] further proposes a dual geometric-consistency constraint that emphasizes global feature extraction to enable more efficient mapping between images from different domains and better match practical scenarios. Nevertheless, these improvements mainly provide targeted remedies for the key weakness and do not fully resolve the range of issues arising from cycle-consistency constraints.

Building on GANs, researchers have also introduced disentangled representation learning. Its key idea is to model the intrinsic factors of variation in data, so as to identify and separate the underlying generative factors hidden in the data, and to encode these factors as independent variables with clear semantic meaning [50–52]. In simple terms, disentangled representation learning can represent the semantic structure of an image and its imaging style as separate variables, and it has been widely used in a range of image-processing tasks [50]. It can learn domain-invariant representations across multiple domains, enabling cross-domain image translation and manipulation [49], and it has shown strong performance in unsupervised domain adaptation [53]. These properties align well with our research on cross-spectral image translation.

Based on the current research landscape, we observe that the differences between visible and thermal infrared imaging mechanisms lead to a substantial domain gap. Moreover, under an unpaired setting, relying only on cycle consistency may result in ill-posed solutions and mapping ambiguity. Therefore, when designing our approach, we aim to ensure that the model can account for both global and local features during cross-spectral image translation while preserving geometric shapes to stabilize cross-domain mapping and maintain structural fidelity. To this end, this paper builds on a dual learning framework and combines the strengths of dual geometric-consistency constraints and cycle-consistency constraints, so as to achieve an accurate mapping of both overall content and local details

across different image domains. We further introduce disentangled representation learning, which enables the GAN to make effective use of training data from two different domains and thus narrow the domain gap. Overall, the symmetric design of the framework also helps improve training stability. Experiments on public datasets, including AVIID [54] and Drone Vehicle Datasets [55], show that the proposed method outperforms several existing image translation approaches to translating visible images into infrared images.

In summary, the main contributions of this paper are as follows. (i) We propose a new model, termed Dual Geometric Cycle Generative Adversarial Networks (DGC GAN), with the primary goal of enabling unpaired cross-spectral image translation from visible images to thermal infrared images. (ii) We design a dual geometric cycle learning scheme that integrates dual geometric consistency and cycle consistency. This scheme preserves geometric structure under image transformations and improves the stability and reliability of global and local cross-spectral mapping. (iii) We introduce disentangled representation learning to separate semantic structure from image style, enabling one-to-many synthesis through recombination and improving the balance between realism and diversity in the generated thermal infrared images. (iv) We conduct comparative and ablation experiments on public datasets, including AVIID and Drone Vehicle, and further verify the feasibility of the proposed method on real drone-collected data. In addition, we analyze intensity distributions to clarify the scope and limitations of GAN-based cross-spectral synthesis.

The remainder of this paper is organized as follows. Section 2 presents the proposed method in detail. Section 3 describes the comparative experiments and reports the results. Section 4 discusses the practical applicability of the proposed approach. Finally, Section 5 concludes the paper.

2. Methodology

The goal of DGC GAN is to translate visible images into infrared images with high visual realism. Although this paper does not incorporate a physical spectral model, it disentangles semantic structure and imaging style for images from different domains and then recombines them across domains. In this way, it models cross-spectral domain differences and enables robust transfer. The architecture of the DGC_GAN model is shown in Figure 2. The DGC_GAN model consists of six cross-domain generators $\{G_U, G_V, G_{U'}, G_{V'}, G_{\widetilde{U}}, G_{\widetilde{V}}\}$, six domain discriminators $\{D_U, D_V, D_{U'}, D_{V'}, D_{\widetilde{U}}, D_{\widetilde{V}}\}$, six semantic structure encoders $\{E_X^{ss}, E_{X'}^{ss}, E_y^{ss}, E_{y'}^{ss}, E_u^{ss}, E_v^{ss}\}$, six imaging style encoders $\{E_X^{is}, E_{X'}^{is}, E_y^{is}, E_{y'}^{is}, E_u^{is}, E_v^{is}\}$, and three additional semantic structure discriminators $\{D_{xy}^{ss}, D_{x'y'}^{ss}, D_{uv}^{ss}\}$.The semantic structure encoders aim to map the input image into a shared, domain-invariant semantic structure space, while the imaging style encoders seek to embed the input image into domain-specific imaging style spaces. The goal of the generators is to produce transformed images based on semantic structure information and imaging style features. The domain discriminators attempt to train the semantic structure discriminators, distinguishing between semantic structure information extracted from the visible or infrared image domains. During training, parameters are shared between the cross-domain generators $\{G_U, G_{U'}, G_{\widetilde{U}}\}$ and $\{G_V, G_{V'}, G_{\widetilde{V}}\}$, the semantic structure encoders $\{E_X^{ss}, E_{X'}^{ss}, E_u^{ss}\}$ and $\{E_y^{ss}, E_{y'}^{ss}, E_v^{ss}\}$, and the imaging style encoders $\{E_X^{is}, E_{X'}^{is}, E_u^{is}\}$ and $\{E_y^{is}, E_{y'}^{is}, E_v^{is}\}$, thereby reducing the computational resource requirements.

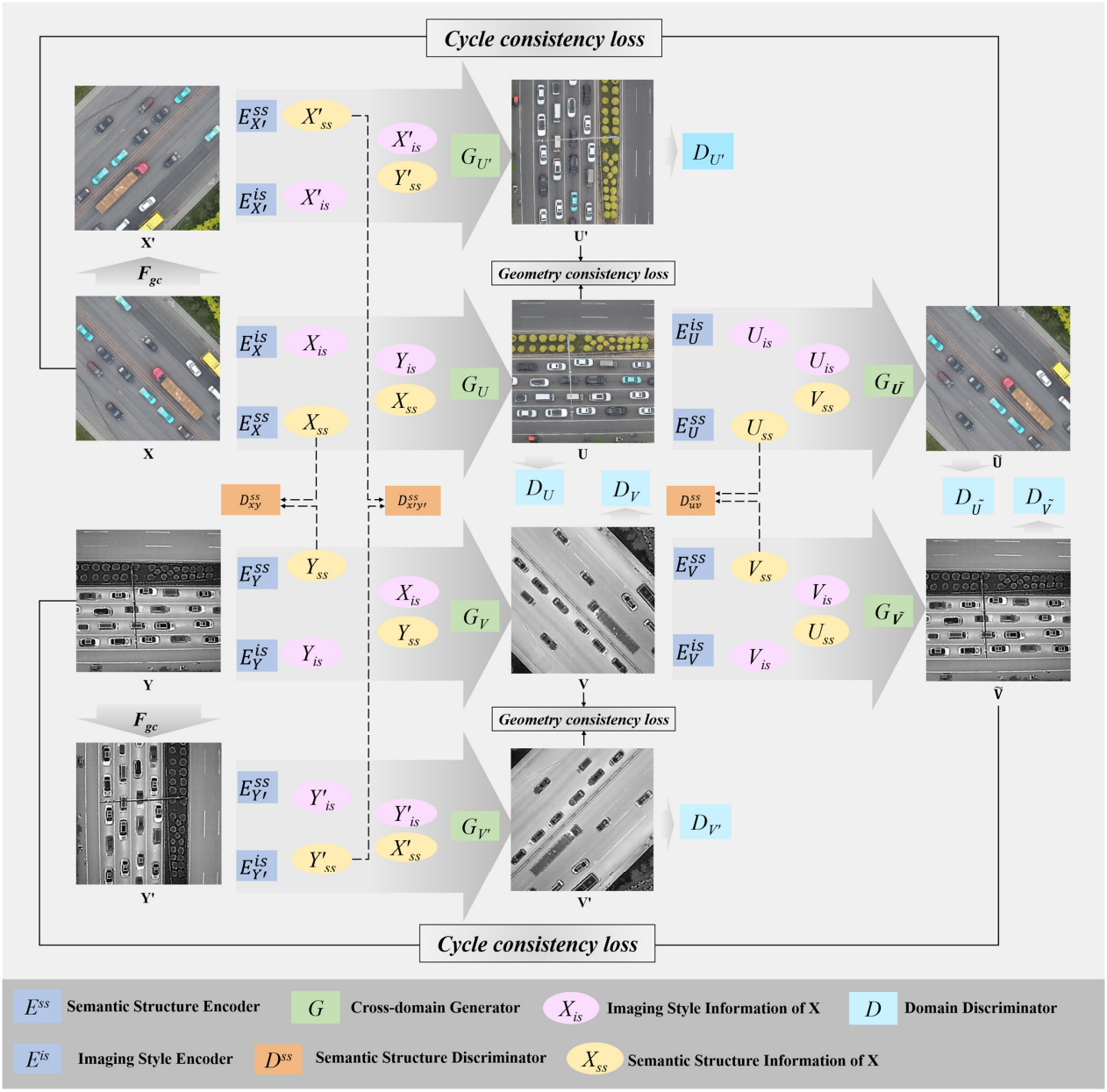

Figure 2. Overall architecture of DGC_GAN.

As shown in Figure 2, during training, the input visible image X and infrared image Y undergo a geometric rotation (prior research indicates that a 90° rotation achieves the best balance between realism and diversity in the generated image results, so we adopt a clockwise 90° rotation as the geometric transformation). This results in the generated images X′ and Y′. The four images, X, Y, X′ and Y′, are passed through the semantic structure encoders and imaging style encoders, which disentangle the corresponding semantic structure and imaging style information into different spaces. These disentangled features are then paired, and the resulting combinations are processed by different generators to produce new images, U, V, U′ and V′. The newly generated images are evaluated by discriminators, D, while the geometric-consistency loss is calculated between U and U′, as well as V and V′. The newly generated images, U and V, are then passed through the semantic structure encoders and imaging-style encoders to disentangle the corresponding semantic structure and imaging style information. These features are processed by the generators to produce the final output, which is then evaluated by the discriminators.

Meanwhile, the cycle consistency loss is computed between $\widetilde{U}$ and X, as well as between $\widetilde{V}$ and Y.

This section provides a detailed description of the strategy for image disentangled representation. It then outlines the dual geometric cycle-consistency constraint framework proposed in this paper, which enables GAN to better learn the cross-domain mapping between visible and infrared images. Finally, the loss function used in the proposed method is introduced.

