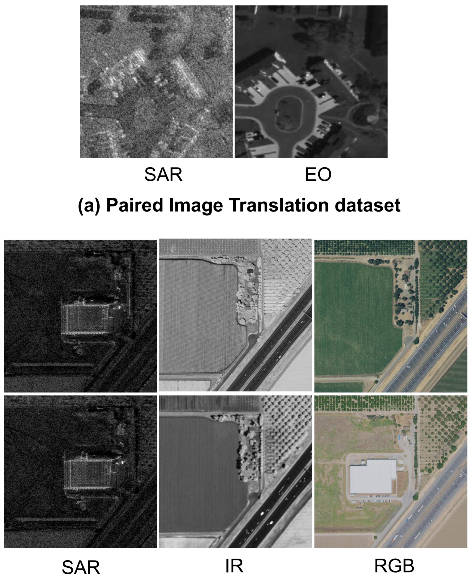
MvAV-pix2pixHD: Multi-view Aerial View Image Translation

Jun Yu[1] Keda Lu[1,2*] Shenshen Du[1*] Lin Xu[1] Peng Chang [3†]
Houde Liu[4] Bin Lan[5] Tianyu Liu[5]
[1]University of Science and Technology of China [2]Ping An Technology Co., Ltd, China
[3]PAII Inc. [4]Tsinghua Shenzhen International Graduate School
[5]Jianghuai Advance Technology Center
harryjun@ustc.edu.cn {lukeda, dushens}@mail.ustc.edu.cn
changpeng805@paii-labs.com liu.hd@sz.tsinghua.edu.cn
liutianyu18@mails.ucas.ac.cn {xulin0114,lanbin.thu}@gmail.com

Abstract

Multi-modal aerial view image translation involves converting aerial images from one modality to another while preserving basic details and features. These modalities encompass Synthetic Aperture Radar (SAR), Infrared (IR), Visible Light (RGB), Electro-Optical (EO), and other image types. Recently, various methods have been proposed to tackle this task, but the focus tends to be on paired image research, overlooking the discrepancies found in aerial images of the same location captured at different times and angles, termed incomplete matching or multi-view image translation. Consequently, we propose MvAV-pix2pixHD to address this issue. For multi-view data sampling, we propose two methods: random sampling and time-priority sampling. Additionally, within the pix2pixHD framework, we introduce an inverse generator to ensure the basic semantic features of the generated images and incorporate three robust loss functions to constrain the authenticity of the generated images. We conduct extensive experiments on two multi-view image translation tasks in the Multi-modal Aerial View Imagery Challenge: Translation (MAVIC-T). Experimental results demonstrate the superiority of our proposed method, and we achieved second place in the MAVIC-T competition in the 20th IEEE Workshop on Perception Beyond the Visible Spectrum of the CVPR 2024.



(a) Paired Image Translation dataset

(b) Multi-view Image Translation dataset

Figure 1. **An illustration on Paired and Multi-view Datasets.** The multi-view datasets exhibit a certain degree of correlation (same location, different times, different angles).

1. Introduction

In recent years, Generative Adversarial Networks (GANs) models [9] have emerged and gradually gained dominance in image translation tasks. These models employ an encoder-decoder structure, where the encoder encodes the input image from the source domain into a low-dimensional feature vector, and the decoder decodes this feature vector into an image with the desired style from the target domain. During the training process, the models utilize paired data and attempt to learn how to transform the style of input images into the style of the target images by minimizing the discrepancy between the generated images and real target images. Due to their ability to generate realistic images in

*Equal contribution.
†Corresponding author.

image translation tasks, GAN-based image translation models have found wide applications in various everyday and entertainment scenarios [13, 21, 28, 34].

Remote sensing technology involves multiple image modalities, including EO, SAR, RGB, and IR images. In remote sensing image translation, it is common to translate SAR images into other modalities; however, due to the presence of significant noise, the translation results often fall short of expectations. Among these translations, the most extensively studied is SAR-to-EO, where Yu et al. [7] have demonstrated the effectiveness of pix2pixHD on SAR-to-EO paired datasets, while the translation between other modalities remains to be further explored.

