Journal of Remote Sensing
A SCIENCE PARTNER JOURNAL

RESEARCH ARTICLE

Aerial Visible-to-Infrared Image Translation: Dataset, Evaluation, and Baseline

Zonghao Han, Ziye Zhang, Shun Zhang, Ge Zhang, and Shaohui Mei*

School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China.

*Address correspondence to: meish@nwpu.edu.cn

Citation: 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:Article 0096. https://doi.org/10.34133/remotesensing.0096

Submitted 5 April 2023
Accepted 19 October 2023
Published 10 November 2023

Aerial visible-to-infrared image translation aims to transfer aerial visible images to their corresponding infrared images, which can effectively generate the infrared images of specific targets. Although some image-to-image translation algorithms have been applied to color-to-thermal natural images and achieved impressive results, they cannot be directly applied to aerial visible-to-infrared image translation due to the substantial differences between natural images and aerial images, including shooting angles, multi-scale targets, and complicated backgrounds. In order to verify the performance of existing image-to-image translation algorithms on aerial scenes as well as advance the development of aerial visible-to-infrared image translation, an Aerial Visible-to-Infrared Image Dataset (AVIID) is created, which is the first specialized dataset for aerial visible-to-infrared image translation and consists of over 3,000 paired visible-infrared images. Over the constructed AVIID, a complete evaluation system is presented to evaluate the generated infrared images from 2 aspects: overall appearance and target quality. In addition, a comprehensive survey of existing image-to-image translation approaches that could be applied to aerial visible-to-infrared image translation is given. We then provide a performance analysis of a set of representative methods under our proposed evaluation system on AVIID, which can serve as baseline results for future work. Finally, we summarize some meaningful conclusions, problems of existing methods, and future research directions to advance state-of-the-art algorithms for aerial visible-to-infrared image translation.

Introduction

With the rapid development of infrared technology, the infrared camera equipped on unmanned aerial vehicles (UAVs) is increasingly applied for aerial photography. Aerial infrared images have been widely used in the military and in industrial, agricultural, and environmental settings, such as moving target detection [1–3] and tracking [4–6], photovoltaic panel error detection [7–9], image registration [10–12], and visible-infrared image fusion [13–16] because of their advantages, including high sensitivity to temperature variation, strong capability to penetrate through the fog, and powerful robustness when encountering the weak light condition.

Due to the high cost of an infrared camera or the limitations of taking photography conditions, obtaining many aerial infrared images of some specific targets is challenging. In this case, the mainstream method to obtain aerial infrared images is to employ the simulation software platform for target scene infrared simulation [17–21]. These methods first analyze the target attributes to obtain a simulated 3D model scene and then compute the infrared radiation distribution of different materials in the scene according to the infrared radiation theory. Next, the radiation attenuation of the infrared radiation to the detector is calculated by the atmospheric transmission model. The imaging characteristics of the imaging sensor are then simulated and added to the infrared radiation distribution. Finally, the simulated scene is gray-scaled to produce the final infrared image.

Compared with actual photography, the use of infrared simulation software to simulate aerial infrared images of targets can significantly save manpower, material resources, and financial capacity. At the same time, the simulated infrared images with various periods and different bands can be obtained by adjusting the parameters of the infrared radiation distribution model and the imaging sensor. However, these methods have problems such as low simulation degree of the target temperature model, huge intermediate parameters, high coupling degree of each system, and complicated processing procedures, which could not be suitable for quickly obtaining many aerial infrared images. In this paper, we propose a new task called aerial visible-to-infrared image translation, which aims to generate aerial infrared images from visible images and has 3 main advantages:

- Due to the easy acquisition and lower photography cost of aerial visible images, aerial visible images can be translated into corresponding infrared images in a fast, efficient, and low-cost manner.
- Additional modality information can be provided by the aerial visible images to improve the performance of the aerial infrared images in downstream tasks.
- The translated aerial infrared and corresponding visible images can provide paired data support for cross-modality and domain adaptation tasks.

Though translating aerial visible images into corresponding infrared images has the advantage in terms of efficiency and speed compared to actually taking photography and infrared

simulation, 3 significant issues seriously limit the development of aerial visible-to-infrared image translation.

• **Lacking an available dataset for aerial visible-to-infrared image translation experiments:** So far, most datasets consist of color images and lack paired infrared images. Although there are several color-to-thermal datasets [22,23], they are all natural images, not taken from an aerial perspective, without diverse targets and complicated backgrounds like aerial images. Therefore, to the best of our knowledge, there are currently no available datasets for aerial visible-to-infrared image translation.

• **Lacking a survey of methods that could apply to aerial visible-to-infrared image translation:** The translation of aerial visible-to-infrared images can be considered as cross-modality learning, which makes it challenging to model the mapping. As far as we know, no specific approaches have been proposed to solve this problem. Therefore, a survey of methods that can be effectively applied to aerial visible-to-infrared image translation remains to be clarified.

• **Lacking a complete evaluation system to evaluate the quality of generated images:** Existing metrics for evaluating the similarity between images are mainly traditional perceptual indicators, such as MSE, peak signal-to-noise ratio (PSNR), and SSIM. However, they are too shallow functions to account for many nuances of human perception. In addition, evaluating the quality of the generated images only from the similarity of the appearance is obviously unreasonable. A more complete evaluation system to evaluate the quality of generated images is necessary.

In order to address the above issues and fully advance the development of aerial visible-to-infrared image translation, we propose a new specific dataset for aerial visible-to-infrared image translation, called AVIID (Aerial Visible-to-Infrared Image Dataset), consisting of over 3,000 paired visible-infrared images. The goal of AVIID is to provide researchers with an available data resource to evaluate and improve state-of-the-art algorithms. The aerial visible-to-infrared image translation aims to learn a mapping between 2 image domains, which can be regarded as a cross-modality image-to-image translation problem. Recently, image-to-image translation algorithms [16,24–42] among color image domains with the application of deep convolutional neural networks (CNNs) [43–45] and generative adversarial networks (GANs) [46–50] have made significant progress in a wide range of tasks, including style transfer [40,51,52], image inpainting [53], colorization [54], super-solution [55–58], dehazing [59,60], and denoising [61,62]. Some researchers have applied image-to-image translation approaches to color-to-thermal image translation tasks [22,63,64] and achieved impressive results. For example, Kniaz and Knyaz [65] achieve multi-spectral person re-identification by using GAN for color-to-thermal image translation. In this paper, we attempt to apply these image-to-image translation approaches to aerial visible-to-infrared image translation and make a comprehensive survey of these methods. In addition, we propose a complete evaluation system to evaluate the generated infrared images from the overall appearance and target quality. The overall appearance aims to determine the similarity between the generated infrared images and real ones from the visual perception. The target quality reflects the quality of the targets in the generated infrared images, which is important for some downstream tasks such as object detection and tracking. We further evaluate several representative image-to-image translation methods on AVIID under this proposed complete evaluation system, and the results can be seen as a baseline to advance the development of aerial visible-to-infrared image translation.

In summary, the major contributions of this paper are as follows:

• The first specific dataset for aerial visible-infrared image translation, AVIID, is constructed, which provides researchers with an available data resource to evaluate and advance state-of-the-art algorithms.

• A comprehensive survey of up-to-date image-to-image translation algorithms that could be applied to aerial visible-to-infrared image translation is proposed to promote the development of this field.

• A complete evaluation system is presented to evaluate the generated infrared images in terms of the overall appearance and target quality. Several representative image-to-image translation methods are evaluated on AVIID under our proposed complete evaluation system. These results can be regarded as a baseline for future work.

• Some meaningful conclusions, problems of existing methods, and future research directions are summarized to advance state-of-the-art algorithms for aerial visible-to-infrared image translation.

The rest of this paper is organized as follows. We first provide a comprehensive survey of image-to-image translation methods that can be applied to aerial visible-to-infrared image translation in the "A Survey of Methods for Aerial Visible-to-Infrared Image Translation" section. The details of AVIID are then described in the "A Specific Dataset for Aerial Visible-to-Infrared Image Translation" section. In the "Experiments and Results" section, the description of our proposed complete evaluation system and baseline results of representative methods on AVIID are given. Finally, the conclusion of our work is given in the "Conclusion" section.

A Survey of Methods for Aerial Visible-to-Infrared Image Translation

In this section, we comprehensively make a survey of image-to-image translation methods that could be applied in aerial visible-to-infrared translation. Based on whether the method depends on paired images or not, we simply classify these methods into supervised and unsupervised categories. Supervised methods aim to learn a pixel-level mapping from the source domain to the target domain with the paired data for training, which limits their applications. In contrast, unsupervised methods only need 2 images from 2 different domains as training data to achieve image-to-image translation by adopting additional constraints. According to whether multi-modal outputs are generated based on one single image as input or not, these unsupervised methods can be further divided into 2 types: one-to-one (single modal) and one-to-many (multi-modal). In addition, depending on the mapping relationship between the source and target domains, one-to-one unsupervised approaches can be further classified into 1-sided and 2-sided methods. One-sided unsupervised image-to-image translation methods can only translate the images from the source domain to the target domain. In contrast, 2-sided ones can achieve a bidirectional mapping between the source domain and the target domain. Figure 1 shows an overview of these methods. In what follows, we will introduce each category of these methods in detail.

Supervised image-to-image translation methods

Supervised image-to-image translation methods aim to learn a pixel-level mapping to achieve image translation from one domain to another based on paired data. Paired data means the training data are paired, and every image from the source domain has a corresponding image in the target domain. In this case, Pix2Pix is the first method to achieve task-agnostic image translation, which uses a conditional generative adversarial network (cGAN) [21] to learn a mapping from input images to output images. Based on the framework of Pix2Pix, BicycleGAN adds a variational autoencoder (VAE) in cGAN to generate multiple outputs from a single input image. Additional details of Pix2Pix and BicycleGAN are as follows.

Pix2Pix [24]**:** Pix2Pix investigates cGANs, a variant of GAN, as a general solution to image-to-image translation problems. The key idea of GAN is to simultaneously train the discriminator and the generator: the discriminator is designed to distinguish between the real data and the generated samples, while the generator aims to generate the fake samples that are as real as possible in order to convince the discriminator that the fake samples come from the real data. Given the paired image data (x, y), where x is from the source domain X and y is from the target domain Y. The cGANs aim to learn a mapping from the image x with a random latent vector z to the image y, $y = G(x, z)$. The generator G is trained to produce outputs that cannot be distinguished from the "real" images in the target domain with an adversarial discriminator, D, which is trained to detect the generator's "fakes" as soon as possible. The full objective of the cGANs can be expressed as

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{(x,y)}\left[\log D(x, y)\right] + \mathbb{E}_{(x,z)}\left[\log(1 - D(x, G(x, z)))\right], \quad (1)$$

where G attempts to minimize this objective versus an adversarial D that tries to maximize it. In addition, Pix2Pix adds an additional L_1 distance constraint to the generator to make the

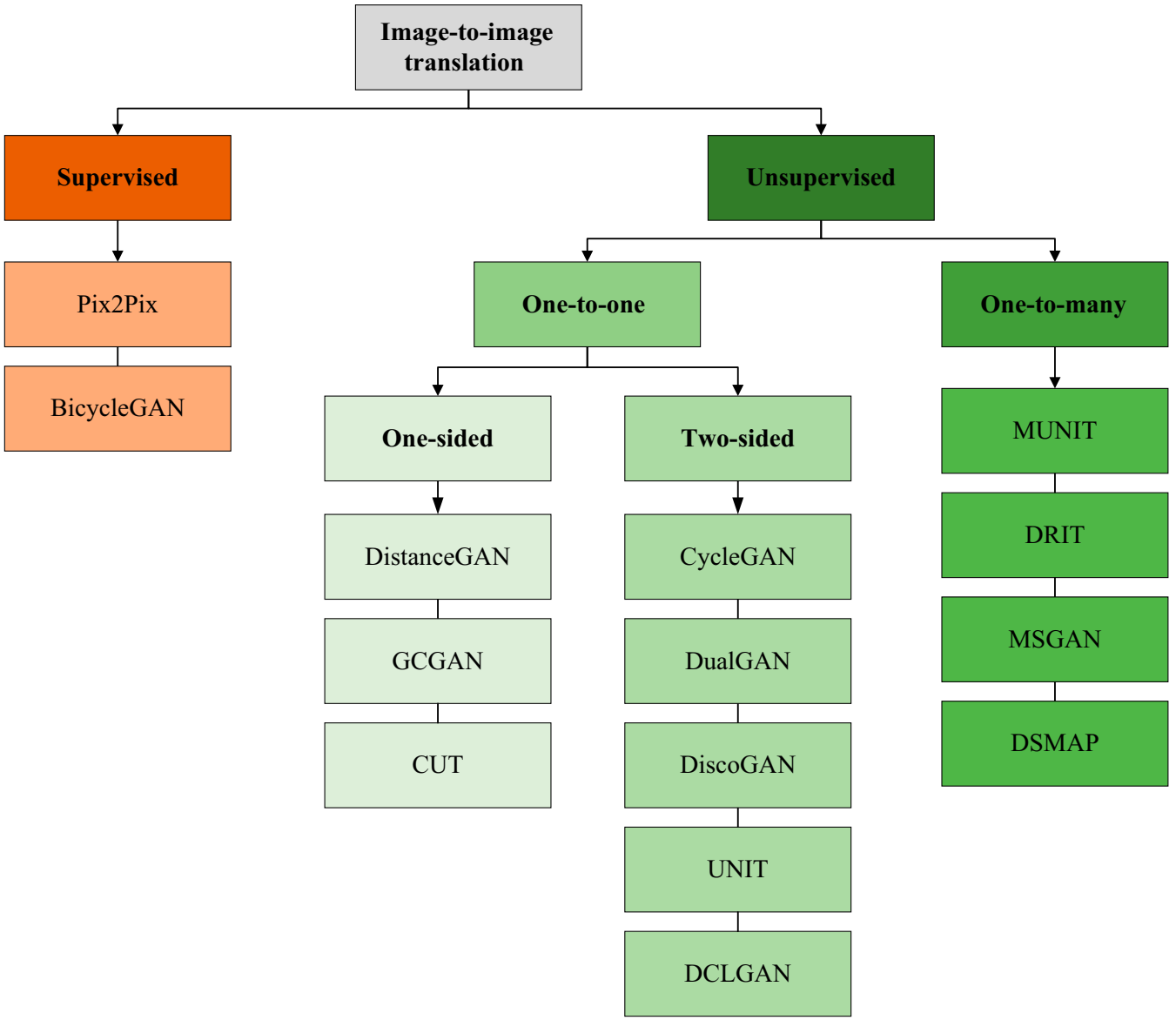

Fig. 1. Overview of image-to-image translation methods that could be applied to aerial visible-to-infrared image translation. Each color represents a category.

translated image visually similar to its corresponding ground truth, which can be formulated as

$$\mathcal{L}_{L_1}(G) = \mathbb{E}_{(x,y,z)}\left[\| y - G(x,z) \|_1\right]. \quad (2)$$

Therefore, the final objective of Pix2Pix can be formulated as

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G,D) + \lambda \mathcal{L}_{L_1}(G), \quad (3)$$

where λ is a hyperparameter.