2.1. Disentangled Representation Learning

This paper introduces disentangled representation learning, through which visible and infrared images are disentangled into semantic structure and imaging style spaces through the semantic structure encoder and imaging style encoder, as detailed below:

$$\begin{aligned} Z_x^{is}, Z_{x'}^{is}, Z_u^{is} &= E_x^{is}(x), E_{x'}^{is}(x'), E_u^{is}(u) \\ Z_y^{is}, Z_{y'}^{is}, Z_v^{is} &= E_y^{is}(y), E_{y'}^{is}(y'), E_v^{is}(v) \\ Z_x^{ss}, Z_{x'}^{ss}, Z_u^{ss} &= E_x^{ss}(x), E_{x'}^{ss}(x'), E_u^{ss}(u) \\ Z_y^{ss}, Z_{y'}^{ss}, Z_v^{ss} &= E_y^{ss}(y), E_{y'}^{ss}(y'), E_v^{ss}(v) \end{aligned} \tag{1}$$

In theory, the semantic structure information encoded from both domains should be shared and domain-invariant, while the imaging style features should preserve the specific style of each domain. To clearly and effectively disentangle the image representations, this paper employs two strategies: weight sharing and semantic structure adversarial loss. First, the weights are shared at the generator's output layer, which is responsible for the high-level semantic structure information of the semantic structure encoders $\{E_x^{ss}, E_{x'}^{ss}, E_u^{ss}\}$ and $\{E_y^{ss}, E_{y'}^{ss}, E_v^{ss}\}$. Through shared weighting, the input images to the semantic structure encoders are mapped into the same space. Secondly, a semantic structure discriminator is introduced to distinguish the domain associations of the encoded semantic structure information $\left\{Z_x^{ss}, Z_{x'}^{ss}, Z_u^{ss}\right\}$ and $\left\{Z_y^{ss}, Z_{y'}^{ss}, Z_v^{ss}\right\}$. This operation ensures that the semantic structure space contains as much semantic structural information as possible from the two different domains being mapped. In summary, the semantic structure encoder encodes the semantic structure information associated with the semantic structure discriminator, and the semantic structure adversarial loss is formulated as follows:

$$\begin{aligned} &L_{adv}^{ss}\left(E_x^{ss}, E_{x'}^{ss}, E_u^{ss}, E_y^{ss}, E_{y'}^{ss}, E_v^{ss}, D_{xy}^{ss}, D_{x'y'}^{ss}, D_{uv}^{ss}\right) \\ &= E_{x\sim x}[\tfrac{1}{2}\log D_{xy}^{ss}(E_x^{ss}(x)) + \tfrac{1}{2}\log(1 - D_{xy}^{ss}(E_x^{ss}(x)))] \\ &+E_{x'\sim x'}[\tfrac{1}{2}\log D_{x'y'}^{ss}(E_{x'}^{ss}(x')) + \tfrac{1}{2}\log(1 - D_{x'y'}^{ss}(E_{x'}^{ss}(x')))] \\ &+E_{u\sim u}[\tfrac{1}{2}\log D_{uv}^{ss}(E_u^{ss}(u)) + \tfrac{1}{2}\log(1 - D_{uv}^{ss}(E_u^{ss}(u)))] \\ &+E_{y\sim y}[\tfrac{1}{2}\log D_{xy}^{ss}(E_y^{ss}(y)) + \tfrac{1}{2}\log(1 - D_{xy}^{ss}(E_y^{ss}(y)))] \\ &+E_{y'\sim y'}[\tfrac{1}{2}\log D_{x'y'}^{ss}(E_{y'}^{ss}(y')) + \tfrac{1}{2}\log(1 - D_{x'y'}^{ss}(E_{y'}^{ss}(y')))] \\ &+E_{v\sim v}[\tfrac{1}{2}\log D_{uv}^{ss}(E_v^{ss}(v)) + \tfrac{1}{2}\log(1 - D_{uv}^{ss}(E_v^{ss}(v)))] \end{aligned} \tag{2}$$

After disentangling the image representations, the cross-domain generators recombine the semantic structure features and imaging style features from different domains to perform image translation, thereby generating new images.

$$\begin{aligned} u &= G_u\left(z_x^{is}, z_y^{ss}\right), u \sim U \\ u' &= G_{u'}\left(z_{x'}^{is}, z_{y'}^{ss}\right), u' \sim U' \\ \widetilde{u} &= G_{\widetilde{u}}\left(z_u^{is}, z_v^{ss}\right), \widetilde{u} \sim \widetilde{U} \\ v &= G_v\left(z_y^{is}, z_x^{ss}\right), v \sim V \\ v' &= G_{v'}\left(z_{y'}^{is}, z_{x'}^{ss}\right), v' \sim V' \\ \widetilde{v} &= G_{\widetilde{v}}\left(z_v^{is}, z_u^{ss}\right), \widetilde{v} \sim \widetilde{V} \end{aligned} \tag{3}$$

Specifically, $\{u, u', \tilde{u}, v, v', \tilde{v}\}$ are the translated images, while $\{U, U', \widetilde{U}, V, V', \widetilde{V}\}$ denote their corresponding translation domains. To make the translated images indistinguishable from real ones, we adopt a domain adversarial loss. Taking the visible-to-infrared direction as an example, the adversarial loss is defined as follows:

$$\begin{aligned} L_{adv}^{X} &= L_{adv}^{cc}\left(L_{adv}^{dg}\right) \\ L_{adv}^{dg}\left(G_X, D_X, E_Y^{ss}, E_X^{is}\right) &= E_{y\sim Y}[\log D_Y(y)] + E_{v\sim V}[\log(1 - D_Y(u))] \\ L_{adv}^{cc}\left(G_U, D_U, E_V^{ss}, E_X^{is}\right) &= E_{u\sim U}[\log D_U(u)] + E_{\tilde{u}\sim\widetilde{U}}[\log(1 - D_U(\tilde{u}))] \end{aligned} \tag{4}$$

For the reverse mapping from the infrared image domain to the visible image domain, the domain adversarial loss is defined as follows:

$$\begin{aligned} L_{adv}^{Y} &= L_{adv}^{cc}\left(L_{adv}^{dg}\right) \\ L_{adv}^{dg}\left(G_Y, D_Y, E_X^{ss}, E_Y^{is}\right) &= E_{x\sim X}[\log D_X(x)] + E_{u\sim U}[\log(1 - D_X(u))] \\ L_{adv}^{cc}\left(G_U, D_U, E_V^{ss}, E_X^{is}\right) &= E_{v\sim V}[\log D_V(v)] + E_{\tilde{v}\sim\widetilde{V}}[\log(1 - D_V(\tilde{v}))] \end{aligned} \tag{5}$$

For X' and Y' after the geometric transformation, the domain adversarial losses in equations $L_{adv}^{X'}$, $L_{adv}^{Y'}$, L_{adv}^{X} and L_{adv}^{Y}, have the same form.

2.2. Dual Geometric Cycle Framework

To produce translated images that are closer to real images, we use the proposed dual geometric cycle framework during GAN training. This framework imposes both dual geometric consistency and cycle consistency, and it takes advantage of the strengths of both constraints. Under this framework, during translation between the visible domain and the infrared domain, the geometric transformation applied to the input images should be preserved. As a result, the geometric relation between the translated images should match the relation between the corresponding input images. After enforcing the dual geometric constraint, we perform one more round of translation using the newly generated images without geometric transformation. We then combine the final outputs with the input images to impose cycle consistency.

The dual geometric-consistency loss in the framework is defined as follows:

$$\begin{aligned} L_{dgc}\left(E^{ss}, E^{is}, G\right) = {} & E_{u\sim U,u'\sim U'}\left|\left|F_{gc}(u) - u'\right|\right|_1 + E_{u\sim U,u'\sim U'}\left|\left|u - F_{gc}^{-1}(u')\right|\right|_1 \\ & + E_{v\sim V,v'\sim V'}\left|\left|F_{gc}(v) - v'\right|\right|_1 + E_{v\sim V,v'\sim V'}\left|\left|u - F_{gc}^{-1}(v')\right|\right|_1 \end{aligned} \tag{6}$$

The cycle-consistency loss in the framework is defined as follows:

$$L_{cc}\left(E^{ss}, E^{is}, G\right) = E_{u\sim U,u'\sim U}\left|\left|G\left\{E^{ss}\left[G\left(E^{ss}(u), E^{is}(v)\right)\right], [E^{ss}(u)]\right\} - u\right|\right|_1 \tag{7}$$

Here, E^{ss} denotes the semantic structure encoder, E^{is} denotes the imaging style encoder, and G denotes the cross-domain generator.

2.3. Loss Functions

In addition to the semantic structure adversarial loss, the domain adversarial loss, the dual geometric-consistency loss, and the cycle-consistency loss, we also use several other losses to make network training easier. The details of these losses are given below.

2.3.1. Identity Loss

As shown in Figure 3, using the semantic structure and imaging style features that are separated from input image V, the cross-domain generator should decode them back to image V′, and we apply an identity loss to enforce this.

$$\begin{aligned} L_{idt}\left(E^{ss}, E^{is}, G\right) = E_{x\sim X}\left[\left|\left|G_X\left(z_x^{ss}, z_x^{is}\right) - x\right|\right|_1\right] + E_{x\sim U}\left[\left|\left|G_U\left(z_U^{ss}, z_U^{is}\right) - U\right|\right|_1\right] \\ + E_{y\sim Y}\left[\left|\left|G_Y\left(z_Y^{ss}, z_Y^{is}\right) - Y\right|\right|_1\right] + E_{y\sim V}\left[\left|\left|G_V\left(z_V^{ss}, z_V^{is}\right) - V\right|\right|_1\right] \end{aligned} \quad (8)$$

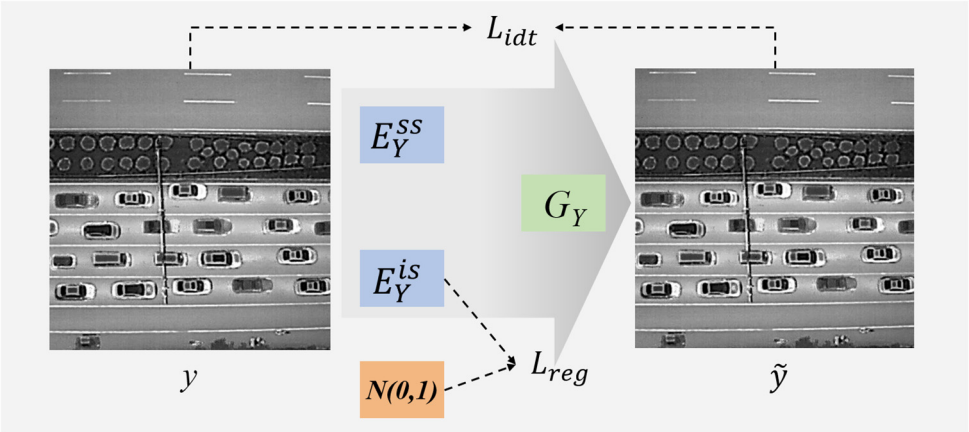

Figure 3. Illustration of the regularization loss and the identity loss.