However, in practical scenarios of aerial image translation tasks, the collection of paired datasets can be challenging. Aerial images are often obtained from different platforms and perspectives, and they can be influenced by factors such as lighting and environment, making it difficult to directly pair them with images from other modalities, such as ground-level images. In such cases, a common approach is to train models using unpaired data and employ unsupervised learning or self-supervised learning methods for image translation. Unsupervised learning methods can automatically learn the correspondence between images from a large set of unpaired images, enabling image translation. Common unsupervised learning methods include CycleGAN [35] and Autoencoders [17], which do not require explicit annotation of paired data but instead learn the transformation relationship between images by minimizing reconstruction errors or adversarial losses.

Multi-modal Aerial View Imagery Challenge: Translation (MAVIC-T) hosts four different modal image translation tracks, including two paired translation tasks, SAR2EO and RGB2IR, and two multi-view translation tasks, SAR2RGB and SAR2IR, as shown in Fig. 1. This paper first validates the effectiveness of pix2pixHD on SAR2EO and RGB2IR tasks, then proposes MvAV-pix2pixHD for multi-view image translation based on pix2pixHD. For multi-view data sampling, we propose two methods: random sampling and time proximity sampling. Additionally, we introduce identity loss and propose high-level perceptual loss to constrain the authenticity of generated images based on the pix2pixHD architecture. Finally, we introduce a reverse generator to ensure the basic semantic features of the generated images.

We conduct extensive experiments on all four image translation tasks in MAVIC-T. Firstly, we validate the effectiveness of pix2pixHD on SAR2EO and RGB2IR tasks, where SAR2EO and RGB2IR individually rank second. Subsequently, we thoroughly validate the effectiveness of MvAV-pix2pixHD on SAR2RGB and SAR2IR tasks, achieving first place in SAR2RGB. In the final leaderboard of MAVIC-T across all four tracks, we achieve an average score of 0.33, ranking the second place.

Overall, our main contribution is as follows:

- We propose two multi-view data sampling methods: random sampling and time proximity sampling, aimed at capturing the representation space of the target domain from different perspectives.
- We propose MvAV-pix2pixHD aimed at addressing multi-view translation tasks, achieved through a inverse generator and three powerful loss functions on top of the pix2pixHD framework.
- We validate the effectiveness of pix2pixHD on two paired datasets and demonstrate significant improvements with our proposed MvAV-pix2pixHD on two multi-view datasets.

2. Related Work

2.1. Paired Image Translation

The objective of image translation is to acquire the mapping relationship between images from a source domain and images from a target domain. Paired image translation methods aim to establish this mapping by training a collection of paired image pairs [12, 21, 28, 34]. These methods rely on datasets that provide a one-to-one correspondence between input and output images.

The Generative Adversarial Network (GAN) model, introduced by Goodfellow et al. [9], has indeed become a cornerstone in various fields, including image translation. Paired image translation, in particular, involves converting an image from one domain to another while preserving its semantic content. GANs are well-suited for this task due to their ability to learn complex mappings between domains without explicit supervision.

The pix2pix model [14] addresses limitations in traditional GANs for paired image translation by enforcing a specific correspondence between input and output images. However, it struggles with generating high-resolution images with realistic details and texture. To overcome these challenges, pix2pixHD [28] introduces enhancements such as a coarse-to-fine generator, multi-scale discriminators, and an improved adversarial loss. The coarse-to-fine generator progressively refines low-resolution images to high resolution, enabling the generation of more detailed textures. Multi-scale discriminators evaluate image details at different resolutions, preserving fine-grained textures. Enhanced adversarial loss encourages the generation of more realistic images. These improvements enhance the performance of paired image translation, resulting in visually compelling and detailed output images.

2.2. Unpaired Image Translation

However, collecting datasets with one-to-one mappings for supervised learning is often difficult in real-world scenarios [1, 4]. Unpaired image translation methods have been developed to address this challenge by allowing mapping between multiple domains where direct matches between images are not available [16, 20, 27, 29, 30]. However, these methods can be affected by unwanted images, hindering their ability to focus on the most relevant parts of the images. Researchers are actively working on improving these models to enhance their reliability and robustness for more accurate and meaningful mappings between diverse domains.