BicycleGAN [26]**:** Though Pix2Pix has achieved ambiguous results for image-to-image translation, it is prone to suffer from mode collapse, resulting in generating very similar images. To address this issue, BicycleGAN aims to enhance the relationship between the output with the latent code, which helps to produce more diverse results. For the paired image data (x, y), BicycleGAN first maps the target domain image y to a specific latent code z by a VAE encoder, $z = E(y)$. The latent code is encoded by the real data in the training process, but a random latent code may not yield realistic images at testing time. To avoid this, an additional *KL* loss is used to align the distribution of the latent code with the standard normal distribution. Then, BicycleGAN combines the latent code with the input image to translate it from the source domain to the target domain by cGANs like in Pix2Pix, $\hat{y} = G\left(E\left(y\right), x\right)$. The translated image $\hat{y}$ is not necessarily needed to be close to the ground truth, which may suffer from mode collapse, but must be realistic. To achieve this, BicycleGAN recovers the latent code by the VAE encoder, $\hat{z} = E\left(\hat{y}\right)$, and utilizes an L_1 loss to keep the consistency between the recovered and the original latent code, which can be expressed as

$$\mathcal{L}_1^{latent} = \| E\left(y\right) - E\left(G\left(E\left(y\right), x\right)\right) \|_1. \quad (4)$$

Unsupervised image-to-image translation methods

One-to-one

Unsupervised image-to-image translation algorithms aim to learn a joint distribution by using images from the marginal distributions in individual domains. Since there exists an infinite set of possible joint distributions that can arrive at the marginal distributions, it is impossible to guarantee that a particular input and output correspond in a meaningful way without additional assumptions or constraints. As a consequence, various constraints have been proposed to achieve unsupervised image-to-image translation.

DistanceGAN assumes that the distance between 2 images in the source domain should be preserved after mapping them to the target domain. GCGAN develops a geometry-consistency constraint from the special property of images that simple geometric transformations will not change the semantic structure of images. CUT proposes a contrastive learning-based constraint to maximize the mutual information between the input and the output. These methods can be seen as one-sided unsupervised image-to-image translation because the mapping from the source domain to the target domain is unidirectional. In addition, some methods construct various specific constraints to achieve 2-sided unsupervised image-to-image translation. For example, CycleGAN, DualGAN, and DiscoGAN employ the cycle-consistency constraint, which aims to transfer an image in the source domain to the target domain, and this translated image can also be transferred back to the source domain. UNIT makes a shared-latent space assumption that also implies the cycle-consistency constraint. DCLGAN takes advantage of CycleGAN and CUT, employing the idea of mutual information maximization to enable 2-sided unsupervised image-to-image translation. More details of these methods are as follows.

DistanceGAN [37]**:** Let $x \in X$ denote a random image from the source domain, and $y \in Y$ represents a random target domain image. Unsupervised training data pairs are expressed as (x_i, y_j), $i = 1, 2, \ldots, N$, where N means the size of the dataset. DistanceGAN presents a distance-preserving mapping, which aims to enforce that the distance between images in the source domain is preserved after mapping them to the target domain and can be formulated as

$$d\left(x_i, x_j\right) \approx a \cdot d\left(G_{XY}\left(x_i\right), G_{XY}\left(x_j\right)\right) + b, \quad (5)$$

where $d(.)$ is a predefined metric function to measure the distance between 2 samples, a and b are the linear coefficient and bias, and $G_{XY}(.)$ is the generator.

GCGAN [36]**:** GCGAN presents a geometry-consistency constraint in that a given specific geometric between the input images should be preserved after transferring them to the target domain. In detail, given a random image x from the source domain X, a specific geometric transformation $f(.)$, and 2 related translators G_{XY} and $G_{\tilde{X}\tilde{Y}}$, the geometry-consistency constraint can be expressed as

$$\begin{aligned} G_{XY}(x) &\approx f^{-1}\left(G_{\tilde{X}\tilde{Y}}\left(f(x)\right)\right), \\ G_{\tilde{X}\tilde{Y}}\left(f(x)\right) &\approx f\left(G_{XY}(x)\right), \end{aligned} \quad (6)$$

where $f^{-1}(.)$ is the inverse of the transformation $f(.)$.

CUT [40]**:** CUT proposes a novel constraint to maximize the mutual information between the corresponding input and output patches based on the intuition that each path in the output should reflect the content of the counterpart patch in the input and be independent of the domain. To achieve this, CUT uses a type of contrastive learning loss function, InfoNCE loss [66], which aims to learn an embedding that associates a patch of the output v and its corresponding patch of the input v^+, while separating it from the other N noncorresponding patches of the input v^-, which can be formulated as

$$\mathcal{L}_{InfoNCE} = -\log\left[\frac{exp\left(v \cdot v^+ / \tau\right)}{exp\left(v \cdot v^+ / \tau\right) + \sum_{n=1}^{N} exp\left(v . v_n^- / \tau\right)}\right], \quad (7)$$

where τ is a temperature hyperparameter. Intuitively, this loss can be seen as a classifier that attempts to classify v as v^+.

CycleGAN [51]**/DualGAN** [29]**/DiscoGAN** [67]**:** CycleGAN, DualGAN, and DiscoGAN propose the cycle-consistency constraint to achieve the 2-sided unsupervised image-to-image translation. These methods construct 2 translators to learn 2 mappings simultaneously via transferring an image to the target domain and back, maintaining the fidelity of the input and the reconstructed image through the cycle-consistency constraint. Mathematically, for an image x from the source domain X, the translator G_{XY} translates it to the target domain Y, and then this translated image is transferred back to the source domain by the translator G_{YX}, and the cycle-consistency constraint is used to preserve the semantic structure of the reconstructed image and the input. For the domain Y, it is an inverse process and the whole objective of cycle-consistency constraint can be expressed as

$$\mathcal{L}_{cycle-consistency} = \| x - G_{YX}(G_{XY}(x)) \|_1 + \| y - G_{XY}(G_{YX}(y)) \|_1. \quad (8)$$

UNIT [25]**:** UNIT presents a shared-latent space assumption, which assumes that a pair of corresponding images from different domains can be mapped to the same latent representation in a shared-latent space. Consequently, the latent code can be computed from each of the images, and these 2 images can also be recovered from the shared latent code. Based on this assumption, UNIT proposes a 2-sided unsupervised image-to-image translation framework consisting of 6 sub-networks, including 2 domain image encoders E_X and E_Y, 2 domain generators G_X and G_Y, and 2 domain discriminators D_X and D_Y. For any given pair of image data (x, y), the shared latent code can be obtained by encoders $z = E_X(x) = E_Y(y)$, and conversely, the images can be recovered from this latent code, $x = G_X(E_Y(y)$ and $y = G_Y(E_X(x)$. In this way, images from the source and target domains can be mutually transferred. However, to achieve this, a necessary condition to exist is the cycle-consistency constraint: $x = G_X(G_Y(E_X(x))$ and $y = G_Y(G_X(E_Y(y))$. Therefore, from this perspective, the shared-latent space assumption also implies the cycle-consistency constraint.

DCLGAN [34]**:** Although the cycle-consistency constraint can ensure that the translated images have similar semantic information compared to the target domain, it enforces the relationship between the 2 domains to be bijective, which is too restrictive. At the same time, CUT has demonstrated the effectiveness of contrastive learning in one-sided unsupervised image-to-image translation. However, one embedding for 2 separate domains may not capture the domain gap. To solve this, DCLGAN takes advantage of CycleGAN and CUT to propose a novel method based on contrastive learning and a dual learning setting to enable an efficient 2-sided domain mapping with unpaired data.

One-to-many

Though several methods have enabled unpaired image-to-image translation, they fail to generate multi-modal results. An effective way to handle multi-modal image-to-image translation is to perform image translation conditioned on the input image and a specific latent code. To achieve this, DRIT/DRIT++ and MUNIT assume that the image representation can be disentangled into 2 spaces: a domain-invariant content space capturing shared information across domains and a domain-specific style space. Then, to achieve translation, they recombine its content information with a random style feature sampled from the style space of the target domain. To improve the diversity, MSGAN presents a mode-seeking regularization term that maximizes the ratio of the distance between translated images with respect to the distance between latent vectors. DSMAP leverages domain-specific mappings for remapping latent features in the shared content space to domain-specific content spaces, which is conducive to achieve more challenging style transfer tasks that require more attention on local and structural–semantic correspondences. These methods are described in detail as follows.

DRIT [52]**/DRIT++**[35]**/MUNIT** [27]**:** DRIT/DRIT++ and MUNIT assume that images from 2 domains can be decomposed into a domain-invariant content space and a domain-specific style space. The domain-invariant content space captures the shared information across 2 domains, while the style space captures domain-specific attributes. To transfer an image from the source domain to the target domain, they recombine its content code with a random style code sampled from the target domain space. Mathematically, for a given unpaired image data (x, y) random sampled from the source domain X and the target domain Y, DRIT/DRIT++ and MUNIT first use the content encoders $\left(E_c^X, E_c^Y\right)$ and style encoders $\left(E_s^X, E_s^Y\right)$ to disentangle the images into the domain-variant content code, $z_c = E_c^X(x) = E_c^Y(y)$, and domain-specific style codes, $x_s = E_s^X(x)$ and $y_s = E_s^Y(y)$. Then, they perform a cross-domain mapping to obtain translated images $\left(\tilde{x}, \tilde{y}\right)$ by recombining the content code with the specific style code to the generator, $\tilde{y} = G_{XY}\left(E_c^X(x), E_s^Y\left(y\right)\right)$, and $\tilde{x} = G_{YX}\left(E_c^Y\left(y\right), E_s^X(x)\right)$, where G_{YX} and G_{XY} are cross-domain generators. After that, they apply the above cross-domain mapping one more time and leverage the cycle-consistency constraint to enforce the consistency between the reconstructed images and the original input images, which can be formulated as

$$\mathcal{L}_{cycle-consistency} = \| x - G_{YX}\left(E_c^Y\left(\tilde{y}\right), E_s^X(\tilde{x})\right) \|_1 + \| y - G_{XY}\left(E_c^X(\tilde{x}), E_s^Y\left(\tilde{y}\right)\right) \|_1. \quad (9)$$

MSGAN [68]**:** Existing cGANs tend to focus on conditional input images but ignore random latent vectors that significantly contribute to the diversity of outputs and thus suffer from mode collapse. To address this issue and improve the diversity of the generated images, MSGAN proposes a simple yet effective mode-seeking regularization term, which aims to maximize the ratio of the distance between generated images with respect to the corresponding latent vectors. Let an input image x from the domain X, 2 latent vectors z_1, z_2 from the latent space Z, and a cross-domain generator G_{XY} that translates the input image with the latent vectors to the target domain, respectively. Then, the mode-seeking regularization term directly maximizes the ratio of the distance between the translated images to the distance between the latent vectors, which can be expressed as

$$\mathcal{L}_{ms} = \max_{G_{XY}} \left(\frac{d\left(G_{XY}, \left(x, z_1\right) G_{XY}\left(x, z_2\right)\right)}{d\left(z_1 - z_2\right)} \right), \quad (10)$$

where $d(.)$ denotes the predefined distance metric.

DSMAP [39]**:** Previous multi-modal unsupervised image-to-image translation methods often assume that the image representation can be decomposed into a shared domain-variant content space and a domain-specific space. However, this content space only considers the shared information across domains but ignores the relationship between content and style, which may weaken the presentation of content. To address this issue, DSMAP leverages 2 additional domain-specific mapping functions to remap the content features in the shared domain-invariant content space into the domain-specific content spaces for different domains, which can be expressed as

$$\begin{aligned} x_c^Y &= \Phi_{C \to Y}\left(E_c^X(x)\right), \\ y_c^X &= \Phi_{C \to X}\left(E_c^Y\left(y\right)\right), \end{aligned} \quad (11)$$

where x, y are an unpaired image data randomly sampled from the domain X and domain Y, $\Phi_{C \to Y}$, $\Phi_{C \to X}$ are the

domain-specific mapping functions, and E_c^Y, E_c^X are the domain-invariant encoders. By these domain-specific mapping functions, the features in the shared content space could be aligned with the target domain to encode the domain-specific content features and thus improve the content representation ability for translation.

A Specific Dataset for Aerial Visible-to-Infrared Image Translation

In this section, we introduce AVIID, a specific dataset for aerial visible-to-infrared image translation in detail. AVIID consists of paired aerial visible and infrared images that are taken by a dual-light camera equipped on the UAV. Figure 2 shows the dual-light camera and the UAV. Table 1 describes the detailed parameters of the dual-light camera. Depending on the shooting time, various scenarios, and conditions of photography, we further divide AVIID into 3 subdatasets named AVIID-1, AVIID-2, and AVIID-3, respectively. Table 2 shows the overall comparison of the 3 subdatasets, and the details of them are described in the following.

AVIID-1

AVIID-1 contains 993 pairs of paired visible-infrared images with an image size of 434 × 434. The scenes of AVIID-1 are the roads, and the targets in the images are common vehicles, including cars, buses, vans, and trucks. These images are taken between 9 a.m. and 12 p.m. with temperatures ranging from 28°C to 32°C. When taking images, the height of the UAV is about 15 m, the distance from the road is about 90 m, and the shooting angle of the dual-light camera is 90° horizontally. The scenarios in these images are very similar, mainly including various cars, trees beside the road, and houses in the distance. Therefore, using this subdataset for aerial visible-infrared image translation is relatively simple. Figure 3 shows some examples of AVIID-1.

AVIID-2

AVIID-2 contains 1,090 pairs of paired visible-infrared images with an image size of 434 × 434. The taking conditions and scenes of AVIID-2 are the same as AVIID-1, except that this subdataset is taken from 8 p.m. to 10 p.m., and the temperatures are between 26°C and 28°C. The images of AVIID-2 are taken under low-light conditions, resulting in much noise in the images, and even blurry targets and backgrounds, which is challenging for aerial visible-to-infrared translation compared with AVIID-1. Some examples of AVIID-2 can be seen in Fig. 4.

AVIID-3

AVIID-3 contains 1,280 pairs of paired visible-infrared images with an image size of 512 × 512. These images are taken by the UAV at 3 different heights, including about 50 m, 100 m, and 150 m, and 2 different shooting angles of 45° and 60° vertically. The taking time is mainly from 2 p.m. to 5 p.m., and the temperatures are between 30°C and 34°C. Compared with AVIID-1 and AVIID-2, this dataset contains more types of vehicles and numerous targets of multiple densities, viewpoints, and scales. In addition, AVIID-3 is collected in various scenarios with more complicated backgrounds, including roads, bridges across rivers, parking lots, and streets of residential communities. Therefore, this dataset is more challenging for aerial visible-to-infrared image translation and can be better used to evaluate the performance of different methods. Some figures of AVIID-3 are displayed in Fig. 5.

Experiments and Results

In this section, we evaluate some representative image-to-image methods on AVIID. First, we present our experiment settings, including dataset usage, baseline methods, and training and testing procedure details. Then, our proposed complete evaluation system that evaluates generated images from 2 aspects, overall appearance and target quality, is introduced in detail. Finally, the baseline results are given for future work.

Settings

We conduct experiments on all 3 subdatasets and set the ratios of the training set to be 50% and 80%, respectively, the left data for testing. We select 10 representative methods as baseline methods for our experiments, 2 supervised methods, including Pix2Pix and BicycleGAN, and 8 unsupervised methods,

A B

Fig 2. (A) Dual-light camera. (B) UAV.