2.3.2. Regularization Loss

To enable a random sampling of imaging style features and to produce different outputs during testing, we apply a regularization loss that forces the distribution of the style features to be close to a prior Gaussian distribution, as shown in Figure 3, which can be written as:

$$\begin{aligned} L_{reg}\left(E_X^{is}, E_U^{is}, E_Y^{is}, E_V^{is}\right) = E_{x\sim X}\left[D_{KL}\left(E_X^{is}(x)\middle|\middle|N(0,1)\right)\right] + E_{x\sim U}\left[D_{KL}\left(E_U^{is}(u)\middle|\middle|N(0,1)\right)\right] \\ + E_{y\sim Y}\left[D_{KL}\left(E_Y^{is}(y)\middle|\middle|N(0,1)\right)\right] + E_{y\sim V}\left[D_{KL}\left(E_V^{is}(v)\middle|\middle|N(0,1)\right)\right] \end{aligned} \quad (9)$$

where $D_{KL}(p||q) = -\int p(z)\log(p(z)/q(z))dz$

2.3.3. Latent Regression Loss

To promote a stable mapping between images and imaging style features, we use a latent regression process. As shown in Figure 4, we randomly sample a latent vector, z, from a prior Gaussian distribution as the imaging style feature, and we combine it with the semantic structure information as the input to the cross-domain generator to perform image translation. We add domain adversarial losses $L_{adv}^{X}(G_X, D_X, E_Y^{ss})$, $L_{adv}^{U}(G_U, D_U, E_U^{ss})$, $L_{adv}^{Y}(G_X, D_X, E_x^{ss})$, and $L_{adv}^{V}(G_X, D_X, E_V^{ss})$ to confirm that the translated images belong to the target domain. We then use the corresponding imaging style encoder to reconstruct the style features from the translated images. Intuitively, the reconstructed style features should be as close as possible to the original features. We use a latent regression loss to achieve this goal.

$$\begin{aligned} L_{latent}\left(E^{ss}, E^{is}, G\right) = E_{x\sim X}\left[\left|\left|E_Y^{is}\left(G_Y\left(E_X^{ss}(x), z\right)\right) - z\right|\right|_1\right] \\ + E_{x\sim U}\left[\left|\left|E_V^{is}\left(G_V\left(E_U^{ss}(x), z\right)\right) - z\right|\right|_1\right] \\ + E_{y\sim Y}\left[\left|\left|E_X^{is}\left(G_X\left(E_Y^{ss}(y), z\right)\right) - z\right|\right|_1\right] \\ + E_{y\sim V}\left[\left|\left|E_U^{is}\left(G_U\left(E_V^{ss}(y), z\right)\right) - z\right|\right|_1\right] \end{aligned} \quad (10)$$

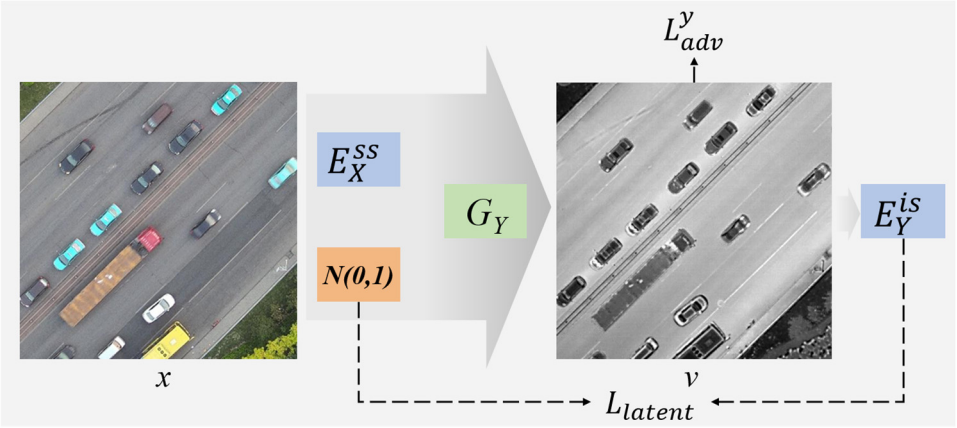

Figure 4. Schematic diagram of the latent regression process.

3. Comparative Experiments and Results

The comparative experiments in this paper mainly introduce several different methods for comparison on public datasets to verify the superiority of the DGC_GAN model. We also conduct ablation studies to demonstrate the effectiveness of the proposed dual geometric cycle framework and perform a sensitivity analysis on the key loss function parameters of DGC_GAN.

3.1. Datasets Description

In this study, the AVIID3 unpaired 8 subset from the publicly available AVIID dataset is used to verify the effectiveness of the proposed model architecture. In addition, data are collected from the public Drone Vehicle Datasets. Invalid images that contain no target objects or suffer from severe interference are removed, and the remaining images are divided into two categories, Day Road and Night Road, according to the acquisition time. The basic statistics of the three datasets are summarized in Table 1.

Table 1. Overview of the basic information of the datasets.

	AVIID	Day Road	Night Road
Number of images (couple)	1280	5724	13,090
Image size	480 × 480	640 × 512	640 × 512
Download	https://github.com/silver-hzh/Averial-visible-to-infrared-image-translation/tree/main (accessed on 8 February 2026)	https://github.com/VisDrone/DroneVehicle (accessed on 8 February 2026)	

Compared with AVIID, the Day Road and Night Road sets contain more images, more complex scenes, and a wider range of objects, which provides strong data for our experiments. The detailed information for AVIID, Day Road, and Night Road is given below.

3.1.1. AVIID

AVIID is the first dataset designed specifically for visible-to-infrared image translation. It contains 1280 couples of aerial visible and infrared images, each with a resolution of 512 by 512. The dataset is split into training and testing sets at a ratio of 4 to 1. The scenes in these images include roads, bridges, parking lots, and streets. Examples from AVIID are shown in Figure 5.



Figure 5. Sample images from the AVIID dataset.

3.1.2. Day Road

This dataset contains 5724 couples aerial visible and infrared images captured during daytime by drones at ten different altitudes and viewing angles. It covers a variety of scenes, including urban roads, residential areas, industrial zones, and parking lots. Each image has a resolution of 640 by 512. The dataset is split into training and testing sets at a ratio of 4 to 1. Example images from the Day Road dataset are shown in Figure 6.

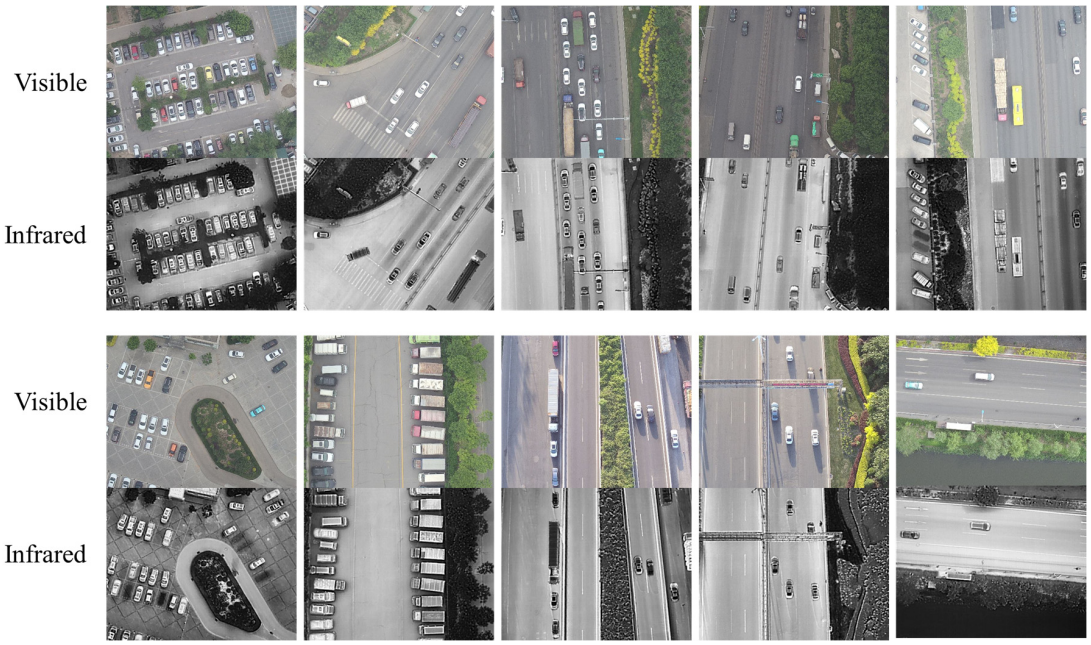

Figure 6. Sample images from the Day Road dataset.

3.1.3. Night Road

Compared with Day Road, the Night Road dataset contains more images, with a total of 13090 couples, and includes more complex scenes such as urban roads, residential areas, industrial zones, and parking lots. It is split into training and testing sets at a ratio of 4 to 1. The influence of nighttime artificial lighting poses additional challenges for translating visible images into infrared images. Example images from the Night Drone dataset are shown in Figure 7.

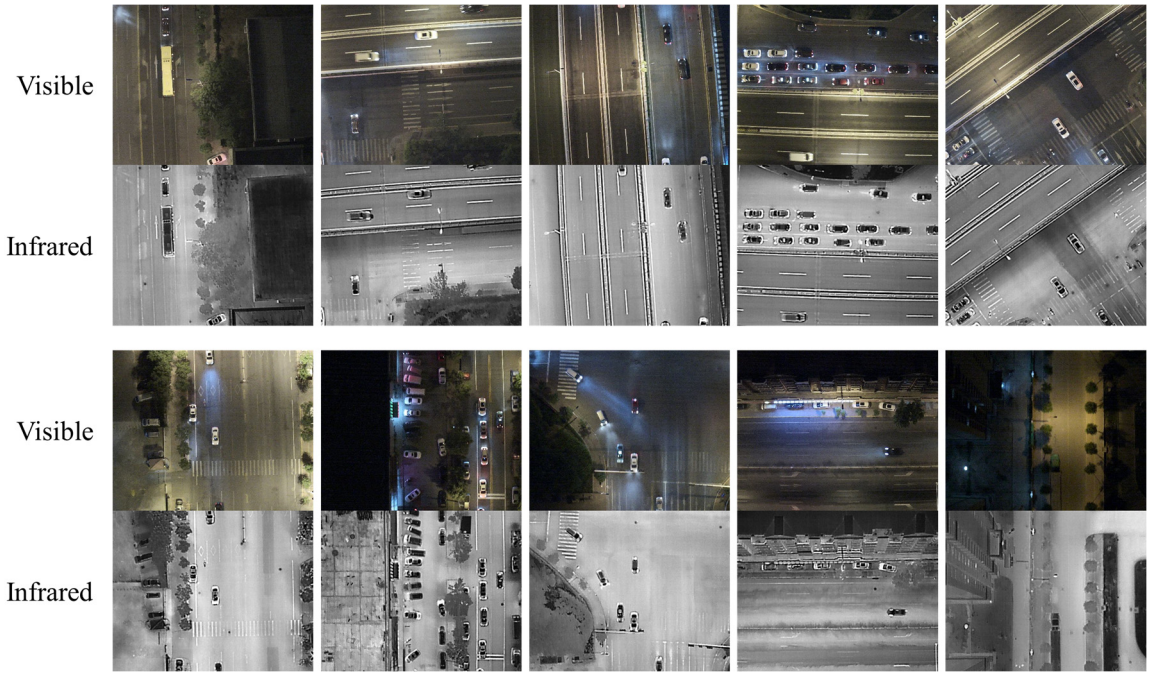

Figure 7. Sample images from the Night Road dataset.

3.2. Compared Algorithms

We compare our DGC_GAN with several recent image translation methods. The compared algorithms include MUNIT [56], DRIT [57], DSMAP [58], ACLGAN [46], and DR_AVIT [49]. Their brief descriptions are as follows.