In the context of unpaired image-to-image translation, various methods have emerged to establish connections between two data domains, labeled as X and Y. Rosales et al. [22] introduced a Bayesian framework incorporating a prior derived from a patch-based Markov random field from a source image, along with a likelihood term derived from multiple style images. More recently, CoGAN [19] and cross-modal scene networks [2] employ a weight-sharing strategy to learn a shared representation across different domains. In parallel, Liu et al. [20] extended this framework by combining variational autoencoders (VAEs) and generative adversarial networks (GANs). Concurrently, another line of research [17] focuses on encouraging shared ”content” features between input and output, despite potential style differences. These methods utilize adversarial networks and incorporate additional terms to enforce output similarity to input within predefined metric spaces, such as class label space [5], image pixel space [24], or image feature space [26]. By leveraging neural networks, CycleGAN [34] modifies image style while preserving content, enabling effective unpaired image translation.

2.3. Multi-view image Translation

When source images or target main have multiple views, achieving satisfactory modeling and image transformation results becomes extremely challenging due to the excessive input information. Zhou et al. [32] proposed a method for generating new views of the same object by learning appearance flows. Recently, inspired by [25], [3] presented a cVAE-GAN model for generating multi-view images of clothing based on a single-view image. However, there are fewer studies related to the multi-view aerial view translation task.

3. Methodology

In this section, we first review the pix2pixHD method applied to paired image translation, focusing on its powerful generator, discriminator, and the loss function it uses. Then, we introduce the proposed MvAV-pix2pixHD, applied to the multi-view aerial view image translation task, as shown in Fig. 2.

3.1. Preliminary : pix2pixHD

Similar to the pix2pix [14] model, pix2pixHD [28] is a conditional GAN framework for image-to-image translation. For our task, the goal of the generator G is to translate the image from the source domain into a realistic image similar to the target domain, while the goal of the discriminator D is to distinguish the real image from the translated image. The framework operates in a supervised environment. In other words, the training dataset is a set of corresponding image pairs $\{(x, y)\}$, where x is an image from the source domain and y is the corresponding target domain image. The goal of conditional GAN is to model the conditional distribution of real images of the input semantic labeling graph by the following minimum game:

$$\min_G \max_D \mathcal{L}_{\text{GAN}}(G, D) \quad (1)$$

The pix2pix model is unable to generate high-resolution images, and the generated images lack details and realistic texture. Thus, pix2pixHD proposes the following solutions: coarse-to-fine generator, multi-scale discriminators, and improved adversarial loss to improve the above problems[28].

Coarse-to-fine generator: The generator is divided into two sub-networks: G_1 and G_2, with G_1 serving as the global generator and G_2 as the local enhancer. The generator is represented as $G = G_1$, G_2, where G_1 operates at a lower resolution and G_2 can be utilized to synthesize higher-resolution images. G_1 consists of three components: a convolutional front-end G_1^F, a set of residual blocks G_1^R, and a transposed convolutional back-end G_1^B. Similarly, G_2 also comprises three components: a convolutional front-end G_2^F, a set of residual blocks[10] G_2^R, and a transposed convolutional back-end G_2^B. The training process involves initially training G_1, followed by training G_2 in ascending order of resolution. Ultimately, all networks undergo fine-tuning together. This generator design effectively integrates global and local information for image synthesis.

Multi-scale discriminators: For a generative network, designing a discriminator is a rather difficult task. Compared to low-resolution images, for high-resolution images, the discriminator requires a large receptive field, which requires a large convolutional kernel or a deeper network structure. Adding a large convolutional kernel or deepening the network is easy to cause overfitting and will increase the computational burden. To solve the above problems, multi-scale discriminators have been proposed. It consists of three different discriminators, $D = D_1, D_2, D_3$, which are the same network structure but operate on different image scales. Then, the generated images are downsampled with a factor of 2 and 4, resulting in three images with different resolutions, which are then inputted into the three identical discriminators. This way, the D corresponding to the image with the smallest resolution will have a larger receptive field, providing a stronger global sense for image