Table 1. Detailed parameters of the dual-light camera

Visible camera parameters		Infrared camera parameters	
Imaging sensor	Exmor CMOS	Lens barrel	25 mm
HDMI output	1,080p/59.94	Horizontal field	24. 6°
Signal-to-noise ratio	Above 50 dB	Vertical field	18. 5°
Optical zoom	30 times	Diagonal field angle	30. 4°
Digital zoom	12 times	Working form	Long wave (8–14 μm)
Viewing angle	63. 7° – 2. 3°	Image resolution	640 × 480
Minimum object distance	10 – 1,200 mm	Spatial resolution	0.617 mrad

including GCGAN, CUT, CycleGAN, UNIT, DCLGAN, MUNIT, DRIT, and MSGAN. In the training time, every image is first resized to 286 × 286, then random cropped to 256 × 256, and finally horizontally flipped with a probability of 0.5 for data augmentation. To train Pix2Pix, BicycleGAN, GCGAN, CUT, CycleGAN, and DCLGAN, we use the Adam optimizer with a learning rate of 0.0002 and a batch size of 4 for 1,000 epochs on NVIDIA RTX3090. For DRIT and MSGAN, the whole networks are also optimized by the Adam optimizer with a learning rate of 0.0001 for 1,200 epochs on GTX1080Ti and the batch size is also set to 4. With respect to UNIT and MUNIT, we use the Adam optimizer to train them for 200,000 iterations on NVIDIA RTX3090, the learning rate is 0.0001, the batch size is 4, and the weight decay is set to 0.0001. In the testing procedure, the input image is resized to 256 × 256 without any data augmentation.

Complete evaluation system

Overall appearance evaluation

In order to evaluate the overall appearance quality of the generated images, we adopt the most widely used traditional perceptual metrics, including MSE, PSNR, and SSIM. The details of these metrics are as follows.

MSE: MSE is used to evaluate the margin of the discrepancy between the pixels of the generated image and its ground truth, which can be defined as

$$MSE(y,\hat{y}) = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\left(y_{i,j} - \hat{y}_{i,j}\right)^2. \quad (12)$$

where y and $\hat{y}$ where y and $\hat{y}$represent the generated image and the corresponding real ones, and H and W are the height and width of the image, respectively.

PSNR: The PSNR aims to measure the degree of distortion for the generated image with respect to its corresponding ground truth, which can be expressed as

$$PSNR(y,\hat{y}) = 10\log_{10}\frac{\max(\hat{y})}{MSE(y,\hat{y})}, \quad (13)$$

where $max\left(\hat{y}\right)$ means the max pixel of the real image. Higher PSNR indicates a smaller distortion of the generated image.

SSIM: SSIM can estimate the structural similarity between the generated image and the real image, which can be formulated as

$$SSIM(y,\hat{y}) = \frac{\left(2\mu_y\mu_{\hat{y}} + c_1\right)\left(2\sigma_{y\hat{y}} + c_2\right)}{\left(\mu_y^2 + \mu_{\hat{y}}^2 + c_1\right)\left(\sigma_y^2 + \sigma_{\hat{y}}^2 + c_2\right)}, \quad (14)$$

where c_1 and c_2 are constant, μ_y, $\mu_{\hat{y}}$, σ_y, and $\sigma_{\hat{y}}$ are the mean and variance of the generated image and the ground truth, respectively, and $\sigma_{y\hat{y}}$ is their covariance. Higher SSIM means the generated image is more similar to its corresponding real image.

Though MSE, PSNR, and SSIM are the most widely used traditional perceptual metrics, they are relatively shallow functions and fail to account for many nuances of human perception. In recent years, regarding the deep features of deep CNN as a perceptual metric have been demonstrated to be an effective way and more consistent with human perception judgment. Therefore, to more accurately evaluate the quality of the generated images, we adopt three CNN-based perceptual metrics, including FID [69], KID [70], and LPIPS [71]. More details of FID, KID and LPIPS are as follows.

LPIPS: LPIPS is a CNN-based perceptual metric and has been demonstrated to coincide greatly with human judgment. It can be computed by a weighted L_2 distance between the deep features extracted by the deep CNN of the generated images and their ground truth

$$LPIPS\left(Y,\hat{Y}\right) = \frac{1}{N}\sum_{i=1}^{N}\sum_{l=1}^{L}\frac{1}{H_l W_l}\sum_{h,w}\| w_l \odot \left(y_{i,hw}^l - \hat{y}_{i,hw}^l\right)\|^2 \quad (15)$$

Table 2. The overall comparison of the 3 subdatasets

Subdataset	Paired	Various scenarios	Multi-scale targets	Time	Temperature	Image size	Images
AVIID-1	✓	✗	✗	Day	28–32 °C	434 × 434	993
AVIID-2	✓	✗	✗	Night	26 – 28°C	434 × 434	1,090
AVIID-3	✓	✓	✓	Day	30 – 34°C	512 × 512	1,280

Fig. 3. Some examples of AVIID-1. The scenes of AVIID-1 contain the roads with various kinds of vehicles, including cars, buses, vans, and trucks.

where Y and $\hat{Y}$ represent the generated images and the real ones, y^l and $\hat{y}^l$ are normalized deep features extracted from the l layer of the deep CNN, w_l means the weighted parameters, and N is the number of the images. We use the AlexNet pre-trained on the ImageNet as the deep feature extractor, and a lower LPIPS score indicates a better quantity of the generated images.

FID: FID is a widely used metric to estimate the distribution of real and generated images through deep features extracted by the last pooling layer of the Inception-V3 model trained on the ImageNet and compute the divergence between them, which can be formulated as

$$FID\left(Y,\hat{Y}\right) = \| m_Y - m_{\hat{Y}} \|_2^2 + Tr\left(C_Y + C_{\hat{Y}} - 2\left(C_Y C_{\hat{Y}}\right)^{\frac{1}{2}}\right) \tag{16}$$

where m indicates the mean of the deep features, C means the covariance matrix, and $Tr(.)$ is the trace operation. Intuitively, if the generated images are similar to the real ones, they should have lower FID values.

KID: KID is a metric similar to the FID, the Kernel Inception Distance, to be the squared MMD [15] between Inception representations and has a simple unbiased estimator. Correspondingly, a lower KID means a better performance.

In the testing process, we randomly sample 150 test images and implement translation on them to get the corresponding infrared images for Pix2Pix, BicycleGAN, CycleGAN, GCGAN, UNIT, CUT and DCLGAN. As for one-to-many methods, we generate 10 examples per input and randomly select one as the final result. These generated images and counterpart real ones are used to calculate the metrics mentioned above for each method. We repeat the experiments 5 times and report the average score and standard variances of each metric.

Target quality evaluation

For aerial infrared images, generating as real targets as possible is essential for many tasks, such as object detection and

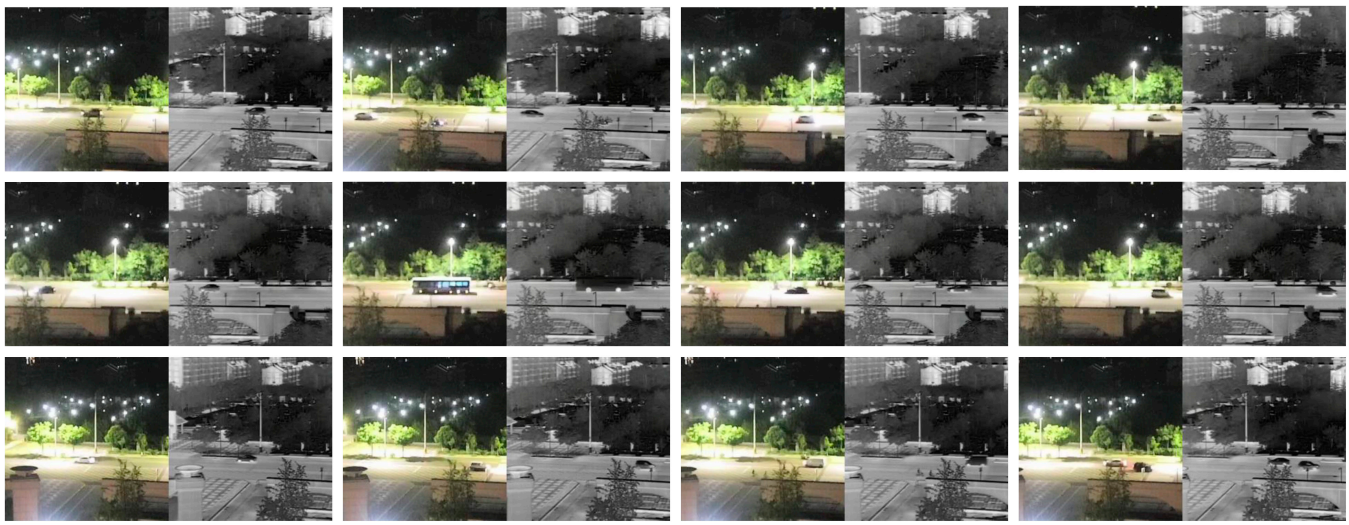

Fig. 4. Some examples of AVIID-2. The scenes of AVIID-2 are the same as AVIID-1.

Fig. 5. Some examples of AVIID-3. (A) The height of the UAV is about 50 m. (B) The height of the UAV is about 100 m. (C) The height of the UAV is about 150 m. The AVIID-3 contains various scenes with complicated background, including roads, bridges, parking lots, and streets, and the number images of each scene are 938, 134, 158, and 50, respectively.

Table 3. Overall appearance evaluation under 50% training ratio on AVIID-1. The best results are highlighted in bold

Methods	Traditional perceptual metrics			CNN-based perceptual metrics		
	MSE ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	KID ↓
Pix2Pix	**13.63 ± 0.13**	**25.46 ± 0.08**	**0.7303 ± 0.0016**	**0.2605 ± 0.0008**	**75.05 ± 2.04**	**0.0787 ± 0.0039**
BicycleGAN	14.12 ± 0.69	25.16 ± 0.42	0.7260 ± 0.0270	0.2776 ± 0.0114	78.64 ± 4.97	0.0829 ± 0.0089
GCGAN	21.89 ± 1.50	21.39 ± 0.62	0.4976 ± 0.0427	0.2592 ± 0.0084	51.30 ± 6.09	0.0387 ± 0.0106
CUT	15.53 ± 0.65	24.37 ± 0.37	0.7073 ± 0.0251	0.1964 ± 0.0133	33.75 ± 3.62	**0.0130 ± 0.0039**
DCLGAN	**14.94 ± 0.67**	**24.71 ± 0.40**	**0.7318 ± 0.0227**	**0.1932 ± 0.0126**	**33.47 ± 4.41**	0.0136 ± 0.0036
CycleGAN	21.13 ± 0.42	21.66 ± 0.17	0.5309 ± 0.0141	0.2745 ± 0.0111	56.42 ± 4.69	0.0435 ± 0.0086
UNIT	17.85 ± 2.20	23.20 ± 1.06	0.6700 ± 0.0804	0.2351 ± 0.0184	42.32 ± 0.92	0.0245 ± 0.0018
DRIT	16.31 ± 0.51	23.92 ± 0.27	0.6824 ± 0.0207	0.2316 ± 0.0057	48.25 ± 2.27	0.0333 ± 0.0057
MUNIT	19.50 ± 1.60	22.39 ± 0.69	0.5958 ± 0.0548	0.2828 ± 0.0132	57.94 ± 2.64	0.0474 ± 0.0054
MSGAN	20.54 ± 0.16	21.98 ± 0.06	0.5592 ± 0.0074	0.3115 ± 0.0035	67.08 ± 1.79	0.0476 ± 0.0045

Table 4. Overall appearance evaluation under 80% training ratio on AVIID-1. The best results are highlighted in bold

Methods	Traditional perceptual metrics			CNN-based perceptual metrics		
	MSE ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	KID ↓
Pix2Pix	**13.56 ± 0.07**	**25.50 ± 0.04**	0.7299 ± 0.0020	**0.2572 ± 0.0014**	**71.75 ± 5.09**	**0.0733 ± 0.0086**
BicycleGAN	13.84 ± 0.32	25.32 ± 0.20	**0.7368 ± 0.0082**	0.2737 ± 0.0075	74.93 ± 2.26	0.0742 ± 0.0025
GCGAN	21.88 ± 1.35	21.39 ± 0.55	0.4975 ± 0.0402	0.2556 ± 0.0097	51.46 ± 3.36	0.0373 ± 0.0049
CUT	14.93 ± 0.71	24.73 ± 0.44	0.7372 ± 0.0249	0.1922 ± 0.0079	**32.68 ± 3.25**	**0.0111 ± 0.0024**
DCLGAN	**14.25 ± 0.91**	**25.11 ± 0.58**	**0.7481 ± 0.0285**	**0.1889 ± 0.0144**	33.54 ± 4.05	0.0121 ± 0.0036
CycleGAN	21.35 ± 0.57	21.57 ± 0.24	0.5281 ± 0.0228	0.2968 ± 0.0128	63.57 ± 2.69	0.0534 ± 0.0047
UNIT	18.22 ± 1.50	22.99 ± 0.72	0.6499 ± 0.0608	0.2408 ± 0.0139	43.49 ± 1.93	0.0266 ± 0.0037
DRIT	16.38 ± 1.08	23.89 ± 0.57	0.6729 ± 0.0410	0.2251 ± 0.0099	46.80 ± 2.90	0.0301 ± 0.0043
MUNIT	18.50 ± 0.76	22.83 ± 0.36	0.6354 ± 0.0314	0.2737 ± 0.0058	55.17 ± 3.05	0.0424 ± 0.0054
MSGAN	19.69 ± 0.47	22.36 ± 0.20	0.5849 ± 0.0154	0.2860 ± 0.0043	58.29 ± 4.01	0.0381 ± 0.0050

Table 5. RmAP under the Faster RCNN object detection model on AVIID-1. The best results are highlighted in bold

Methods	50% Training ratio			80% Training ratio		
	RmAP, IOU:			RmAP, IOU:		
	0.5:0.95	0.5	0.75	0.5:0.95	0.5	0.75
Pix2Pix	**0.260 ± 0.005**	**0.041 ± 0.007**	**0.326 ± 0.019**	**0.229 ± 0.011**	**0.031 ± 0.008**	**0.282 ± 0.033**
BicycleGAN	0.344 ± 0.022	0.063 ± 0.013	0.506 ± 0.062	0.306 ± 0.028	0.080 ± 0.025	0.431 ± 0.051
GCGAN	0.441 ± 0.030	0.085 ± 0.043	0.735 ± 0.064	0.364 ± 0.036	**0.041 ± 0.013**	0.574 ± 0.082
CUT	0.548 ± 0.033	0.206 ± 0.054	0.814 ± 0.024	0.487 ± 0.009	0.143 ± 0.019	0.727 ± 0.021
DCLGAN	0.459 ± 0.043	0.105 ± 0.040	0.747 ± 0.045	**0.351 ± 0.025**	0.044 ± 0.017	**0.547 ± 0.060**
CycleGAN	0.494 ± 0.018	0.163 ± 0.040	0.779 ± 0.010	0.504 ± 0.037	0.163 ± 0.047	0.766 ± 0.041
UNIT	0.426 ± 0.052	0.071 ± 0.035	0.706 ± 0.053	0.403 ± 0.019	0.077 ± 0.030	0.644 ± 0.018
DRIT	**0.398 ± 0.016**	**0.064 ± 0.009**	**0.677 ± 0.025**	0.396 ± 0.021	0.058 ± 0.010	0.642 ± 0.047
MUNIT	0.441 ± 0.031	0.083 ± 0.021	0.711 ± 0.041	0.425 ± 0.013	0.063 ± 0.011	0.705 ± 0.036
MSGAN	0.441 ± 0.006	0.075 ± 0.015	0.711 ± 0.011	0.402 ± 0.008	0.073 ± 0.013	0.631 ± 0.026