3.2.1. MUNIT

MUNIT is an unsupervised image translation model. Its key idea is to split the latent representation into a content code and a style code. By combining the content code from one domain with a style code that is randomly sampled from the target domain, it produces cross-domain translations with high quality and good diversity. This approach addresses the limitation of CycleGAN, which struggles to generate diverse style variants in the target domain.

3.2.2. DRIT

DRIT centers on learning a domain-invariant content representation with a shared content encoder. It then uses domain-specific attribute representations and cross-domain translators to decouple style and attributes and to control them with diversity. It performs well on translation between domains with large semantic differences. It is an important advance and a useful complement after MUNIT.

3.2.3. DSMAP

DSMAP is a generative adversarial model for fine-grained and high-fidelity image-to-image translation. Its main goal is to address content distortion, detail loss, and local inconsistency in complex scenes. It is especially useful when the precise control of specific region attributes is needed or when the objects are highly structured.

3.2.4. ACLGAN

ACLGAN is a generative adversarial model for multi-attribute image editing. Its main goal is to modify selected attributes in a precise and local way while keeping other regions unchanged as much as possible. It integrates class activation maps and an attention mechanism into the GAN framework to locate the editing regions with high accuracy.

3.2.5. DR_AVIT

DR_AVIT is a generative adversarial model for translation between visible and infrared images in drone aerial scenes. Its core idea is to build on a dual learning framework and use disentangled representation learning to extract semantic structure and imaging style from images. It also adds a new dual geometric-consistency constraint to support efficient mapping between different image domains.

3.3. Evaluation Metrics

To assess the visual realism of the infrared images produced by different methods, we use two perception metrics: the Fréchet Inception Distance (FID) and the Kernel Inception Distance (KID). To evaluate the diversity among multiple generated results from the same input, we adopt two metrics, namely the Learned Perceptual Image Patch Similarity (LPIPS) and the Structural Similarity Index Measure (SSIM). The details of these metrics are described below.

3.3.1. FID

FID is computed by measuring the divergence of deep features taken from the last pooling layer of the InceptionV3 model, which is trained on ImageNet. It is widely used to evaluate how similar the distribution of generated images is to the distribution of real images. The formula is defined as follows:

$$\mathrm{FID}(Y, \widehat{Y}) = ||m_Y - m_{\widehat{Y}}||_2^2 + \mathrm{Tr}(C_Y + C_{\widehat{Y}} - 2(C_Y C_{\widehat{Y}})^{\frac{1}{2}}) \tag{11}$$

m is the mean deep feature, C is the covariance matrix, and Tr is the trace operator.

3.3.2. KID

KID is an improved metric for measuring the similarity between generated images and real images. It compares kernel matrices derived from feature representations and quantifies the difference using the largest eigenvalue of a Gaussian kernel. It also uses the InceptionV3 model to extract deep features from the translated images and the real images. Intuitively, if the generated images are more similar to the real images, the FID and KID score should be lower.

3.3.3. LPIPS

LPIPS can be represented by the weighted L2 distance between the deep features extracted by the deep convolutional neural network and the ground truth of the generated image, which is expressed as follows:

$$\mathrm{LPIPS}(y, \widehat{y}) = \sum_{l=1}^{l} \frac{1}{H_l W_l} \sum_{h,w} ||w_l \Theta (y_{i,hw}^l - \widehat{y}_{i,hw}^l)||^2 \tag{12}$$

Here, y and $\hat{y}$ represent the generated image and its corresponding ground truth, while y^l and $\hat{y}^l$ are the normalized deep features extracted from the l layer of the deep convolutional neural network. w_l denotes the weighting parameter. A smaller LPIPS distance between two corresponding images indicates better similarity. Conversely, a higher average LPIPS distance signifies greater diversity between the transformation results.

3.3.4. SSIM

SSIM is a metric for measuring the similarity between two images. It takes into account the brightness, contrast, and structural information of the images and can be expressed as follows:

$$\mathrm{SSIM}(y, \hat{y}) = \frac{(2\mu_y\mu_{\hat{y}} + c_1)(2\sigma_{y\hat{y}} + c_2)}{\left(\mu_y^2 + \mu_{\hat{y}}^2 + c_1\right)(\sigma_y^2 + \sigma_{\hat{y}}^2 + c_2)} \tag{13}$$

Here, c_1 and c_2 are constants, while μ_y, $\mu_{\hat{y}}$, σ_y, and $\sigma_{\hat{y}}$ represent the mean and variance of the generated image and its corresponding ground truth, respectively. Additionally, $\sigma_{y\hat{y}}$ denotes the covariance between them. A higher SSIM value indicates better similarity between the two images, while a lower SSIM value suggests better diversity between the transformed images.

3.4. Experimental Settings

The experiments were conducted on a workstation equipped with an Intel (R) Core (TM) i9 14900KF CPU (3.20 GHz) and an RTX 5090 GPU. During training, each input image is first cropped into 256 by 256 patches and then rotated by 90 degrees with a probability of one half to introduce geometric variation. The base learning rate is set to 1×10^{-4}, while the learning rate of the semantic structure discriminator is set to 4×10^{-5}. The weight decay is kept the same for all optimizers and set to 1×10^{-4}. Training is performed for 100 epochs. The learning rate is kept unchanged for the first 50 epochs and then linearly decayed for the remaining 50 epochs using a lambda schedule. Model parameters are updated using the Adam optimizer, with beta one set to 0.5 and beta two set to 0.999. During testing, we use the model parameters saved at the last epoch of training. After evaluation, the cross-domain generator produces ten thermal infrared translation results for each input visible image.

For the test subsets, we select 200 images from the AVIID test set, and 1000 images from each of the Day Road and Night Road test sets for evaluation. The FID and KID scores between all generated infrared images and their corresponding ground-truth values were calculated using PyTorch 11.8's fidelity metrics. To calculate the average LPIPS and SSIM, five couples of generated infrared images were extracted for each input visible image. Due to the instability of GAN training, to ensure a fair and comparable evaluation, we repeat the experiments five times for each network and select the best translation results among the five runs for comparison.

3.5. Experimental Results and Discussion

3.5.1. Quantitative Results

Different methods exhibit different performances on the same dataset. We report the quantitative results of various methods on the AVIID, Day Road, and Night Road datasets in Tables 2–4. The best results are highlighted in bold. (↑) indicates that larger values are better; (↓) indicates that smaller values are better. From the tables, it is evident that the method proposed in this paper, DGC_GAN, performs exceptionally well in both realism and diversity. On the Day Road dataset, DGC_GAN ranks first in FID and KID evaluation scores, with values of 63.727 and 0.008711, respectively, and the highest average LPIPS value of 0.268151. On the Night Road dataset, DGC_GAN also achieves the lowest FID

and KID scores, with values of 69.419 and 0.019352, respectively. DGC_GAN performs excellently on the AVIID dataset as well, surpassing other algorithms in all four metrics. We attribute this to two key factors: First, the use of disentangled representation learning to decouple semantic structures and randomly sample imaging styles significantly improves the diversity of the transformation results. On the other hand, the new model framework proposed in this paper leverages the advantages of dual geometric-consistency constraints and cycle-consistency constraints, allowing the model to effectively learn the mapping between the visible image domain and the infrared image domain. It also accounts for transformation information at different levels, making the results closer to real images.

Table 2. Quantitative results of different algorithms on the Day Road.

	FID (↓)	**KID (↓)**	**LPIPS (↑)**	**SSIM (↓)**
MUNIT	65.385	0.017531	0.173115	0.853941
DRIT	65.037	0.009714	0.248961	0.817952
DSMAP	68.082	0.050617	0.148656	0.841365
ACLGAN	70.321	0.063272	0.163974	0.971135
DR_AVIT	65.244	0.012619	0.194701	**0.791172**
DGC_GAN	**63.727**	**0.008711**	**0.268151**	0.791963

Table 3. Quantitative results of different algorithms on the Night Road.

	FID (↓)	**KID (↓)**	**LPIPS (↑)**	**SSIM (↓)**
MUNIT	77.461	0.041763	0.059854	0.846612
DRIT	72.805	0.027516	0.069517	0.761354
DSMAP	81.758	0.048192	0.068431	0.896551
ACLGAN	98.541	0.054913	0.054132	0.942844
DR_AVIT	71.864	0.024185	**0.076951**	0.809112
DGC_GAN	**69.419**	**0.019352**	0.075657	**0.794335**

Table 4. Quantitative results of different algorithms on the AVIID.

	FID (↓)	**KID (↓)**	**LPIPS (↑)**	**SSIM (↓)**
MUNIT	67.406	0.009012	0.097363	0.82386
DRIT	68.851	0.011051	0.097921	0.830268
DSMAP	77.008	0.019383	0.050415	0.869289
ACLGAN	79.534	0.043439	0.01485	0.959509
DR_AVIT	66.210	0.008468	0.102164	0.80293
DGC_GAN	**63.038**	**0.007495**	**0.159644**	**0.784365**

3.5.2. Qualitative Results

Figures 8–10 show the transformation results of different algorithms on the Day Road, Night Road, and AVIID datasets. In the ground-truth infrared images, we use red boxes to highlight local regions that are prone to errors, which makes it easier to visually observe the advantages of the DGC_GAN translation results. As can be seen from the figures, DGC_GAN, with its model structure integrating dual geometric-consistency constraints and cycle-consistency constraints, generates infrared images that outperform other methods in both diversity and realism. The MUNIT and DRIT algorithms rely solely on the cycle-consistency constraint in the cycle reconstruction process, resulting in generated images with a lot of texture features. The ACGAN algorithm introduces

reverse consistency constraints, eliminating the need for a cycle reconstruction process, but it retains more features from the input visible image domain. The DSMAP algorithm decomposes image representations into domain-specific semantic structures and imaging styles, but this sometimes leads to mismatched imaging styles and semantic structures, reducing the realism of the generated images. The DR_AVIT algorithm focuses on global imaging style features but tends to overlook the complex and subtle variations in local imaging styles, which causes the generated results to deviate from reality.

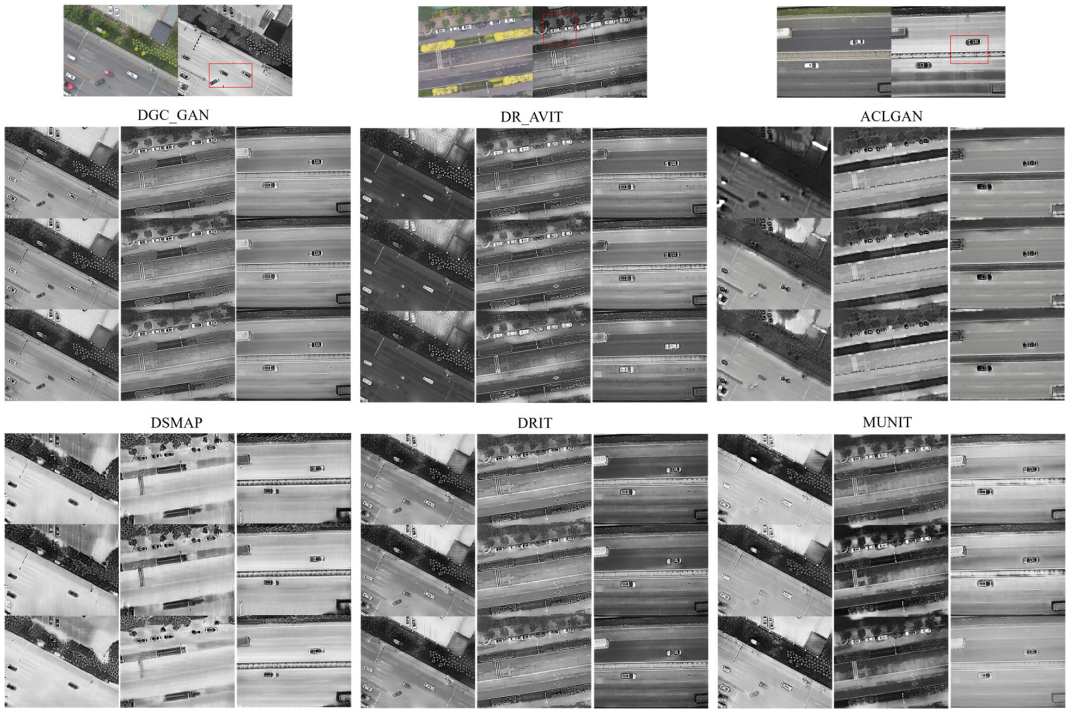

Figure 8. Translation results of different methods on Day Drone dataset.

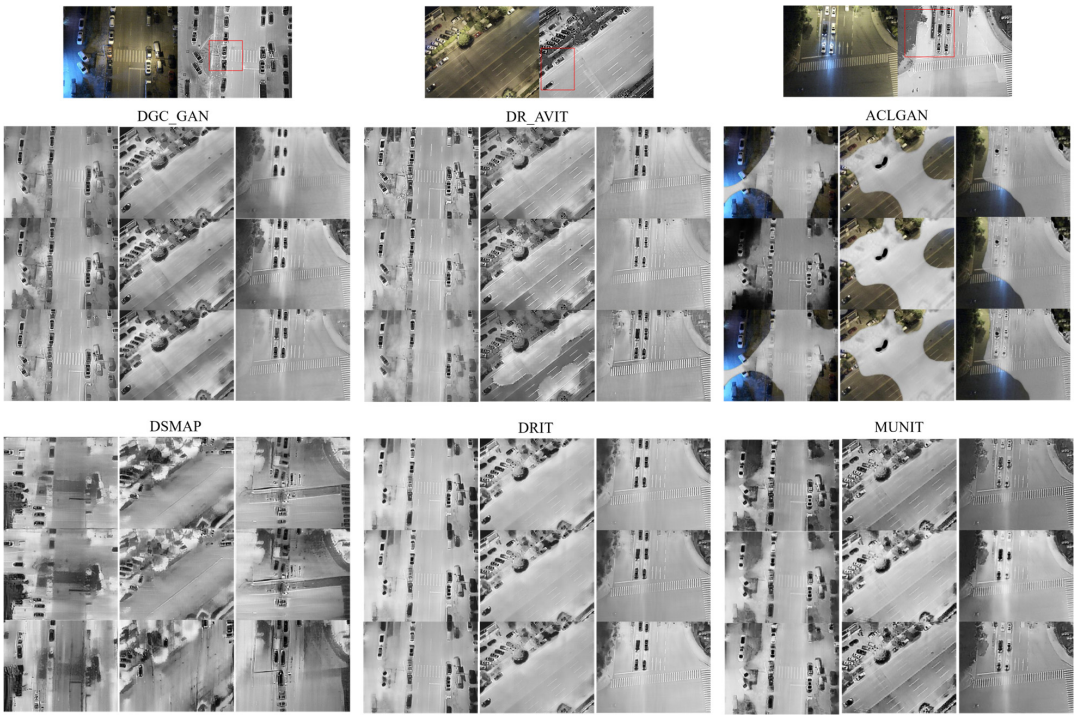

Figure 9. Translation results of different methods on Night Drone dataset.

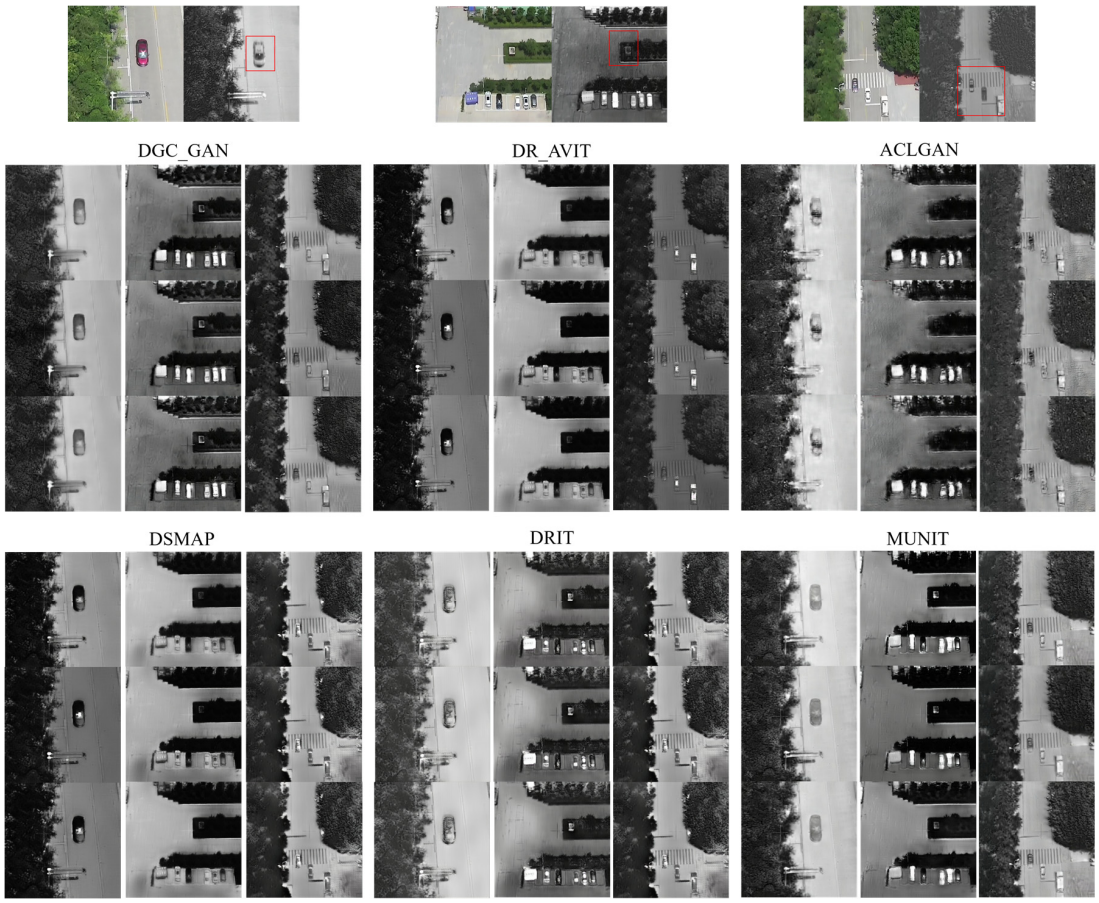

Figure 10. Translation results of different methods on AVIID dataset.

3.6. Ablation Experimental

To further demonstrate that the proposed dual-geometric-cycle framework can improve the accuracy of GAN-based cross-spectral mapping between visible and thermal infrared images, we progressively simplify DGC_GAN and derive two variants: a geometry-consistent GAN and a cycle-consistent GAN, both incorporating disentangled representation learning. The network architectures of these two GANs are illustrated in Figure 11. We then compare these two networks with DGC_GAN on the AVIID dataset. To ensure a fair evaluation, each experiment is repeated five times for each network, and the best translation results among the five runs are selected for comparison.

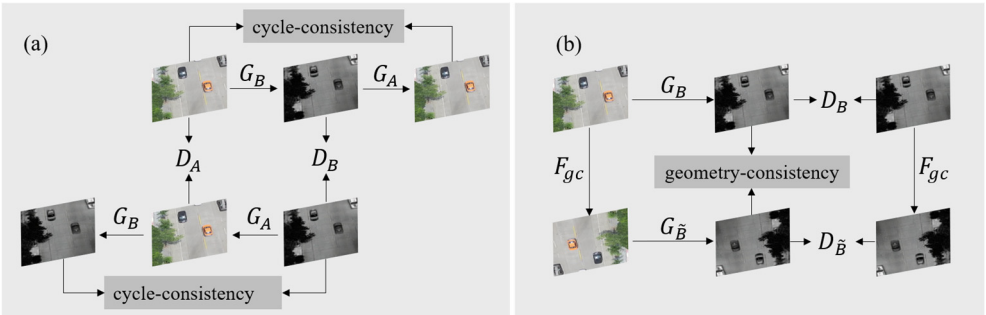

Figure 11. Architectures of the ablation models: (**a**) cycle-consistency constrained GAN; (**b**) geometric-consistency constrained GAN.

Using FID, KID, LPIPS, and SSIM as evaluation metrics, Table 5 reports the quantitative results of different methods on the AVIID dataset. As shown in the table, the

disentanglement-based CycleGAN, which learns the mapping between visible and thermal infrared images only with the cycle-consistency constraint, does not achieve satisfactory performance. Its visual realism is limited, with an FID of 77.291 and a KID of 0.035972. In contrast, GGAN, by combining the geometric-consistency constraint with disentangled representation learning, achieves clear improvements in both realism and diversity, reaching an FID of 68.043 and a KID of 0.014848. Compared with CycleGAN and GGAN, DGC_GAN shows more evident advantages in terms of realism and diversity. Overall, the proposed dual geometric cycle constraint leverages the complementary strengths of the dual geometric-consistency constraint and the cycle-consistency constraint. While preserving the global style of the image, it can also recover multi-level details, making the translation results more convincing.

Table 5. Ablation study results.

	FID (↓)	KID (↓)	LPIPS (↑)	SSIM (↓)
CycleGAN	77.291	0.035972	0.197371	0.942721
GCGAN	68.043	0.014848	0.247426	0.851142
DGC_GAN	**63.727**	**0.008711**	**0.268151**	**0.791963**

3.7. Sensitivity Analysis of Loss Function Parameters

In DGC_GAN, the dual geometric-consistency loss and the cycle-consistency loss are key to achieving accurate cross-domain mapping between visible images and thermal infrared images. Therefore, we investigate how different values of the two loss weights, λ_{gc} and λ_{cc}, affect the performance of DGC_GAN. Sensitivity experiments are conducted on the AVIID dataset using a one factor at a time strategy. When analyzing the impact of λ_{gc}, we set λ_{cc} to 10, and the results are reported in Table 6. When analyzing the impact of λ_{cc}, we set λ_{gc} to 20, and the results are reported in Table 7.

Table 6. Sensitivity analysis results for λ_{gc}.