Table 6. RmAP under the YOLOv3 object detection model on AVIID-1. The best results are highlighted in bold

Methods	50% Training ratio			80% Training ratio		
	RmAP, IOU:			RmAP, IOU:		
	0.5:0.95	0.5	0.75	0.5:0.95	0.5	0.75
Pix2Pix	**0.177 ± 0.009**	**0.039 ± 0.021**	**0.376 ± 0.022**	**0.171 ± 0.013**	**0.040 ± 0.007**	**0.372 ± 0.042**
BicycleGAN	0.212 ± 0.020	0.040 ± 0.020	0.470 ± 0.057	0.189 ± 0.020	0.050 ± 0.011	0.412 ± 0.045
GCGAN	0.327 ± 0.029	0.076 ± 0.011	0.711 ± 0.075	**0.239 ± 0.036**	**0.025 ± 0.010**	**0.572 ± 0.077**
CUT	0.424 ± 0.039	0.211 ± 0.034	0.753 ± 0.032	0.359 ± 0.012	0.135 ± 0.015	0.710 ± 0.031
DCLGAN	0.325 ± 0.048	0.091 ± 0.048	0.688 ± 0.048	0.244 ± 0.021	0.040 ± 0.012	0.596 ± 0.040
CycleGAN	0.374 ± 0.034	0.148 ± 0.042	0.726 ± 0.035	0.363 ± 0.036	0.136 ± 0.060	0.739 ± 0.039
UNIT	0.286 ± 0.045	**0.059 ± 0.027**	**0.604 ± 0.055**	0.266 ± 0.019	0.048 ± 0.011	0.636 ± 0.040
DRIT	**0.280 ± 0.021**	0.071 ± 0.010	0.630 ± 0.058	0.270 ± 0.023	0.057 ± 0.014	0.653 ± 0.018
MUNIT	0.320 ± 0.021	0.078 ± 0.030	0.668 ± 0.029	0.306 ± 0.018	0.050 ± 0.018	0.719 ± 0.026
MSGAN	0.324 ± 0.018	0.065 ± 0.014	0.698 ± 0.033	0.284 ± 0.012	0.076 ± 0.022	0.661 ± 0.023

tracking. However, existing perceptual metrics mainly consider the overall appearance of the generated images but ignore the evaluation of the targets in the generated images. To address this issue, we propose a new metric named RmAP, which aims to measure the similarity of the targets between the generated images and the real ones and can be obtained by computing the absolute value of the mAP between the real and generated images on the same object detection framework as

$$RmAP = \mid mAP(Y) - mAP\left(\hat{Y}\right) \mid, \tag{17}$$

where mAP is a widely used metric for evaluating the performance of object detection algorithms [72–74].

At testing time, we first use 80% of the real aerial infrared images to train 4 object detection models, including Faster RCNN [75], YOLOv3 [76], YOLOv5 [77], and YOLOx [78]. Then, we randomly select 150 generated images with their ground truth for each method and compute the absolute value of their mAP on every object detection model with 3 kinds of IOU settings. Similar to the overall appearance evaluation, we also repeat the experiments 5 times and report the average score and standard variances of RmAP.

Results and discussion

AVIID-1

Tables 3 and 4 show the means and standard variances of overall appearance evaluation metrics under 50% and 80% training ratio on AVIID-1, respectively. The results show that Pix2Pix performs better than BicycleGAN on both traditional and CNN-based perceptual metrics. DCLGAN and CUT perform similarly, outperforming other unsupervised methods on all appearance evaluation metrics, while CUT performs slightly worse. These results reveal that contrastive learning constraints can achieve a patch-level alignment by maximizing the mutual information between the corresponding input and output

Table 7. RmAP under the YOLOv5 object detection model on AVIID-1. The best results are highlighted in bold

Methods	50% Training ratio			80% Training ratio		
	RmAP, IOU:			RmAP, IOU:		
	0.5:0.95	0.5	0.75	0.5:0.95	0.5	0.75
Pix2Pix	**0.154 ± 0.003**	**0.016 ± 0.003**	**0.277 ± 0.027**	**0.153 ± 0.015**	**0.013 ± 0.008**	**0.267 ± 0.058**
BicycleGAN	0.202 ± 0.025	0.028 ± 0.008	0.369 ± 0.081	0.171 ± 0.017	0.027 ± 0.006	0.296 ± 0.063
GCGAN	0.314 ± 0.024	0.048 ± 0.011	0.678 ± 0.061	0.270 ± 0.038	**0.010 ± 0.004**	0.563 ± 0.107
CUT	0.441 ± 0.029	0.188 ± 0.057	0.736 ± 0.023	0.383 ± 0.013	0.100 ± 0.019	0.661 ± 0.038
DCLGAN	0.347 ± 0.040	0.078 ± 0.034	0.663 ± 0.077	**0.269 ± 0.021**	0.022 ± 0.010	0.523 ± 0.047
CycleGAN	0.377 ± 0.018	0.118 ± 0.013	0.686 ± 0.030	0.376 ± 0.030	0.090 ± 0.032	0.691 ± 0.066
UNIT	0.309 ± 0.058	0.056 ± 0.036	0.588 ± 0.086	0.292 ± 0.022	0.038 ± 0.014	**0.541 ± 0.036**
DRIT	**0.287 ± 0.023**	**0.041 ± 0.009**	**0.584 ± 0.054**	0.291 ± 0.021	0.028 ± 0.011	0.564 ± 0.041
MUNIT	0.316 ± 0.031	**0.041 ± 0.014**	0.614 ± 0.028	0.316 ± 0.017	0.045 ± 0.008	0.626 ± 0.019
MSGAN	0.317 ± 0.019	0.043 ± 0.019	0.652 ± 0.033	0.314 ± 0.016	0.048 ± 0.017	0.613 ± 0.021

Table 8. RmAP under the YOLOx object detection model on AVIID-1. The best results are highlighted in bold

Methods	50% Training ratio			80% Training ratio		
	RmAP, IOU:			RmAP, IOU:		
	0.5:0.95	0.5	0.75	0.5:0.95	0.5	0.75
Pix2Pix	**0.065 ± 0.005**	**0.008 ± 0.007**	**0.211 ± 0.045**	**0.048 ± 0.018**	**0.015 ± 0.011**	**0.190 ± 0.041**
BicycleGAN	0.114 ± 0.021	0.038 ± 0.010	0.314 ± 0.034	0.090 ± 0.014	0.049 ± 0.023	0.249 ± 0.012
GCGAN	**0.152 ± 0.025**	0.049 ± 0.013	**0.367 ± 0.060**	**0.082 ± 0.016**	**0.034 ± 0.018**	0.289 ± 0.036
CUT	0.256 ± 0.027	0.199 ± 0.061	0.428 ± 0.046	0.191 ± 0.015	0.121 ± 0.029	0.364 ± 0.015
DCLGAN	0.198 ± 0.021	0.119 ± 0.038	0.403 ± 0.040	0.116 ± 0.012	0.041 ± 0.008	**0.282 ± 0.023**
CycleGAN	0.203 ± 0.028	0.109 ± 0.036	0.392 ± 0.042	0.214 ± 0.051	0.171 ± 0.089	0.384 ± 0.046
UNIT	0.196 ± 0.043	0.088 ± 0.066	0.390 ± 0.017	0.177 ± 0.009	0.056 ± 0.013	0.385 ± 0.033
DRIT	0.174 ± 0.020	**0.044 ± 0.017**	0.387 ± 0.029	0.170 ± 0.015	0.050 ± 0.009	0.383 ± 0.030
MUNIT	0.190 ± 0.027	**0.044 ± 0.027**	0.397 ± 0.054	0.195 ± 0.020	0.062 ± 0.024	0.421 ± 0.047
MSGAN	0.223 ± 0.022	0.094 ± 0.036	0.437 ± 0.023	0.206 ± 0.013	0.123 ± 0.023	0.394 ± 0.024

patches, thereby improving the overall appearance quality of generated images.

Tables 5 to 8 illustrate the means and standard variances of target quality evaluation metric under 4 objection detection models with 3 IOU settings on AVIID-1. The RmAP results indicate that supervised methods give significantly superior performance compared with unsupervised ones in terms of target quality, which is contrary to the conclusions drawn from the overall appearance evaluation. This suggests that the pixel-level mapping learned from the paired data is beneficial for generating fine-grained targets, while also indicating that the RmAP metric complements overall appearance evaluation metrics and thus more effectively evaluate the performance of algorithms. For unsupervised methods, contrastive learning-based methods do not achieve as excellent performance in target quality as excellent a performance in target quality as in overall appearance

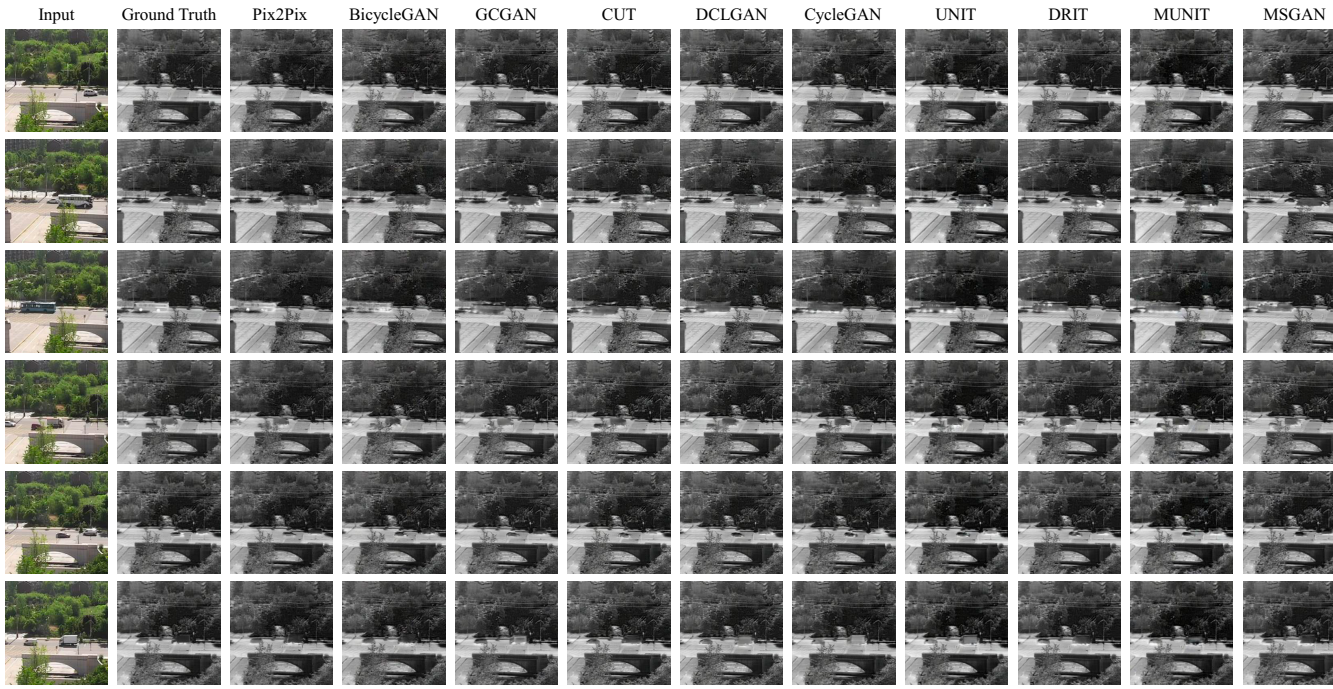

Fig. 6. Some generated images for each method under 50% training ratio on AVIID-1.

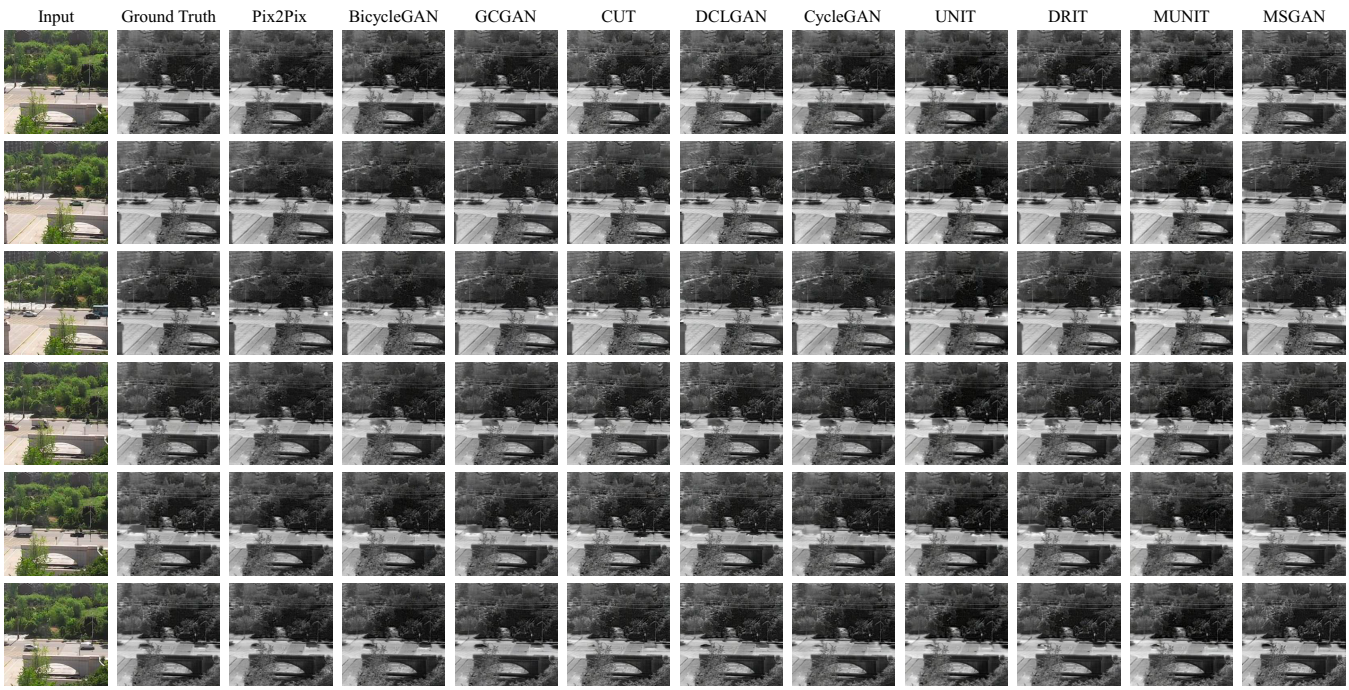

Fig. 7. Some generated images for each method under 80% training ratio on AVIID-1.