λ_{cc} \ Metrics	FID (↓)	KID (↓)	LPIPS (↑)	SSIM (↓)
5	66.027	0.009335	0.094712	0.841353
10	64.774	0.008106	0.109475	0.845469
20	63.038	0.007495	**0.159644**	**0.784365**
30	**62.703**	**0.007114**	0.144962	0.803372

Table 7. Sensitivity analysis results for λ_{cc}.

λ_{cc} \ Metrics	FID (↓)	KID (↓)	LPIPS (↑)	SSIM (↓)
5	68.504	0.010547	0.104271	0.827731
10	**63.038**	**0.007495**	**0.159644**	**0.784365**
20	66.207	0.008462	0.128426	0.819725

As shown in Table 6, as the value of λ_{gc} increases, the FID and KID values gradually decrease, indicating that the realism of the translation results improves. This is because a stronger geometric-consistency constraint helps the GAN learn a more accurate mapping

between different domains. However, when λ_{gc} increases from 20 to 30, the LPIPS value decreases, and the SSIM value increases, suggesting reduced diversity for the translated results. This implies that an overly strict geometric-consistency constraint may reduce the efficiency of disentangled representation learning. As shown in Table 7, when λ_{cc} is set to 10, FID, KID, LPIPS, and SSIM all achieve their best values. Considering both realism and diversity, the model performs best when λ_{gc} is set to 20 and λ_{cc} is set to 10.

4. Discussion on the Practical Use of the Method

The comparative experiments have demonstrated the advantages of the proposed DGC_GAN method on public datasets. To further examine its practical usefulness, we conducted additional tests using real image data collected by a drone to evaluate the performance of DGC_GAN in generating thermal infrared images for real-world applications.

4.1. Experimental Equipment

For outdoor data acquisition, a DJI M350 RTK drone was used as the data collection platform, equipped with a DJI H30t dual-spectrum camera, as shown in Figure 12. The thermal imaging sensor in the H30t is an uncooled vanadium oxide VOx microbolometer, with an imaging wavelength range of 8 μm to 14 μm and a temperature measurement range of minus 20 degrees Celsius to 150 degrees Celsius. The lens has a focal length of 24 mm, with a 32 times equivalent zoom, and the photo resolution is 1280 by 1024. The camera supports simultaneous visible and thermal infrared imaging. For the indoor evaluation of practical applicability, the same computing equipment as that used in the comparative experiments was adopted, including an Intel Core i9 14900KF CPU at 3.20 GHz and an RTX 5090 GPU.



Figure 12. Data acquisition equipment: (**a**) DJI M350 RTK; (**b**) DJI H30t; (**c**) integrated configuration.

4.2. Overview of the Experimental Area

The data for this algorithm application experiment were collected on 11 November 2025. The weather was clear on that day, with a maximum temperature of 18.5 degrees Celsius, a minimum temperature of 10.3 degrees Celsius, and wind speeds between 3.4 and 5.4 m/s. The data collection area covered part of the Science Island campus, and the specific location is shown in Figure 13a. The overall size of the experimental area was approximately 1000 m by 500 m. The area contains diverse land-cover types, including both natural and human-made features, such as trees, buildings, roads, water bodies, grassland, and parking lots. Examples of these land-cover types are shown in Figure 13b. The land-cover distribution in the study area is relatively complex, which imposes higher requirements on the robustness and information extraction capability of the proposed method and, therefore, meets the needs of this application experiment.

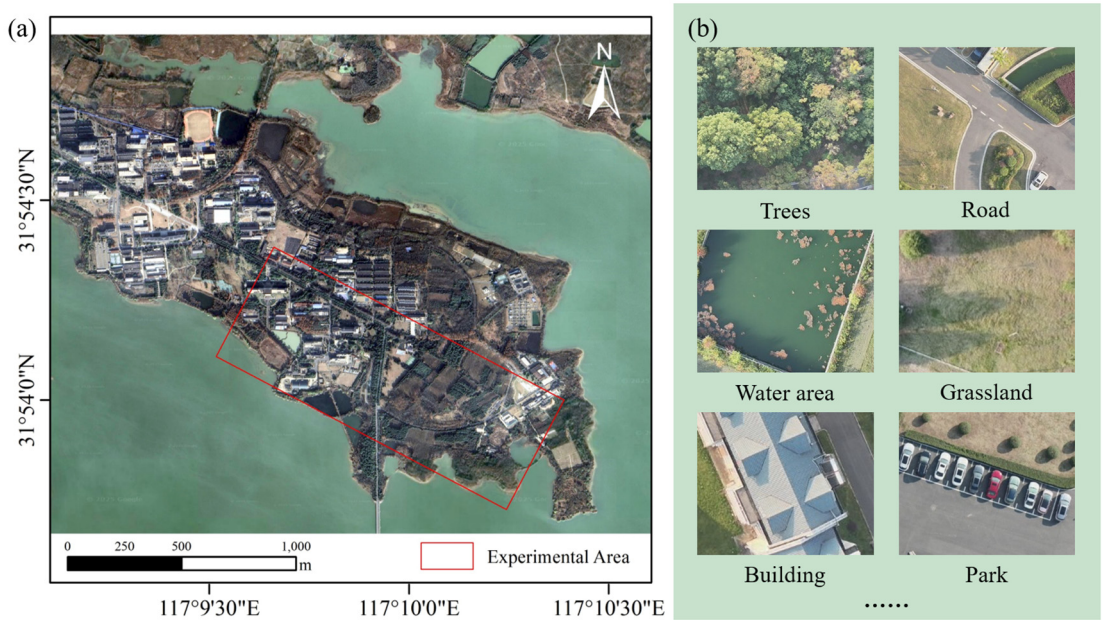

Figure 13. Location of the study area and examples of land-cover types: (**a**) study area location; (**b**) examples of land-cover types.

4.3. Experimental Procedure

The overall workflow of this algorithm application experiment is shown in Figure 14. During data acquisition, the drone flight altitude was set to 120 m. In this experiment, we collected 1080 couples of visible and infrared images. These images were first preprocessed and cropped to a resolution of 640 by 512 and then split into training and test sets at a ratio of 4 to 1. Example images are shown in Figure 15. Under the same experimental environment and parameter settings as in the comparative experiments, we trained DGC_GAN using the training set constructed from the collected data and then evaluated it on the test set after training. Finally, we assessed accuracy by comparing the translated infrared images with the corresponding real captured infrared images, in order to evaluate the quality of the generated infrared images.

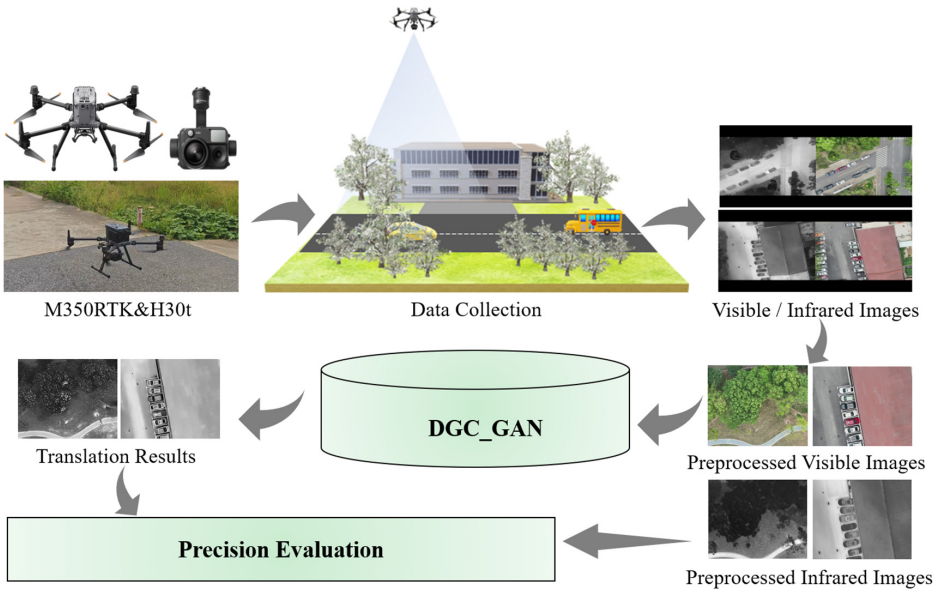

Figure 14. Application experiment flow chart.



Figure 15. Examples of the collected data in the experiment.

4.4. Results Analysis

The final quantitative evaluation results of the DGC_GAN application experiment are reported in Table 8. In this experiment, DGC_GAN achieves strong performance across all metrics, and the results are better than those obtained in the comparative experiments on public datasets. Specifically, the FID is 62.923, the KID is 0.004372, the LPIPS is 0.280954, and the SSIM is 0.763358. This improvement is mainly attributed to the high imaging quality of the H30t camera, which provides a stronger data foundation than the public datasets used in the comparative experiments. The results also indicate that DGC_GAN remains effective on high-resolution data. Overall, these quantitative results further demonstrate the practical usefulness of the proposed DGC_GAN method.

Table 8. Quantitative evaluation results of DGC_GAN in practical application.

Evaluation Metrics	FID	KID	LPIPS	SSIM
Value	62.923	0.004372	0.280954	0.763358

The translation results of the DGC_GAN application experiment are shown in Figure 16. As can be seen, the generated thermal infrared images exhibit strong overall realism. In particular, for land-cover targets with salient spatial distribution patterns, the generated images show a high degree of consistency with the original images in terms of shape, structure, and fine details, without obvious deformation or distortion. In addition, DGC_GAN demonstrates an excellent ability to recover shadow details that are often overlooked in thermal infrared images. The shadow regions in the generated images maintain contour boundaries consistent with those in the original thermal infrared images, further confirming the effectiveness and reliability of the proposed method for thermal infrared image generation.

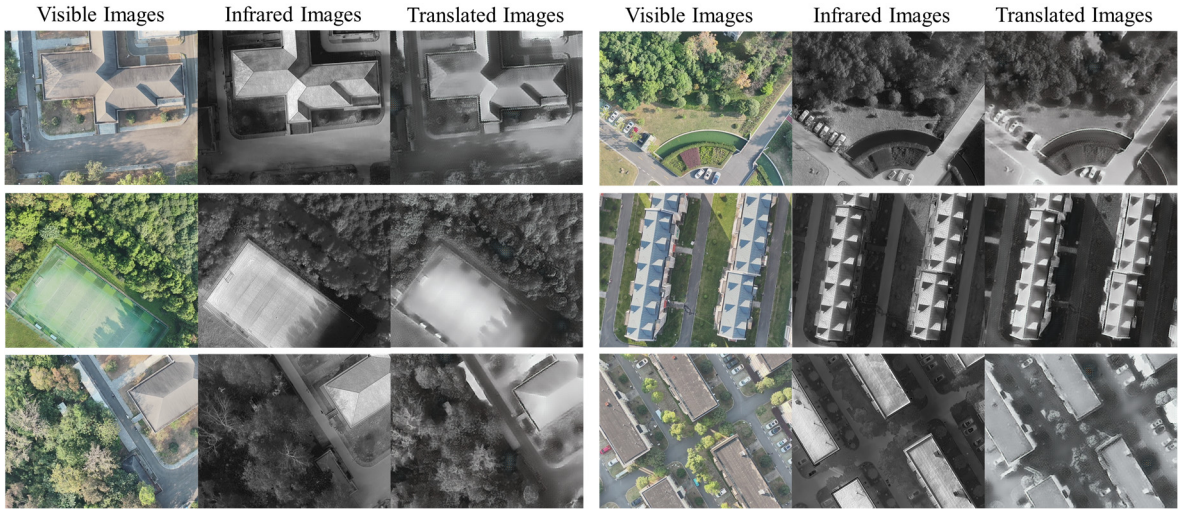

Figure 16. Examples of DGC_GAN translation results in practical application.

A closer look at the image intensity values focuses on the temperature distribution of different land-cover types in the scene. Figure 17 presents several thermal infrared images generated by DGC_GAN, the corresponding original thermal infrared images, and their respective intensity value distributions. Overall, the original and generated images show similar distribution trends, indicating consistent temperature distribution patterns across different land-cover types. However, the specific intensity values do not match one to one. In some local details, the intensity differences between the generated and original thermal infrared images are relatively large, suggesting a clear temperature discrepancy in those regions. This phenomenon can be attributed to the underlying objective of GAN-based methods, which aims to reproduce a similar appearance distribution, rather than achieving quantitative radiometric recovery at the level of physical measurements.

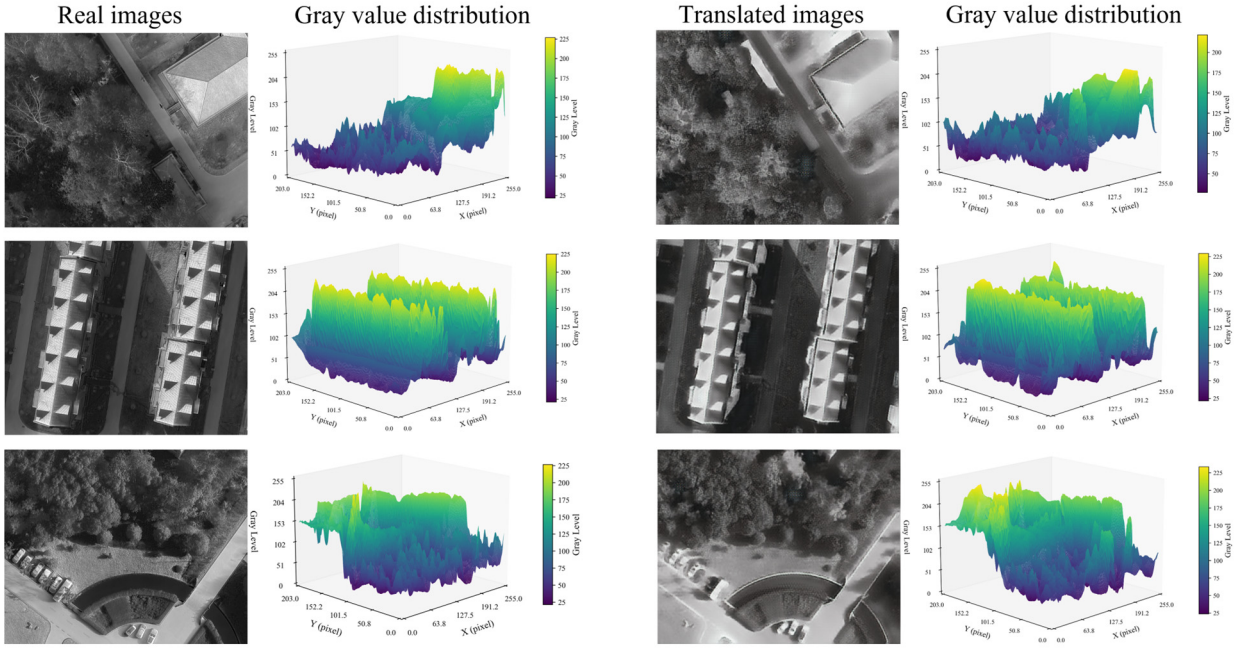

Figure 17. Comparison of intensity values between real and translated images.

5. Conclusions

To address the shortage of publicly available thermal infrared datasets, this paper has proposed a GAN-based model named DGC_GAN for cross-spectral translation from

visible images to thermal infrared images. The proposed approach is a deep learning method that can generate translated results with high realism and strong diversity. We developed a framework that combines dual geometric-consistency constraints with cycle-consistency constraints, which leverages the strengths of both constraints. This design improves the stability of GAN training while enabling the efficient learning of accurate mappings between the visible image domain and the thermal infrared image domain. In addition, DGC_GAN introduces disentangled representation learning to decompose an input image into two components, semantic structure and imaging style. The semantic structure from the visible image is then combined with randomly sampled imaging style features from thermal infrared images to obtain diverse translation results. Furthermore, we used visible and thermal infrared images from public datasets, including AVIID and Drone Vehicle Datasets, to evaluate the model performance. In addition, we designed ablation experiments to verify that the proposed framework can leverage the strengths of both the dual geometric-consistency constraint and the cycle-consistency constraint, and we conducted comparative experiments to analyze the sensitivity of the loss functions. We also designed an application-oriented experiment using data collected with a dual spectral drone platform, and we discussed the practical usefulness of DGC_GAN in scenarios with complex and diverse ground object distributions.

Comparative experiments on public datasets demonstrate that the proposed DGC_GAN method produces images with advantages in both realism and diversity. An application experiment using data collected with a dual-spectrum drone platform also confirms the applicability of DGC_GAN in complex scenes. However, several issues were observed during the experiments. On the one hand, DGC_GAN includes a relatively large number of loss functions and involves a high computational cost, which leads to demanding hardware requirements and a long training time. On the other hand, although the translated thermal infrared images generated by DGC_GAN show higher realism and diversity than those produced by other methods and achieve improved consistency with the original thermal infrared images in terms of intensity distribution trends, the exact intensity values still differ from those in the original images. Based on these observations, future work will extend visible-to-thermal infrared translation in two directions. First, we will explore lightweight model designs to reduce computational costs. Second, we will introduce land cover-based surface temperature simulation methods to constrain pixel intensity values in the generated thermal infrared images.

Author Contributions: Conceptualization, S.Y. and X.L.; methodology, S.Y. and X.S.; software, S.Y. validation, Y.W. (Yichen Wei) and B.S.; formal analysis, Y.W. (Yuyao Wang) and Y.L.; investigation, X.S.; resources, X.L.; data curation, Y.W. (Yichen Wei); writing—original draft preparation, S.Y.; writing—review and editing, X.S. and X.L.; visualization, X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Aerospace Science and Technology Innovation Application Research Project (E23Y0H555S1), the Aviation Science and Technology Innovation Application Research Project (62502510201), and the Key Laboratory Project of Chinese Academy of Sciences (E33Y0HB42P1).

Data Availability Statement: The data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

Acknowledgments: The authors thank Zonghao Han from Northwestern Polytechnical University for publishing the AVIID public dataset (https://pan.baidu.com/s/1M2WlHt1qqAuQ5QnUQSQ5lw?pwd=si8j (accessed on 8 February 2026)) and Yiming Sun from Tianjin University for publishing the Drone Vehicle Datasets public dataset (https://github.com/VisDrone/DroneVehicle (accessed on 8 February 2026)).

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

References

1. Wang, N.; Liu, H.; Zhuo, Q.; Li, M.; Shang, C. A Brief Overview of the Application of Infrared Intelligent Detection in Precision Guidance. In *Tenth Symposium on Novel Optoelectronic Detection Technology and Applications*; SPIE: Bellingham, WA, USA, 2025; Volume 13511, pp. 8–14.
2. Pineda, M.; Barón, M.; Pérez-Bueno, M.L. Thermal Imaging for Plant Stress Detection and Phenotyping. *Remote Sens.* **2021**, *13*, 68. [CrossRef]
3. Li, N.; Wang, H.; Zhao, H.; Ou, W. Cross-Modal Visible-to-Infrared Image Translation in Remote Sensing Guided by Thermal Features. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5628916. [CrossRef]
4. Zhao, D.; Zhou, L.; Li, Y.; He, W.; Arun, P.V.; Zhu, X.; Hu, J. Visibility Estimation via Near-Infrared Bispectral Real-Time Imaging in Bad Weather. *Infrared Phys. Technol.* **2024**, *136*, 105008. [CrossRef]
5. Liu, J.; Wu, G.; Liu, Z.; Wang, D.; Jiang, Z.; Ma, L.; Zhong, W.; Fan, X.; Liu, R. Infrared and Visible Image Fusion: From Data Compatibility to Task Adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 2349–2369. [CrossRef] [PubMed]
6. Ji, M.; Wang, Y.; Li, X. IR-NeLF: Infrared Neural Light Field on Mobile Devices. *Opt. Express* **2025**, *33*, 45967–45981. [CrossRef] [PubMed]
7. Bie, S.H.; Xie, J.T.; Zhang, Y.X.; Li, M.F.; Yu, W.K.; Chen, X.H.; Zhang, J. Cross-Waveband Optical Computing Imaging. *Opt. Lett.* **2024**, *49*, 2413–2416. [CrossRef]
8. Zhao, D.; Zhong, W.; Ge, M.; Jiang, W.; Zhu, X.; Arun, P.V.; Zhou, H. SiamBSI: Hyperspectral Video Tracker Based on Band Correlation Grouping and Spatial–Spectral Information Interaction. *Infrared Phys. Technol.* **2025**, *151*, 106063. [CrossRef]
9. Liu, Y.; Liu, Y.; Yan, S.; Chen, C.; Zhong, J.; Peng, Y.; Zhang, M. A Multi-View Thermal–Visible Image Dataset for Cross-Spectral Matching. *Remote Sens.* **2022**, *15*, 174. [CrossRef]
10. Zhao, D.; Hu, B.; Jiang, W.; Zhong, W.; Arun, P.V.; Cheng, K.; Zhao, Z.; Zhou, H. Hyperspectral Video Tracker Based on Spectral Difference Matching Reduction and Deep Spectral Target Perception Features. *Opt. Lasers Eng.* **2025**, *194*, 109124. [CrossRef]
11. Sokolik, I.N.; Toon, O.B.; Bergstrom, R.W. Modeling the Radiative Characteristics of Airborne Mineral Aerosols at Infrared Wavelengths. *J. Geophys. Res. Atmos.* **1998**, *103*, 8813–8826. [CrossRef]
12. Moncet, J.L.; Uymin, G.; Lipton, A.E.; Snell, H.E. Infrared Radiance Modeling by Optimal Spectral Sampling. *J. Atmos. Sci.* **2008**, *65*, 3917–3934. [CrossRef]
13. Allamandola, L.J.; Hudgins, D.M.; Sandford, S.A. Modeling the Unidentified Infrared Emission with Combinations of Polycyclic Aromatic Hydrocarbons. *Astrophys. J.* **1999**, *511*, L115. [CrossRef] [PubMed]
14. Jacobs, P.A.M. *Simulation of the Thermal Behaviour of an Object and Its Nearby Surroundings*; Physics Laboratory TNO (TNO Physisch Laboratorium) 1980-08; TNO: The Hague, The Netherlands, 1980.
15. Sheffer, A.D.; Cathcart, J.M. Computer Generated IR Imagery: A First Principles Modeling Approach. In *Multispectral Image Processing and Enhancement*; SPIE: Bellingham, WA, USA, 1988; Volume 933, pp. 199–206.
16. Balfour, L.S.; Bushlin, Y. Semi-Empirical Model-Based Approach for IR Scene Simulation. In *Infrared Technology and Applications XXIII*; SPIE: Bellingham, WA, USA, 1997; Volume 3061, pp. 616–623.
17. Gonda, T.G.; Less, D.M.; Filbee, D.R.; Polsen, E.S.; Edwards, T. An Exploration of Vehicle-Terrain Interaction in IR Synthetic Scenes. In *Targets and Backgrounds IX: Characterization and Representation*; SPIE: Bellingham, WA, USA, 2003; Volume 5075, pp. 9–19.
18. Luo, X.; Wang, S.; Liu, J.; Dong, X.; He, P.; Yang, Q.; Chen, X.; Zhou, F.; Zhang, T.; Feng, S.; et al. Revolutionizing Optical Imaging: Computational Imaging via Deep Learning. *Photonics Insights* **2025**, *4*, R03. [CrossRef]
19. Zhang, Y.; Cai, Z. CE-RetinaNet: A Channel Enhancement Method for Infrared Wildlife Detection in UAV Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4104012. [CrossRef]
20. Wang, H.; Li, N.; Zhao, H.; Wen, Y.; Su, Y.; Fang, Y. MappingFormer: Learning Cross-Modal Feature Mapping for Visible-to-Infrared Image Translation. In *Proceedings of the 32nd ACM International Conference on Multimedia*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 10745–10754.
21. Xu, Z.D. Study on Coupled Thermal Infrared Characteristics of Tank Target and Complex Ground Background. Ph.D. Thesis, Nanjing University of Science & Technology, Nanjing, China, 2023.
22. Zhou, K.; Hassan, F.H.; Hoon, G.K. The State of the Art for Cross-Modal Retrieval: A Survey. *IEEE Access* **2023**, *11*, 138568–138589. [CrossRef]
23. Zhao, D.; Tang, L.; Arun, P.V.; Asano, Y.; Zhang, L.; Xiong, Y.; Tao, X.; Hu, J. City-Scale Distance Estimation via Near-Infrared Trispectral Light Extinction in Bad Weather. *Infrared Phys. Technol.* **2023**, *128*, 104507. [CrossRef]
24. Han, T.; Wu, J.; Luo, W.; Wang, H.; Jin, Z.; Qu, L. Review of Generative Adversarial Networks in Mono- and Cross-Modal Biomedical Image Registration. *Front. Neuroinform.* **2022**, *16*, 933230. [CrossRef]
25. Jiang, J.; Li, P.; Wang, J.; Chen, H.; Zhang, T. Asphalt Pavement Crack Detection Based on Infrared Thermography and Deep Learning. *Int. J. Pavement Eng.* **2024**, *25*, 2295906. [CrossRef]