Table 9. Overall appearance evaluation under 50% training ratio on AVIID-2. The best results are highlighted in bold

Methods	Traditional perceptual metrics			CNN-based perceptual metrics		
	MSE ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	KID ↓
Pix2Pix	**17.23 ± 0.11**	**23.43 ± 0.05**	**0.6939 ± 0.0026**	**0.2745 ± 0.0015**	**66.72 ± 2.24**	**0.0633 ± 0.0038**
BicycleGAN	22.60 ± 0.31	21.07 ± 0.12	0.6114 ± 0.0087	0.3417 ± 0.0021	92.70 ± 3.99	0.1035 ± 0.0071
GCGAN	22.75 ± 0.40	21.00 ± 0.15	0.5941 ± 0.0060	0.2685 ± 0.0058	59.02 ± 5.99	0.0549 ± 0.0106
CUT	16.16 ± 1.23	24.04 ± 0.69	0.7050 ± 0.0350	0.2246 ± 0.0181	**37.84 ± 4.35**	**0.0170 ± 0.0048**
DCLGAN	**15.75 ± 1.23**	**24.26 ± 0.73**	**0.7230 ± 0.0333**	**0.2234 ± 0.0153**	39.67 ± 2.79	0.0202 ± 0.0042
CycleGAN	26.86 ± 2.21	19.59 ± 0.67	0.5068 ± 0.0261	0.3236 ± 0.0161	68.60 ± 4.80	0.0641 ± 0.0085
UNIT	19.45 ± 1.04	22.38 ± 0.48	0.6566 ± 0.0327	0.2870 ± 0.0120	56.86 ± 4.05	0.0439 ± 0.0052
DRIT	20.82 ± 0.91	21.78 ± 0.37	0.6364 ± 0.0214	0.2575 ± 0.0056	45.70 ± 1.04	0.0310 ± 0.0029
MUNIT	19.32 ± 1.22	22.44 ± 0.56	0.6515 ± 0.0330	0.3017 ± 0.0087	67.02 ± 3.87	0.0622 ± 0.0069
MSGAN	23.02 ± 0.43	20.95 ± 0.17	0.5827 ± 0.0120	0.3197 ± 0.0033	63.88 ± 2.54	0.0453 ± 0.0033

Table 10. Overall appearance evaluation under 80% training ratio on AVIID-2. The best results are highlighted in bold

Methods	Traditional perceptual metrics			CNN-based perceptual metrics		
	PSNR ↑	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	KID ↓
Pix2Pix	**17.45 ± 0.04**	**23.32 ± 0.02**	**0.6915 ± 0.0023**	**0.2763 ± 0.0015**	**69.05 ± 1.77**	**0.0666 ± 0.0037**
BicycleGAN	22.57 ± 0.28	21.08 ± 0.11	0.6120 ± 0.0061	0.3433 ± 0.0025	97.22 ± 3.72	0.1097 ± 0.0077
GCGAN	23.66 ± 0.65	20.66 ± 0.24	0.5798 ± 0.0175	0.2718 ± 0.0108	53.11 ± 4.83	0.0408 ± 0.0067
CUT	19.63 ± 2.82	22.42 ± 1.09	0.6504 ± 0.0171	0.2519 ± 0.0134	48.42 ± 5.72	0.0307 ± 0.0083
DCLGAN	**14.69 ± 1.23**	**24.86 ± 0.74**	**0.7456 ± 0.0287**	**0.2195 ± 0.0122**	**43.49 ± 3.11**	**0.0226 ± 0.0034**
CycleGAN	25.03 ± 1.60	20.19 ± 0.53	0.5505 ± 0.0113	0.3022 ± 0.0026	64.58 ± 5.72	0.0552 ± 0.0102
UNIT	20.21 ± 2.03	22.08 ± 0.85	0.6393 ± 0.0541	0.2864 ± 0.0190	55.41 ± 4.56	0.0416 ± 0.0052
DRIT	20.82 ± 0.42	21.78 ± 0.17	0.6436 ± 0.0202	0.2452 ± 0.0063	44.87 ± 3.70	0.0286 ± 0.0054
MUNIT	20.52 ± 0.20	21.90 ± 0.08	0.6186 ± 0.0092	0.3173 ± 0.0098	80.07 ± 0.95	0.0762 ± 0.0054
MSGAN	23.36 ± 0.23	20.82 ± 0.08	0.5819 ± 0.0054	0.2942 ± 0.0023	52.28 ± 3.10	0.0286 ± 0.0054

Table 11. RmAP under the Fater RCNN object detection model on AVIID-2. The best results are highlighted in bold

Methods	50% Training ratio			80% Training ratio		
	RmAP, IOU:			RmAP, IOU:		
	0.5:0.95	0.5	0.75	0.5:0.95	0.5	0.75
Pix2Pix	**0.266 ± 0.009**	**0.025 ± 0.008**	**0.370 ± 0.043**	**0.279 ± 0.023**	**0.024 ± 0.005**	**0.418 ± 0.069**
BicycleGAN	0.438 ± 0.011	0.076 ± 0.028	0.777 ± 0.039	0.424 ± 0.020	0.098 ± 0.020	0.732 ± 0.032
GCGAN	**0.385 ± 0.007**	**0.037 ± 0.010**	0.685 ± 0.021	0.365 ± 0.041	0.049 ± 0.020	0.647 ± 0.071
CUT	0.461 ± 0.080	0.106 ± 0.044	0.702 ± 0.119	0.499 ± 0.040	0.151 ± 0.041	0.761 ± 0.027
DCLGAN	0.388 ± 0.033	0.062 ± 0.022	**0.669 ± 0.068**	**0.350 ± 0.011**	0.048 ± 0.012	**0.597 ± 0.031**
CycleGAN	0.476 ± 0.044	0.103 ± 0.038	0.795 ± 0.056	0.392 ± 0.031	0.075 ± 0.018	0.669 ± 0.050
UNIT	0.419 ± 0.037	0.074 ± 0.028	0.708 ± 0.045	0.392 ± 0.056	0.070 ± 0.024	0.668 ± 0.085
DRIT	0.412 ± 0.007	0.039 ± 0.012	0.724 ± 0.017	0.385 ± 0.051	0.045 ± 0.024	0.666 ± 0.105
MUNIT	0.464 ± 0.019	0.081 ± 0.026	0.767 ± 0.049	0.444 ± 0.045	0.088 ± 0.036	0.758 ± 0.042
MSGAN	0.424 ± 0.032	0.049 ± 0.020	0.733 ± 0.057	0.392 ± 0.018	**0.034 ± 0.012**	0.708 ± 0.030

Table 12. RmAP under the YOLOv3 object detection model on AVIID-2. The best results are highlighted in bold

Methods	50% Training ratio			80% Training ratio		
	RmAP, IOU:			RmAP, IOU:		
	0.5:0.95	0.5	0.75	0.5:0.95	0.5	0.75
Pix2Pix	**0.148 ± 0.013**	**0.027 ± 0.016**	**0.333 ± 0.073**	**0.151 ± 0.012**	**0.020 ± 0.001**	**0.382 ± 0.035**
BicycleGAN	0.284 ± 0.013	0.046 ± 0.014	0.657 ± 0.045	0.259 ± 0.009	0.073 ± 0.016	0.557 ± 0.032
GCGAN	**0.246 ± 0.015**	**0.038 ± 0.008**	**0.584 ± 0.026**	0.245 ± 0.043	0.041 ± 0.017	0.564 ± 0.102
CUT	0.314 ± 0.080	0.097 ± 0.054	0.603 ± 0.109	0.368 ± 0.049	0.151 ± 0.050	0.677 ± 0.036
DCLGAN	0.260 ± 0.029	0.053 ± 0.017	0.595 ± 0.049	**0.236 ± 0.008**	**0.032 ± 0.008**	**0.530 ± 0.017**
CycleGAN	0.321 ± 0.060	0.082 ± 0.033	0.676 ± 0.076	0.264 ± 0.028	0.059 ± 0.012	0.594 ± 0.052
UNIT	0.281 ± 0.030	0.056 ± 0.016	0.624 ± 0.049	0.261 ± 0.047	0.062 ± 0.024	0.575 ± 0.049
DRIT	0.272 ± 0.013	0.040 ± 0.012	0.633 ± 0.013	0.269 ± 0.034	0.041 ± 0.013	0.596 ± 0.041
MUNIT	0.327 ± 0.027	0.089 ± 0.041	0.682 ± 0.020	0.308 ± 0.043	0.103 ± 0.053	0.652 ± 0.032
MSGAN	0.289 ± 0.028	0.041 ± 0.015	0.631 ± 0.036	0.257 ± 0.015	0.034 ± 0.018	0.588 ± 0.018

quality and even perform worse than other approaches. For instance, DRIT has achieved much lower RmAP values than DCLGAN on Faster RCNN and YOLOv5 object detection algorithms under 50% training ratio with 3 kinds of IOU settings. Similarly, GCGAN also gives better results on the YOLOv3 model under the 80% training ratio for all IOU settings. The possible reason for this phenomenon may be that the patch-level alignment can be seen as the coarse-grained mapping between the input and output images compared with the pixel-level mapping, which could lead to blurriness and distortion of targets in the generated images. This phenomenon becomes more serious in aerial images, mainly because there often exist many small and geometric discrepancy targets (such as cars and buses in our dataset).

Figures 6 and 7 display some generated images for each method under 50% and 80% training ratio on AVIID-1, respectively. By comparing these generated examples, we can find that the vehicles generated by DCLGAN and CUT have geometric distortion and blurred edges compared with Pix2Pix, especially CUT, which further confirms our assumption.

AVIID-2

Tables 9 and 10 show the means and standard variances of overall appearance evaluation metrics under 50% and 80% training ratio on AVIID-2, respectively. Through the results, we can get conclusions similar to those of AVIID-1 that Pix2Pix performs superiorly to BicycleGAN, and DCLGAN achieves the best performance followed by CUT in the unsupervised methods. It is worth noting that BicycleGAN gets a much lower performance than Pix2Pix, which is different from AVIID-1. The reason may be that the visible images in AVIID-2 are seriously affected by weak light and noise, resulting in large discrepancies between them and their corresponding infrared images, especially the backgrounds. As a result, the generator

Table 13. RmAP under the YOLOv5 object detection model on AVIID-2. The best results are highlighted in bold

Methods	50% Training ratio			80% Training ratio		
	RmAP, IOU:			RmAP, IOU:		
	0.5:0.95	0.5	0.75	0.5:0.95	0.5	0.75
Pix2Pix	**0.164 ± 0.013**	**0.010 ± 0.005**	**0.326 ± 0.038**	**0.150 ± 0.016**	**0.007 ± 0.005**	**0.334 ± 0.052**
BicycleGAN	0.296 ± 0.014	0.027 ± 0.015	0.688 ± 0.032	0.268 ± 0.016	0.028 ± 0.006	0.624 ± 0.040
GCGAN	**0.275 ± 0.006**	0.020 ± 0.011	0.650 ± 0.014	0.270 ± 0.039	0.022 ± 0.006	0.607 ± 0.077
CUT	0.349 ± 0.099	0.083 ± 0.047	0.685 ± 0.126	0.399 ± 0.060	0.090 ± 0.032	0.748 ± 0.051
DCLGAN	0.284 ± 0.021	0.025 ± 0.013	**0.635 ± 0.033**	**0.247 ± 0.009**	**0.018 ± 0.005**	**0.550 ± 0.031**
CycleGAN	0.343 ± 0.058	0.041 ± 0.026	0.721 ± 0.074	0.278 ± 0.030	0.032 ± 0.005	0.607 ± 0.058
UNIT	0.296 ± 0.042	0.034 ± 0.017	0.641 ± 0.046	0.284 ± 0.055	0.031 ± 0.017	0.593 ± 0.094
DRIT	0.295 ± 0.012	**0.017 ± 0.010**	0.690 ± 0.010	0.295 ± 0.045	0.025 ± 0.007	0.625 ± 0.072
MUNIT	0.357 ± 0.033	0.052 ± 0.027	0.726 ± 0.040	0.345 ± 0.045	0.063 ± 0.038	0.708 ± 0.027
MSGAN	0.319 ± 0.023	0.027 ± 0.025	0.703 ± 0.019	0.295 ± 0.022	0.030 ± 0.015	0.640 ± 0.036

Table 14. RmAP under the YOLOx object detection model on AVIID-2. The best results are highlighted in bold

Methods	50% Training ratio			80% Training ratio		
	RmAP, IOU:			RmAP, IOU:		
	0.5:0.95	0.5	0.75	0.5:0.95	0.5	0.75
Pix2Pix	**0.123 ± 0.012**	**0.012 ± 0.006**	**0.341 ± 0.063**	**0.126 ± 0.012**	**0.014 ± 0.009**	**0.364 ± 0.034**
BicycleGAN	0.236 ± 0.011	0.055 ± 0.010	0.450 ± 0.035	0.219 ± 0.016	0.062 ± 0.021	0.401 ± 0.035
GCGAN	0.206 ± 0.018	**0.030 ± 0.014**	0.449 ± 0.029	0.186 ± 0.019	**0.025 ± 0.013**	0.394 ± 0.027
CUT	0.244 ± 0.075	0.125 ± 0.073	0.442 ± 0.052	0.288 ± 0.037	0.174 ± 0.068	0.483 ± 0.023
DCLGAN	0.194 ± 0.019	0.050 ± 0.023	0.432 ± 0.057	0.167 ± 0.013	0.033 ± 0.011	0.356 ± 0.037
CycleGAN	**0.185 ± 0.027**	0.049 ± 0.018	**0.391 ± 0.044**	**0.161 ± 0.020**	0.029 ± 0.004	**0.346 ± 0.045**
UNIT	0.249 ± 0.033	0.064 ± 0.028	0.475 ± 0.031	0.229 ± 0.046	0.065 ± 0.032	0.468 ± 0.020
DRIT	0.239 ± 0.011	0.068 ± 0.012	0.482 ± 0.049	0.236 ± 0.030	0.060 ± 0.025	0.444 ± 0.023
MUNIT	0.288 ± 0.031	0.115 ± 0.055	0.487 ± 0.045	0.278 ± 0.044	0.101 ± 0.061	0.461 ± 0.040
MSGAN	0.256 ± 0.022	0.106 ± 0.035	0.448 ± 0.025	0.233 ± 0.012	0.057 ± 0.034	0.452 ± 0.028

may pay too much attention to the latent vector encoded from infrared images in the translating process, which leads to the distortion of details in the generated images. In addition, the values of overall appearance evaluation metrics obtained by each method are significantly lower than those on AVIID-1, indicating that AVIID-2 is more challenging.

Tables 11 to 14 illustrate the means and standard variances of target quality evaluation metric under 4 objection detection models with 3 IOU settings on AVIID-2. From the RmAP results, we can find that Pix2Pix has achieved a better performance than all other methods by a large margin in terms of target quality, which is similar to AVIID-1. As for the unsupervised approaches, DCLGAN achieves superior results on AVIID-2. For example, it gives the best performance on the Faster RCNN, YOLOv3, and YOLOv5 object detection models under 80% training ratio and a lower RmAP on the Faster RCNN and YOLOv5 under 50% training ratio when the IOU is set to 0.75.

Figures 8 and 9 display some generated images for each method under 50% and 80% training ratio on AVIID-2, respectively. From these figures, we can see that some generated images have blurred backgrounds and geometric distortion of the targets, which is more severe in supervised methods. This phenomenon may indicate that pixel-level mapping becomes too strict when the visible images are severely disturbed by weak light and noise, thus degrading the quality of the generated

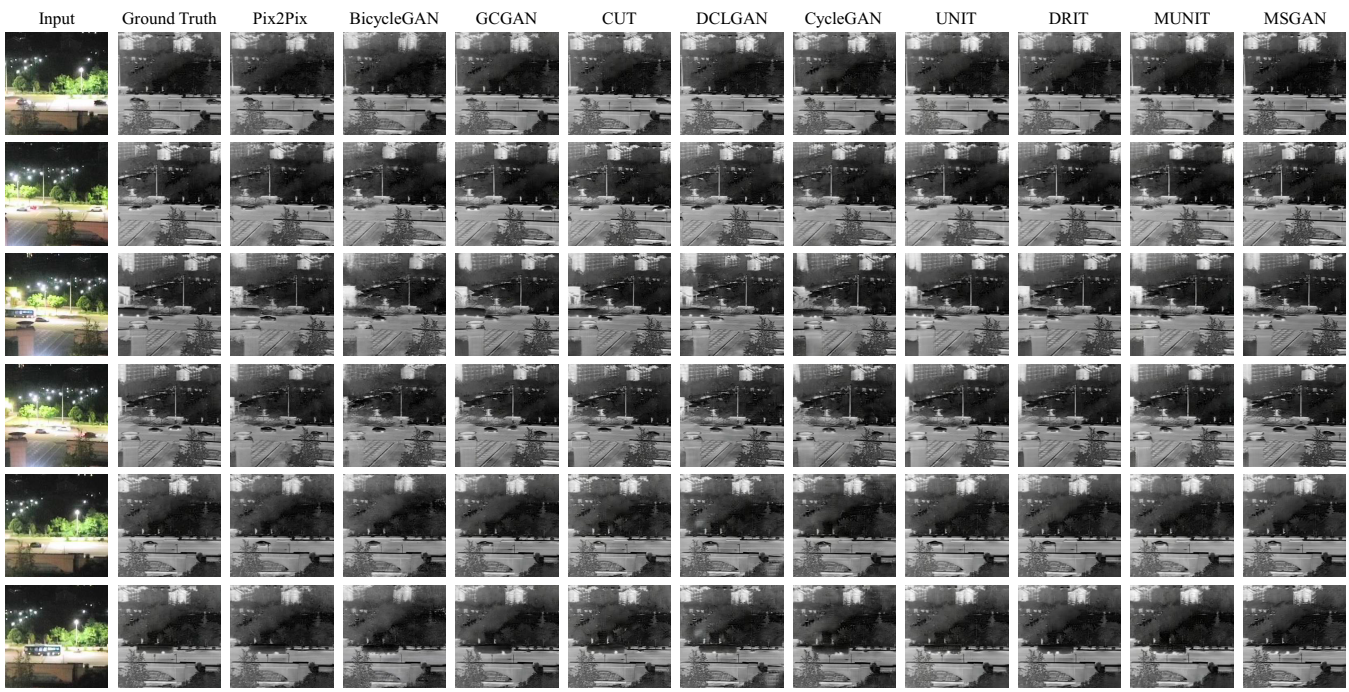

Fig. 8. Some generated images for each method under 50% training ratio on AVIID-2.

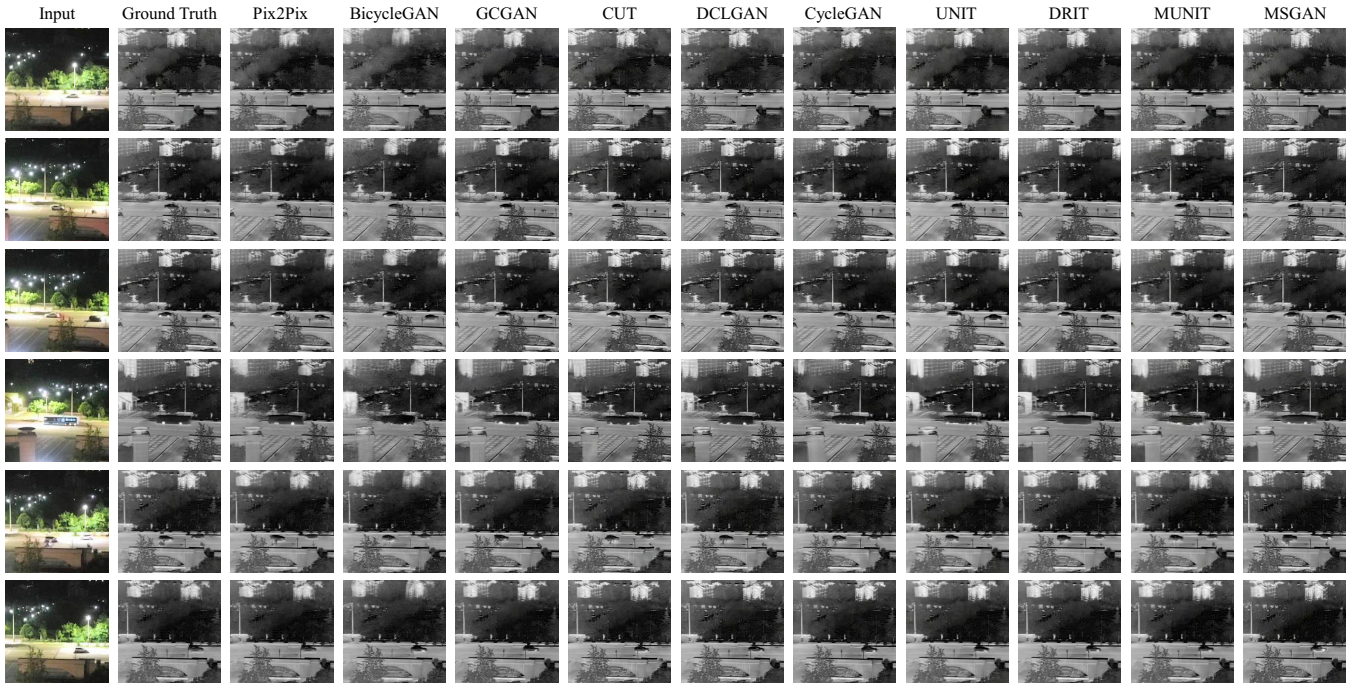

Fig. 9. Some generated images for each method under 80% training ratio on AVIID-2.

images. In this case, patch-level alignment is less strict than pixel-level mapping; thus, contrastive learning-based methods can better preserve the clarity of backgrounds and the geometry of targets in the generated images.

AVIID-3

Tables 15 and 16 show the means and standard variances of overall appearance evaluation metrics under 50% and 80% training ratio on AVIID-3, respectively. From the results, we can find that Pix2Pix still performs better than BicycleGAN, which is similar to AVIID-1 and AVIID-2. However, in the unsupervised methods, GCGAN gives a significantly improved performance of DCLGAN, which performs best on AVIID-1 and AVIID-2 under all overall appearance quality metrics. This phenomenon may result in the conclusion that simple geometry-consistency constraint can effectively maintain the geometric shape of the targets (particularly tiny and dense cars in AVIID) during the translating process, which is beneficial to reduce the blur and detail distortions of the generated images in the case of various scenarios with more complicated backgrounds, while contrastive learning and cycle-consistency constraint are too strict.

Tables 17 to 20 illustrate the means and standard variances of target quality evaluation metric under 4 objection detection models with 3 IOU settings on AVIID-3. From the RmAP results, we can see that GCGAN achieves an overwhelming superiority in target quality compared with other unsupervised methods, which further reflects the effectiveness of

Table 15. Overall appearance evaluation under 50% training ratio on AVIID-3. The best results are highlighted in bold

Methods	Traditional perceptual metrics			CNN-based perceptual metrics		
	MSE ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	KID ↓
Pix2Pix	**22.29 ± 0.50**	**21.38 ± 0.19**	**0.5706 ± 0.0073**	**0.3699 ± 0.0030**	**96.05 ± 3.13**	0.0201 ± 0.0016
BicycleGAN	26.96 ± 0.74	19.80 ± 0.26	0.4974 ± 0.0141	0.3980 ± 0.0061	97.11 ± 3.50	**0.0183 ± 0.0036**
GCGAN	**24.95 ± 0.46**	**20.47 ± 0.13**	**0.5495 ± 0.0034**	**0.3305 ± 0.0034**	**77.01 ± 1.36**	**0.0065 ± 0.0014**
CUT	28.90 ± 0.81	19.18 ± 0.20	0.4856 ± 0.0075	0.3825 ± 0.0027	87.41 ± 1.63	0.0088 ± 0.0008
DCLGAN	26.57 ± 1.06	19.86 ± 0.33	0.5087 ± 0.0134	0.3820 ± 0.0111	91.62 ± 4.61	0.0155 ± 0.0052
CycleGAN	28.63 ± 1.75	19.19 ± 0.49	0.4939 ± 0.0187	0.3617 ± 0.0070	85.92 ± 3.83	0.0068 ± 0.0017
UNIT	29.12 ± 0.93	19.06 ± 00.26	0.4919 ± 0.0171	0.3690 ± 0.0055	89.90 ± 3.38	0.0088 ± 0.0013
DRIT	28.03 ± 1.57	19.41 ± 0.51	0.5078 ± 0.0215	0.3860 ± 0.0104	114.19 ± 2.74	0.0295 ± 0.0043
MUNIT	29.04 ± 0.58	19.08 ± 0.16	0.4806 ± 0.0159	0.3843 ± 0.0037	90.28 ± 1.50	0.0070 ± 0.0006
MSGAN	37.55 ± 1.03	16.96 ± 0.19	0.4595 ± 0.0157	0.4209 ± 0.0033	125.60 ± 3.48	0.0353 ± 0.0024

Table 16. Overall appearance evaluation under 80% training ratio on AVIID-3. The best results are highlighted in bold

Methods	Traditional perceptual metrics			CNN-based perceptual metrics		
	MSE ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	KID ↓
Pix2Pix	**20.66 ± 0.49**	**22.00 ± 0.18**	**0.5861 ± 0.0048**	**0.3628 ± 0.0029**	**96.50 ± 2.07**	0.0183 ± 0.0022
BicycleGAN	25.37 ± 0.89	20.35 ± 0.32	0.5190 ± 0.0061	0.3826 ± 0.0045	100.81 ± 4.22	**0.0180 ± 0.0029**
GCGAN	**22.86 ± 0.71**	**21.39 ± 0.24**	**0.5718 ± 0.0092**	**0.3146 ± 0.0052**	**75.31 ± 1.04**	**0.0076 ± 0.0005**
CUT	27.02 ± 0.59	19.70 ± 0.18	0.5011 ± 0.0111	0.3723 ± 0.0100	88.32 ± 4.73	0.0094 ± 0.0026
DCLGAN	25.37 ± 0.93	20.25 ± 0.28	0.5161 ± 0.0167	0.3705 ± 0.0059	88.83 ± 2.54	0.0128 ± 0.0025
CycleGAN	27.95 ± 2.07	19.39 ± 0.63	0.5039 ± 0.0213	0.3610 ± 0.0111	86.51 ± 3.66	0.0081 ± 0.0019
UNIT	28.60 ± 0.50	19.22 ± 0.15	0.5129 ± 0.0175	0.3615 ± 0.0051	90.89 ± 1.56	0.0091 ± 0.0009
DRIT	24.94 ± 0.75	20.41 ± 0.27	0.5377 ± 0.0226	0.3683 ± 0.0058	112.83 ± 3.45	0.0293 ± 0.0041
MUNIT	29.47 ± 0.75	18.95 ± 0.25	0.4796 ± 0.0220	0.3880 ± 0.0010	93.36 ± 2.46	0.0085 ± 0.0017
MSGAN	32.66 ± 0.16	18.17 ± 0.31	0.4981 ± 0.0194	0.3965 ± 0.0023	114.28 ± 3.16	0.0288 ± 0.0029

geometry-consistency constraint on generating high-quality targets.

Figures 10 and 11 display some generated images for each method under 50% and 80% training ratio on AVIID-3, respectively. From the figures, we can find that GCGAN can maintain the geometric shape of targets to reduce distortions and blur, especially in the case of dense cars, which further proves our conclusion.

Conclusion

From the above experimental results and discussion, we can sum up some meaningful conclusions as follows.

• The pixel-level mapping learned from the paired data is beneficial for generating fine-grained targets. Therefore, supervised methods give significantly superior performance in target quality evaluation compared with unsupervised approaches.

• The contrastive learning constraint can be seen as a patch-level mapping by maximizing mutual information between the corresponding input and output patches. This patch-level alignment can enhance the correspondence of the input and output patches, which helps to improve the quality of generating images, especially in weak light and noisy conditions.

• The geometry-consistency constraint is a simple and effective way to maintain the geometric shape of the targets (particularly tiny and dense targets) during the translating process, which can meaningfully reduce the blur and detail distortions of the generated images in the case of various scenarios with complicated backgrounds.

In addition, several problems of existing methods can be summarized from the experiment results and discussion, which can be seen as follows.

• Current approaches only consider migrating the global styles or attributes onto the entire images but ignore the considerable discrepancy between targets and backgrounds in infrared attributes, resulting in unrealistic targets in the generated images.