26. Zhao, D.; Asano, Y.; Gu, L.; Sato, I.; Zhou, H. City-Scale Distance Sensing via Bispectral Light Extinction in Bad Weather. *Remote Sens.* **2020**, *12*, 1401. [CrossRef]
27. Qu, J.; Chen, W.; Luo, Y. Research on Ship Recognition Technology Based on Infrared Image. In *International Conference on Image, Signal Processing, and Pattern Recognition (ISPP 2023)*; SPIE: Bellingham, WA, USA, 2023; Volume 12707, pp. 40–49.
28. Mao, F.; Mei, J.; Lu, S.; Liu, F.; Chen, L.; Zhao, F.; Hu, Y. PID: Physics-Informed Diffusion Model for Infrared Image Generation. *Pattern Recognit.* **2026**, *169*, 111816. [CrossRef]
29. Huang, Y.; Miyazaki, T.; Liu, X.; Omachi, S. Infrared Image Super-Resolution: A Systematic Review and Future Trends. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 25439–25463. [CrossRef]
30. Özer, S.; Ndigande, A.P. VisIRNet: Deep Image Alignment for UAV-Taken Visible and Infrared Image Pairs. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5403111. [CrossRef]
31. Yang, X.; Xie, D.; Wang, X. Crossing-Domain Generative Adversarial Networks for Unsupervised Multi-Domain Image-to-Image Translation. In *MM'18, Proceedings of the 26th ACM International Conference on Multimedia*; Association for Computing Machinery: New York, NY, USA, 2018; pp. 374–382.
32. Chen, Z.; He, G.; Li, J.; Liao, Y.; Gryllias, K.; Li, W. Domain Adversarial Transfer Network for Cross-Domain Fault Diagnosis of Rotary Machinery. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 8702–8712. [CrossRef]
33. Tumanyan, N.; Geyer, M.; Bagon, S.; Dekel, T. Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 1921–1930.
34. He, Y.; Jin, X.; Jiang, Q.; Cheng, Z.; Wang, P.; Zhou, W. LKAT-GAN: A GAN for Thermal Infrared Image Colorization Based on Large Kernel and AttentionUNet-Transformer. *IEEE Trans. Consum. Electron.* **2023**, *69*, 478–489. [CrossRef]
35. Kniaz, V.V.; Knyaz, V.A.; Hladuvka, J.; Kropatsch, W.G.; Mizginov, V. ThermalGAN: Multimodal Color-to-Thermal Image Translation for Person Re-Identification in Multispectral Dataset. In *Computer Vision – ECCV 2018 Workshops, Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8 September 2018*; Springer: Cham, Switzerland, 2019; pp. 606–624.
36. Ma, D.; Xian, Y.; Li, B.; Li, S.; Zhang, D. Visible-to-Infrared Image Translation Based on an Improved CGAN. *Vis. Comput.* **2024**, *40*, 1289–1298. [CrossRef]
37. Ma, Y.; Hua, Y.; Zuo, Z. Infrared Image Generation by Pix2pix Based on Multi-Receptive Field Feature Fusion. In *International Conference on Control, Automation and Information Sciences (ICCAIS)*; IEEE: New York, NY, USA, 2021; pp. 1029–1036.
38. Yang, G.; Stewart, C.V.; Sofka, M.; Tsai, C.L. Registration of Challenging Image Pairs: Initialization, Estimation, and Decision. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 1973–1989. [CrossRef]
39. Jin, C.-B.; Kim, H.; Liu, M.; Jung, W.; Joo, S.; Park, E.; Ahn, Y.S.; Han, I.H.; Lee, J.I.; Cui, X. Deep CT to MR Synthesis Using Paired and Unpaired Data. *Sensors* **2019**, *19*, 2361. [CrossRef]
40. Pizzati, F.; de Charette, R.; Zaccaria, M.; Cerri, P. Domain Bridge for Unpaired Image-to-Image Translation and Unsupervised Domain Adaptation. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*; IEEE: New York, NY, USA, 2020; pp. 2990–2998.
41. Alqahtani, H.; Kavakli-Thorne, M.; Kumar, G. Applications of Generative Adversarial Networks (GANs): An Updated Review. *Arch. Comput. Methods Eng.* **2021**, *28*, 525–552. [CrossRef]
42. Zhu, J.-Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *IEEE International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2017; pp. 2223–2232.
43. Kim, T.; Cha, M.; Kim, H.; Lee, J.K.; Kim, J. Learning to Discover Cross-Domain Relations with Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2017; pp. 1857–1865.
44. Yi, Z.; Zhang, H.; Tan, P.; Gong, M. DualGAN: Unsupervised Dual Learning for Image-to-Image Translation. In *IEEE International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2017; pp. 2849–2857.
45. Nelson, S.; Menon, R. Bijective-Constrained Cycle-Consistent Deep Learning for Optics-Free Imaging and Classification. *Optica* **2022**, *9*, 26–31. [CrossRef]
46. Zhao, Y.; Wu, R.; Dong, H. Unpaired Image-to-Image Translation Using Adversarial Consistency Loss. In *Computer Vision—ECCV 2020*; Springer: Cham, Switzerland, 2020; pp. 800–815.
47. Park, T.; Efros, A.A.; Zhang, R.; Zhu, J.Y. Contrastive Learning for Unpaired Image-to-Image Translation. In *Computer Vision—ECCV 2020*; Springer: Cham, Switzerland, 2020; pp. 319–345.
48. Zhang, Y.; Wang, S.; Chen, B.; Cao, J. GCGAN: Generative Adversarial Nets with Graph CNN for Network-Scale Traffic Prediction. In *International Joint Conference on Neural Networks (IJCNN)*; IEEE: New York, NY, USA, 2019; pp. 1–8.
49. Han, Z.; Zhang, S.; Su, Y.; Chen, X.; Mei, S. DR-AVIT: Toward Diverse and Realistic Aerial Visible-to-Infrared Image Translation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5004213. [CrossRef]
50. Wang, X.; Chen, H.; Tang, S.A.; Wu, Z.; Zhu, W. Disentangled Representation Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 9677–9696. [CrossRef] [PubMed]

51. Chartsias, A.; Joyce, T.; Papanastasiou, G.; Semple, S.; Williams, M.; Newby, D.E.; Dharmakumar, R.; Tsaftaris, S.A. Disentangled Representation Learning in Cardiac Image Analysis. *Med. Image Anal.* **2019**, *58*, 101535. [CrossRef]
52. Gonzalez-Garcia, A.; Van De Weijer, J.; Bengio, Y. Image-to-Image Translation for Cross-Domain Disentanglement. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 1294–1305.
53. Liu, A.H.; Liu, Y.C.; Yeh, Y.Y.; Wang, Y.C.F. A Unified Feature Disentangler for Multi-Domain Image Translation and Manipulation. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montréal, QC, Canada, 3–8 December 2018; 31, pp. 2595–2604.
54. Han, Z.; Zhang, Z.; Zhang, S.; Zhang, G.; Mei, S. Aerial Visible-to-Infrared Image Translation: Dataset, Evaluation, and Baseline. *J. Remote Sens.* **2023**, *3*, 0096. [CrossRef]
55. Sun, Y.; Cao, B.; Zhu, P.; Hu, Q. Drone-Based RGB-Infrared Cross-Modality Vehicle Detection via Uncertainty-Aware Learning. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 6700–6713. [CrossRef]
56. Huang, X.; Liu, M.Y.; Belongie, S.; Kautz, J. Multimodal Unsupervised Image-to-Image Translation. In *Computer Vision—ECCV 2018*; Springer: Cham, Switzerland, 2018; pp. 179–196.
57. Lee, H.Y.; Tseng, H.Y.; Huang, J.B.; Singh, M.; Yang, M.H. Diverse Image-to-Image Translation via Disentangled Representations. In *Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018*; 2018; pp. 35–51.
58. Chang, H.Y.; Wang, Z.; Chuang, Y.Y. Domain-Specific Mappings for Generative Adversarial Style Transfer. In *European Conference on Computer Vision*; Computer Vision Foundation: New York, NY, USA, 2020; pp. 573–589.

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.