Table 17. RmAP under the Faster RCNN object detection model on AVIID-3. The best results are highlighted in bold

Methods	50% Training ratio			80% Training ratio		
	RmAP, IOU:			RmAP, IOU:		
	0.5:0.95	0.5	0.75	0.5:0.95	0.5	0.75
Pix2Pix	**0.272 ± 0.010**	**0.134 ± 0.016**	**0.344 ± 0.021**	**0.268 ± 0.012**	**0.098 ± 0.007**	**0.329 ± 0.035**
BicycleGAN	0.396 ± 0.015	0.174 ± 0.012	0.607 ± 0.022	0.396 ± 0.020	0.161 ± 0.011	0.592 ± 0.030
GCGAN	**0.328 ± 0.019**	**0.114 ± 0.016**	**0.495 ± 0.042**	**0.320 ± 0.014**	**0.108 ± 0.011**	**0.476 ± 0.032**
CUT	0.427 ± 0.025	0.282 ± 0.034	0.570 ± 0.038	0.408 ± 0.027	0.238 ± 0.019	0.574 ± 0.048
DCLGAN	0.430 ± 0.045	0.280 ± 0.040	0.586 ± 0.065	0.396 ± 0.021	0.237 ± 0.032	0.557 ± 0.031
CycleGAN	0.410 ± 0.035	0.195 ± 0.033	0.604 ± 0.051	0.356 ± 0.033	0.169 ± 0.031	0.526 ± 0.047
UNIT	0.430 ± 0.052	0.244 ± 0.038	0.608 ± 0.067	0.374 ± 0.050	0.178 ± 0.040	0.547 ± 0.066
DRIT	0.424 ± 0.054	0.216 ± 0.077	0.617 ± 0.070	0.340 ± 0.042	0.155 ± 0.028	0.512 ± 0.085
MUNIT	0.476 ± 0.018	0.282 ± 0.012	0.658 ± 0.035	0.469 ± 0.054	0.333 ± 0.101	0.672 ± 0.078
MSGAN	0.441 ± 0.016	0.267 ± 0.044	0.627 ± 0.039	0.367 ± 0.037	0.182 ± 0.017	0.541 ± 0.072

Table 18. RmAP under the YOLOv3 object detection model on AVIID-3. The best results are highlighted in bold

Methods	50% Training ratio			80% Training ratio		
	RmAP, IOU:			RmAP, IOU:		
	0.5:0.95	0.5	0.75	0.5:0.95	0.5	0.75
Pix2Pix	**0.249 ± 0.019**	**0.123 ± 0.025**	**0.462 ± 0.042**	**0.213 ± 0.011**	**0.066 ± 0.010**	**0.420 ± 0.028**
BicycleGAN	0.396 ± 0.025	0.214 ± 0.036	0.715 ± 0.032	0.357 ± 0.012	0.157 ± 0.032	0.658 ± 0.033
GCGAN	**0.315 ± 0.009**	**0.141 ± 0.015**	**0.600 ± 0.033**	**0.300 ± 0.004**	**0.116 ± 0.026**	**0.580 ± 0.007**
CUT	0.427 ± 0.016	0.376 ± 0.060	0.683 ± 0.032	0.407 ± 0.026	0.309 ± 0.036	0.663 ± 0.056
DCLGAN	0.430 ± 0.024	0.362 ± 0.038	0.683 ± 0.064	0.403 ± 0.030	0.327 ± 0.038	0.641 ± 0.042
CycleGAN	0.370 ± 0.030	0.190 ± 0.021	0.686 ± 0.068	0.351 ± 0.030	0.229 ± 0.043	0.599 ± 0.044
UNIT	0.415 ± 0.050	0.292 ± 0.070	0.692 ± 0.080	0.368 ± 0.064	0.233 ± 0.075	0.632 ± 0.079
DRIT	0.405 ± 0.064	0.253 ± 0.092	0.697 ± 0.068	0.329 ± 0.053	0.203 ± 0.057	0.568 ± 0.089
MUNIT	0.461 ± 0.032	0.343 ± 0.034	0.755 ± 0.064	0.456 ± 0.061	0.379 ± 0.104	0.716 ± 0.046
MSGAN	0.424 ± 0.034	0.310 ± 0.080	0.714 ± 0.044	0.351 ± 0.040	0.209 ± 0.035	0.623 ± 0.080

• Existing methods can only transfer styles or attributes between aerial visible and infrared images without taking into account the different properties of each modality. Consequently, the authenticity of generated images is poor.

• For aerial images with multi-scale dense targets, complex backgrounds, and diverse scenes, current methods struggle to capture the spatial differences between images, resulting in distortion and blurring of generated targets and backgrounds, significantly reducing the quality of generated images.

The above conclusions can provide meaningful guidance for investigating more efficient methods on more challenging datasets to facilitate the process of aerial visible-to-infrared image translation.

Conclusion

In this paper, we find that there lacks a benchmark dataset for aerial visible-to-infrared image translation experiments, thus severely limiting the development of this field. To solve the problem, we construct the first specific dataset, AVIID, consisting of paired aerial visible and infrared images for aerial visible-to-infrared image translation. The purpose of AVIID is to provide researchers with an available data resource to evaluate and advance state-of-the-art algorithms for aerial visible-to-infrared image translation. Based on AVIID, we also propose a complete evaluation system to evaluate the generated infrared images from the overall appearance and target quality. In particular, a new metric named RmAP is proposed to evaluate the quality of targets in the generated images. Then, a comprehensive survey on image-to-image translation methods that could be applied to aerial visible-to-infrared image translation is given. After that, several typical image-to-image translation approaches are evaluated using our proposed evaluation system on AVIID. These results can be seen as a baseline for future work. Finally, some meaningful conclusions and problems of existing methods are

Table 19. RmAP under the YOLOv5 object detection model on AVIID-3. The best results are highlighted in bold

Methods	50% Training ratio			80% Training ratio		
	RmAP, IOU:			RmAP, IOU:		
	0.5:0.95	0.5	0.75	0.5:0.95	0.5	0.75
Pix2Pix	**0.237 ± 0.012**	**0.090 ± 0.014**	**0.384 ± 0.025**	**0.205 ± 0.010**	**0.043 ± 0.006**	**0.317 ± 0.021**
BicycleGAN	0.393 ± 0.018	0.153 ± 0.031	0.677 ± 0.021	0.377 ± 0.019	0.123 ± 0.019	0.641 ± 0.025
GCGAN	**0.341 ± 0.004**	**0.106 ± 0.011**	**0.605 ± 0.032**	**0.331 ± 0.003**	**0.087 ± 0.018**	**0.603 ± 0.009**
CUT	0.443 ± 0.020	0.295 ± 0.039	0.690 ± 0.031	0.422 ± 0.029	0.238 ± 0.027	0.696 ± 0.039
DCLGAN	0.438 ± 0.014	0.267 ± 0.041	0.689 ± 0.023	0.418 ± 0.037	0.254 ± 0.045	0.670 ± 0.052
CycleGAN	0.407 ± 0.030	0.174 ± 0.024	0.681 ± 0.051	0.392 ± 0.036	0.203 ± 0.046	0.645 ± 0.049
UNIT	0.465 ± 0.047	0.262 ± 0.064	0.731 ± 0.049	0.403 ± 0.071	0.200 ± 0.090	0.670 ± 0.068
DRIT	0.441 ± 0.068	0.221 ± 0.074	0.712 ± 0.084	0.356 ± 0.061	0.151 ± 0.053	0.617 ± 0.098
MUNIT	0.486 ± 0.040	0.296 ± 0.038	0.746 ± 0.049	0.493 ± 0.066	0.345 ± 0.108	0.760 ± 0.057
MSGAN	0.443 ± 0.041	0.234 ± 0.076	0.724 ± 0.046	0.361 ± 0.041	0.147 ± 0.018	0.633 ± 0.084

Table 20. RmAP under the YOLOx object detection model on AVIID-3. The best results are highlighted in bold

Methods	50% Training ratio			80% Training ratio		
	RmAP, IOU:			RmAP, IOU:		
	0.5:0.95	0.5	0.75	0.5:0.95	0.5	0.75
Pix2Pix	**0.134 ± 0.008**	**0.131 ± 0.023**	**0.221 ± 0.025**	**0.104 ± 0.005**	**0.084 ± 0.006**	**0.194 ± 0.020**
BicycleGAN	0.299 ± 0.007	0.355 ± 0.023	0.359 ± 0.028	0.279 ± 0.006	0.288 ± 0.022	0.377 ± 0.024
GCGAN	**0.228 ± 0.012**	**0.210 ± 0.013**	**0.329 ± 0.035**	**0.224 ± 0.008**	**0.196 ± 0.007**	0.326 ± 0.035
CUT	0.281 ± 0.008	0.377 ± 0.035	0.363 ± 0.035	0.281 ± 0.027	0.351 ± 0.045	0.350 ± 0.027
DCLGAN	0.278 ± 0.025	0.365 ± 0.023	0.358 ± 0.052	0.272 ± 0.033	0.359 ± 0.062	0.347 ± 0.035
CycleGAN	0.284 ± 0.022	0.336 ± 0.043	0.380 ± 0.024	0.269 ± 0.030	0.339 ± 0.069	0.330 ± 0.042
UNIT	0.310 ± 0.029	0.431 ± 0.053	0.355 ± 0.052	0.266 ± 0.054	0.332 ± 0.118	0.329 ± 0.044
DRIT	0.283 ± 0.043	0.358 ± 0.106	0.356 ± 0.037	0.231 ± 0.045	0.265 ± 0.085	**0.298 ± 0.031**
MUNIT	0.330 ± 0.034	0.453 ± 0.078	0.384 ± 0.032	0.324 ± 0.050	0.475 ± 0.121	0.356 ± 0.026
MSGAN	0.285 ± 0.032	0.366 ± 0.073	0.370 ± 0.032	0.219 ± 0.032	0.241 ± 0.056	0.306 ± 0.031

summarized to advance state-of-the-art algorithms for aerial visible-to-infrared image translation. In addition, several future research directions of this field are analyzed and summarized as follows.

• Current image-to-image translation methods are not concerned with the imaging mechanism between the visible and infrared image. How to construct reasonable imaging mechanism constraints to improve the realism of generated infrared images is a future research direction.

• The AVIID dataset proposed in this article are aerial remote sensing images taken by infrared camera equipped on the UAV. The visible-to-infrared image translation in satellite platform also deserves to be researched in the future.

• Existing image-to-image translation methods are mainly based on deep CNNs. However, due to the limitation computational resource, the parameters of the model cannot be infinitely large, so the size of the generated image is limited. Therefore, finding an effective way to transfer these approaches to large-scale areas is necessary.

• The quality of the generated images through image-to-image translation methods is highly correlated with the similarity between training and test data. Therefore, improving the transferability and generalizability of these methods is one of the research directions in the future.

• The radiation value of thermal images has a great relationship with the atmospheric conditions, and when the infrared

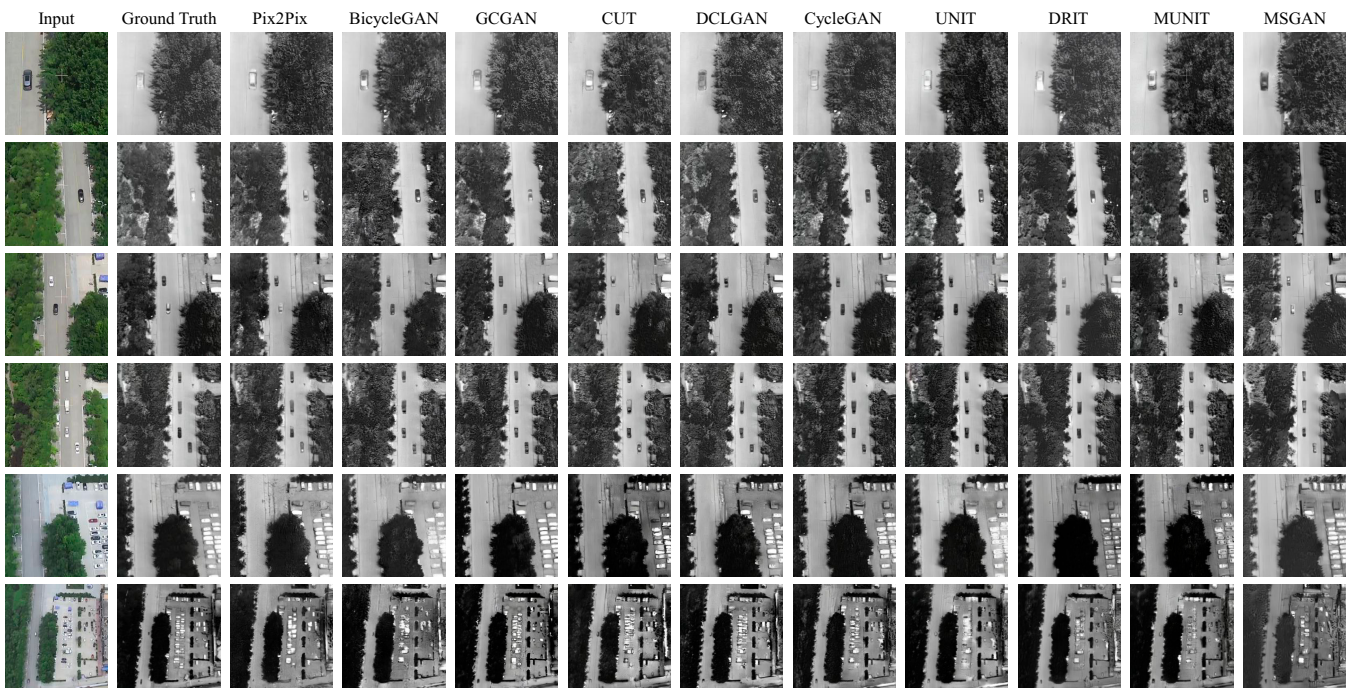

Fig. 10. Some generated images for each method under 50% training ratio on AVIID-3.

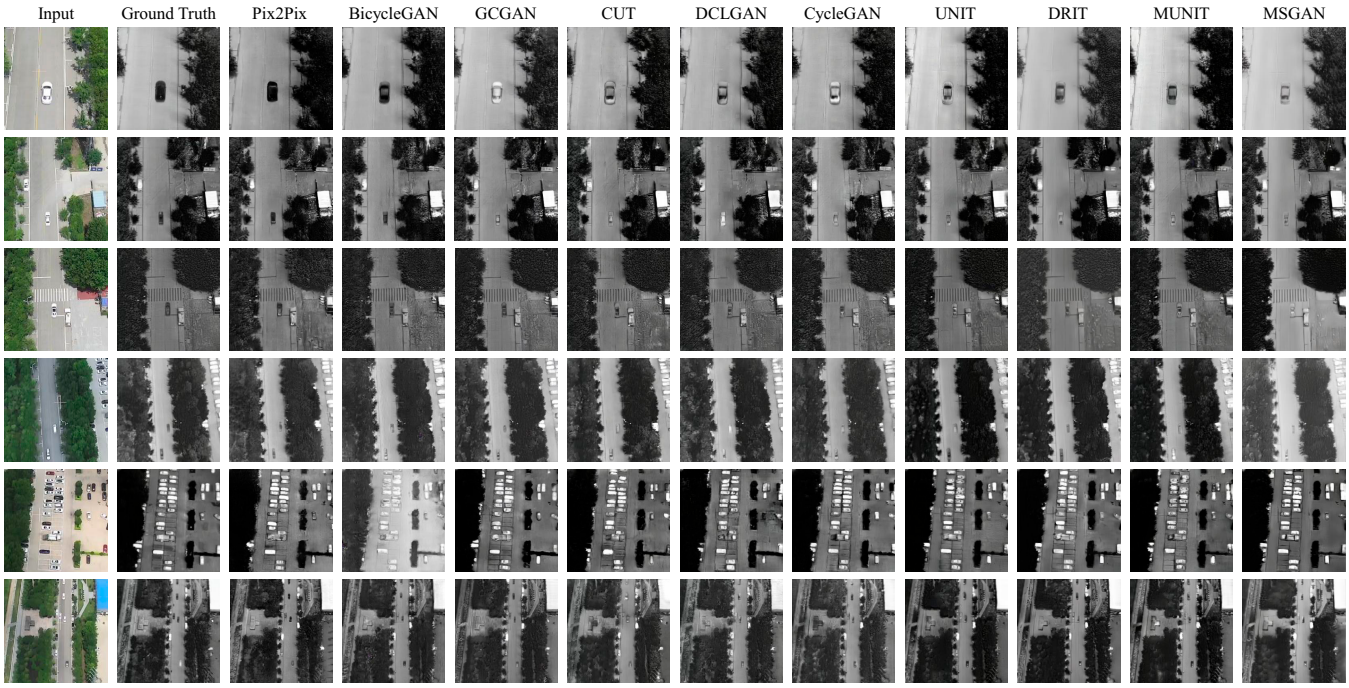

Fig. 11. Some generated images for each method under 80% training ratio on AVIID-3.

images are taken at a very high height above the ground, solving the atmospheric compensation is a worthwhile problem.

Moreover, AVIID and PyTorch codes of these methods can be freely downloaded to advance the process of aerial visible-to-infrared image translation.

Acknowledgments

Funding: This work was supported in part by the National Natural Science Foundation of China (62271409 and 62171381).
Author contributions: Z.H. designed the study and the experiment. Z.Z. performed the experiments and data analysis. Z.H., S.Z., and G.Z. contributed to the writing of the manuscript. S.M. helped in revising the manuscript.
Competing interests: The authors declare that they have no competing interests.

Data Availability

The dataset related to this article is publicly available online and can be downloaded from https://github.com/silver-hzh/Averial-visible-to-infrared-image-translation.

References

1. Jung YS, Song TL. Aerial-target detection using the recursive temporal profile and spatiotemporal gradient pattern in infrared image sequences. *Opt Eng*. 2012;51:Article 066401.
2. Chrétien L-P, Théau J, Ménard P. Visible and thermal infrared remote sensing for the detection of white-tailed deer using an unmanned aerial system. *Wildl Soc Bull*. 2016;40(1):181.
3. Xu W, Zhong S, Yan L, Wu F, Zhang W. Moving object detection in aerial infrared images with registration accuracy prediction and feature points selection. *Infrared Phys Technol*. 2018;92:318–326.
4. Hu Y, Xiao M, Zhang K, Wang X. Aerial infrared target tracking in complex background based on combined tracking and detecting. *Math Probl Eng*. 2019;2019(28):1–17.
5. Lega M, Kosmatka J, Ferrara C, Russo F, Napoli RMA, Persechino G. Using advanced aerial platforms and infrared thermography to track environmental contamination. *Environ Forensic*. 2012;13:332.
6. Wang X, Zhang K, Zhang X, Li S, Yan J. Aerial infrared object tracking via an improved long-term correlation filter with optical flow estimation and surf matching. *Infrared Phys Technol*. 2021;116:Article 103790.
7. Dotenco S, Dalsass M, Winkler L, Würzner T, Brabec C, Maier A, Gallwitz F. Automatic detection and analysis of photovoltaic modules in aerial infrared imagery. In: *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Lake Placid (NY): IEEE; 2016. p. 1–9.
8. de Oliveira AKV, Aghaei M, Rüther R. Aerial infrared thermography for low-cost and fast fault detection in utility-scale PV power plants. *Sol Energy*. 2020;211:712–724.
9. Lee DH, Park JH. Developing inspection methodology of solar energy plants by thermal infrared sensor on board unmanned aerial vehicles. *Energies*. 2019;12(15):2928.
10. Rahaghi AI, Lemmin U, Sage D, Barry DA. Achieving high-resolution thermal imagery in low-contrast lake surface waters by aerial remote sensing and image registration. *Remote Sens Environ*. 2019;221:773–783.
11. Meng L, Zhou J, Liu S, Ding L, Zhang J, Wang S, Lei T. Investigation and evaluation of algorithms for unmanned aerial vehicle multispectral image registration. *Int J Appl Earth Obs Geoinf*. 2021;102(8-10):Article 102403.
12. Liu X, Ai Y, Zhang J, Wang Z. A novel affine and contrast invariant descriptor for infrared and visible image registration. *Remote Sens*. 2018;10(4):658.
13. Li H, Ding W, Cao X, Liu C. Image registration and fusion of visible and infrared integrated camera for medium-altitude

unmanned aerial vehicle remote sensing. *Remote Sens*. 2017;9(5):441.
14. Ma J, Zhang H, Shao Z, Liang P, Xu H. GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion. *IEEE Trans Instrum Meas*. 2020;70:1.
15. Batur E, Maktav D. Assessment of surface water quality by using satellite images fusion based on PCA method in the Lake Gala, Turkey. *IEEE Trans Geosci Remote Sens*. 2018;57:2983.
16. Rao D, Xu T, Wu X-J. Tgfuse: An infrared and visible image fusion approach based on transformer and generative adversarial network. *IEEE Trans Image Process*. 2023.
17. Latger J, Cathala T, Douchin N, Le Goff A. Simulation of active and passive infrared images using the se-workbench. In: *Infrared imaging systems: Design, analysis, modeling, and testing XVIII 6543*. Orlando (FL): SPIE; 2007. p. 11–25.
18. Cathala T, Douchin N, Joly A, Perzon S. The use of se-workbench for aircraft infrared signature, taken into account body, engine, and plume contributions. In: *Infrared imaging systems: Design, analysis, modeling, and testing XXI 7662*. Orlando (FL):SPIE; 2010. p. 269–276.
19. Jian-xun L. Calculation and image simulation of aircraft infrared radiation characteristic. *Acta Armamentarii*. 2012;33:1310.
20. Bezerra L, Oliveira MM, Rolim TL, Conci A, Santos FGS, Lyra PRM, Lima RCF. Estimation of breast tumor thermal properties using infrared images. *Signal Process*. 2013;93(10):2851.
21. Mielikainen J, Huang B, Huang H-LA. GPU-accelerated multi-profile radiative transfer model for the infrared atmospheric sounding interferometer. *IEEE J Sel Top Appl Earth Observ Remote Sens*. 2011;4(3):691–700.
22. Kniaz VV, Knyaz VA, Hladuvka J, Kropatsch WG, Mizginov V, Thermalgan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In: *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*; 2018; Munich, Germany.
23. Jia X, Zhu C, Li M, Tang W, Zhou W. Llvip: A visible-infrared paired dataset for low-light vision. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Montreal, BC, Canada; 2021. p. 3496–3504.
24. Isola P, Zhu J-Y, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, HI, USA; 2017. p. 1125–1134.
25. Liu M-Y, Breuel T, Kautz J. Unsupervised image-to-image translation networks. *Adv Neural Info Process Syst*. 2017;30.
26. Zhu J-Y, Zhang R, Pathak D, Darell T, Efros AA, Wang O, Sechtman E. Toward multimodal image-to-image translation. *Adv Neural Info Process Syst*. 2017;30.
27. Huang X, Liu M-Y, Belongie S, Kautz J. Multimodal unsupervised image-to-image translation. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Munich, Germany; 2018. p. 172–189.
28. Lin J, Xia Y, Qin T, Chen Z, Liu T-Y. Conditional image-to-image translation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT, USA; 2018. p. 5524–5532.
29. Yi Z, Zhang H, Tan P, Gong M. Dualgan: Unsupervised dual learning for image-to-image translation. In: *Proceedings of the IEEE International Conference on Computer Vision*. Venice, Italy; 2017. p. 2849–2857.
30. Choi Y, Choi M, Kim M, Ha J-W, Kim S, Choo J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT, USA; 2018. p. 8789–8797.
31. Richardson E. Encoding in style: A stylegan encoder for image-to-image translation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Online; 2021. p. 2287–2296.
32. Wu P-W, Lin Y-J, Chang C-H, Chang EY, Liao S-W. Relgan: Multi-domain image-to-image translation via relative attributes. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Seoul, South Korea; 2019. p. 5914–5922.
33. Liu M-Y. Few-shot unsupervised image-to-image translation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Seoul, South Korea; 2019. p. 10551–10560.
34. Han J, Shoeiby M, Petersson L, Armin MA. Dual contrastive learning for unsupervised image-to-image translation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Online; 2021. p. 746–755.
35. Lee H-Y, Tseng HY, Mao Q, Huang JB, Lu YD, Singh M, Yang MH. Drit++: Diverse image-to-image translation via disentangled representations. *Int J Comput Vis*. 2020;128:2402.
36. Fu H. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach, CA, USA; 2019. p. 2427–2436.
37. Benaim S, Wolf L. One-sided unsupervised domain mapping. *Adv Neural Inf Proces Syst*. 2017;30:752–762.
38. Shen Z, Huang M, Shi J, Xue X, Huang TS. Towards instance-level image-to-image translation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. Long Beach, CA, USA; 2019. p. 3683–3692.
39. H.-Y. Chang, Z. Wang, Y.-Y. Chuang. Domain-specific mappings for generative adversarial style transfer. In: *European Conference on Computer Vision*. Glasgow (KY); Springer; 2020. p. 573–589.
40. T. Park, A. A. Efros, R. Zhang, J.-Y. Zhu. Contrastive learning for unpaired image-to-image translation. In: *European Conference on Computer Vision*. Glasgow (KY); Springer; 2020. p. 319–345.
41. Wang W, Yu X, Fang B, Zhao Y, Chen Y, Wei W, Chen J. Cross-modality LGE-CMR segmentation using image-to-image translation based data augmentation. *IEEE/ACM Trans Comput Biol Bioinform*. 2022;20(4):2367–2375.
42. 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, Canada; 2023. p. 1921–1930.
43. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Commun ACM*. 2017;60:84.
44. He K, Zhang X, Ren S, Sun J. *Proc IEEE Conf Comput Vis Pattern Recognit*. 2016;770–778.
45. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv. 2014. https://doi.org/10.48550/arXiv.1409.1556
46. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial networks. *Commun ACM*. 2020;63:139.
47. Mao X. Least squares generative adversarial networks. In: *Proceedings of the IEEE International Conference on Computer Vision*. Venice, Italy; 2017. p. 2794–2802.

48. Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv. 2015. https://doi.org/10.48550/arXiv.1511.06434
49. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC. Improved training of Wasserstein GANs. *Adv Neural Inf Proces Syst*. 2017;30:5769–5779.
50. Cai Z, Xiong Z, Xu H, Wang P, Li W, Pan Y. Generative adversarial networks: A survey toward private and secure applications. *ACM Computing Surveys (CSUR)*. 2021;54:1.
51. Zhu J-Y, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *Proceedings of the IEEE International Conference on Computer Vision*. Venice, Italy; 2017. p. 2223–2232.
52. 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; 2018. p. 35–51.
53. Pathak D, Krahenbuhl P, Donahue J, Darrell T, Efros AA. Context encoders: Feature learning by inpainting. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV, USA; 2016. p. 2536–2544.
54. Zhang R, Isola P, Efros AA. Colorful image colorization. In: *European Conference on Computer Vision*. Amsterdam (Netherlands); Springer; 2016. p. 649–666.
55. Yuan Y. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. Salt Lake City, UT, USA; 2018. p. 701–710.
56. Kim J, Lee JK, Lee KM. Accurate image super-resolution using very deep convlolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV, USA; 2016. p. 1646–1654.
57. Wang J, Gao K, Zhang Z, Ni C, Hu Z, Chen D, Wu Q. Multisensor remote sensing imagery super-resolution with conditional Gan. *J Remote Sens*. 2021;2021.
58. Wang B, Zhang S, Feng Y, Mei S, Jia S, du Q. Hyperspectral imagery spatial super-resolution using generative adversarial network. *IEEE Trans Comput Imag*. 2021;7:948.
59. Li R, Pan J, Li Z, Tang J. Single image dehazing via conditional generative adversarial network. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT, USA; 2018. p. 8202–8211.
60. Cho Y, Jang H, Malav R, Pandey G, Kim A. Underwater image dehazing via unpaired image-to-image translation. *Int J Control Autom Syst*. 2020;18:605.
61. Chen J, Chen J, Chao H, Yang M. Image blind denoising with generative adversarial network based noise modeling. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT, USA; 2018. p. 3155–3164.
62. Yan L, Zheng W, Wang F-Y, Gou C. Joint image-to-image translation with denoising using enhanced generative adversarial networks. *Signal Process Image Commun*. 2021;91:Article 116072.
63. Wang Y, Zhang Z, Hao W, Song C. Multi-domain image-to-image translation via a unified circular framework. *IEEE Trans Image Process*. 2020;30:670.
64. Lu Y, Lu G. Bridging the invisible and visible world: Translation between rgb and ir images through contour cycle gan. In: *2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. Virtual: IEEE; 2021. p. 1–8.
65. Kniaz VV, Knyaz VA. Multispectral person re-identification using gan for color-to-thermal image translation. In: *Multimodal Scene Understanding*. London (UK): Elsevier; 2019. p. 135–158.
66. He K, Fan H, Wu Y, Xie S, Girshick R. Momentum contrast for unsupervised visual representation learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle, WA, USA; 2020. p. 9729–9738.
67. Kim T, Cha M, Kim H, Lee JK, Kim J. Learning to discover cross-domain relations with generative adversarial networks. In: *International Conference on Machine Learning*. Sydney (Australia): PMLR; 2017. p. 1857–1865.
68. Mao Q, Lee H-Y, Tseng H-Y, Ma S, Yang M-H. Mode seeking generative adversarial networks for diverse image synthesis. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach, CA, USA; 2019. p. 1429–1437.
69. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. Gans trained by a two time-scale update rule converge to a local Nash equilibrium. *Adv Neural Inf Proces Syst*. 2017;30:6629–6640.
70. Bińkowski M, Sutherland DJ, Arbel M, Gretton A. Demystifying MMD GANs. arXiv. 2018. https://doi.org/10.48550/arXiv.1801.01401
71. Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT, USA; 2018. p. 586–595.
72. Lian J, Mei S, Zhang S, Ma M. Benchmarking adversarial patch against aerial detection. *IEEE Trans Geosci Remote Sens*. 2022;60:5634616.
73. Cheng G, Zhou P, Han J. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. *IEEE Trans Geosci Remote Sens*. 2016;54(12):7405.
74. Mei S, Jiang R, Ma M, Song C. Rotation-invariant feature learning via convolutional neural network with cyclic polar coordinates convolutional layer. *IEEE Trans Geosci Remote Sens*. 2023;61:5600713.
75. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. *Adv Neural Inf Proces Syst*. 2015;28:91–99.
76. J. Redmon, A. Farhadi, Yolov3: An incremental improvement. arXiv. 2018. https://doi.org/10.48550/arXiv.1804.02767
77. Jocher G. ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements; 2020.
78. Z. Ge, S. Liu, F. Wang, Z. Li, J. Sun, Yolox: Exceeding yolo series in 2021. arXiv. 2021. https://doi.org/10.48550/arXiv.2107.08430

Journal of Remote Sensing
A SCIENCE PARTNER JOURNAL

Aerial Visible-to-Infrared Image Translation: Dataset, Evaluation, and Baseline

Zonghao Han, Ziye Zhang, Shun Zhang, Ge Zhang, and Shaohui Mei

Citation: 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. DOI: 10.34133/remotesensing.0096

Aerial visible-to-infrared image translation aims to transfer aerial visible images to their corresponding infrared images, which can effectively generate the infrared images of specific targets. Although some image-to-image translation algorithms have been applied to color-to-thermal natural images and achieved impressive results, they cannot be directly applied to aerial visible-to-infrared image translation due to the substantial differences between natural images and aerial images, including shooting angles, multi-scale targets, and complicated backgrounds. In order to verify the performance of existing image-to-image translation algorithms on aerial scenes as well as advance the development of aerial visible-to-infrared image translation, an Aerial Visible-to-Infrared Image Dataset (AVIID) is created, which is the first specialized dataset for aerial visible-to-infrared image translation and consists of over 3,000 paired visible-infrared images. Over the constructed AVIID, a complete evaluation system is presented to evaluate the generated infrared images from 2 aspects: overall appearance and target quality. In addition, a comprehensive survey of existing image-to-image translation approaches that could be applied to aerial visible-to-infrared image translation is given. We then provide a performance analysis of a set of representative methods under our proposed evaluation system on AVIID, which can serve as baseline results for future work. Finally, we summarize some meaningful conclusions, problems of existing methods, and future research directions to advance state-of-the-art algorithms for aerial visible-to-infrared image translation.
Image

Journal of Remote Sensing (ISSN 2694-1589) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005.