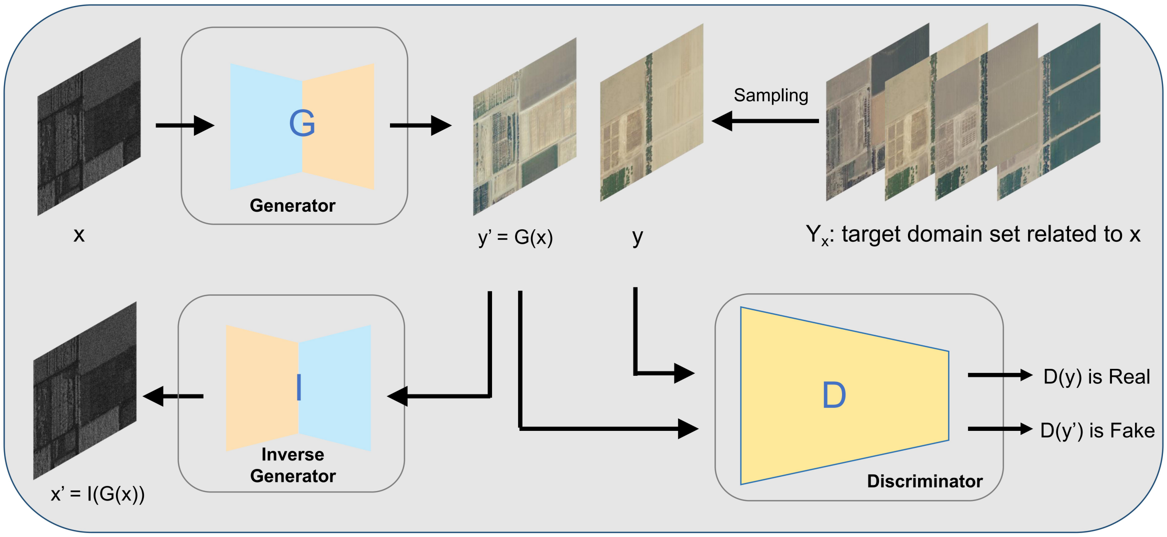


Figure 2. **The pipeline of our proposed MvAV-pix2pixHD method applied to multi-view image translation tasks.** Using the SAR2RGB task as an example, SAR is the input domain and RGB is the output target domain. The generator G and discriminator D use coarse-to-fine generator and multi-scale discriminators which are the same as those proposed by pix2pixHD [28], and the structure of the inverse generator I is exactly similar to that of the generator G, but the training parameters are not shared.

generation, while the D corresponding to the image with the largest resolution will capture finer features.

With the discriminators, the learning problem in Eq. (1) then becomes a multi-task learning problem of

$$\min_{G} \max_{D_1,D_2,D_3} \sum_{k=1}^{3} \mathcal{L}_{\text{GAN}}(G, D_k) \tag{2}$$

where the objective function $\mathcal{L}_{\text{GAN}}(G, D_k)$[1] is given by

$$\mathbb{E}(x,y)[\log D_k(x,y)] + \mathbb{E}_s[\log(1 - D_k(x, G(x)))]. \tag{3}$$

Improved adversarial loss: To match the generator that can produce natural statistics at multiple scales, Wang et al. [28] proposed feature matching loss based on the multi-scale discriminator. Specifically, they extract features from the multi-scale discriminator and learn how to match real image y and synthetic image $G(x)$ with these intermediate representations. For ease of presentation, we denote the i-th layer feature extractor of the discriminator D_k as $D_k^{(i)}$ (from input to the i-th layer of D_k). The feature matching loss $\mathcal{L}_{\text{FM}}(G, D_k)$ is then calculated as:

$$\mathcal{L}_{\text{FM}}(G, D_k) = \mathbb{E}_{(x,y)\sim p_{\text{data}}(x,y)} \sum_{i=1}^{T} \frac{1}{N_i} \cdot \\ [\| D_k^{(i)}(x,y) - D_k^{(i)}(x, G(x)) \|_1] \tag{4}$$

where T is the total number of layers and N_i denotes the number of elements in each layer. This GAN discriminator feature matching loss is related to perceptual loss [6, 8, 15], which has been shown to be useful for image super-resolution [18] and style transfer [15]. Wang et al. [28] discuss how the discriminator feature matching loss and the perceptual loss can be jointly used for further improving the performance. For ease of presentation, we denote the i-th layer feature extractor of the pretrained VGG-19 loss network V as $\mathrm{V}^{(i)}$ (from input to the i-th layer of V). The perceptual loss $\mathcal{L}_p(G)$ is then calculated as:

$$\mathcal{L}_p(G) = \mathbb{E}_{(x,y)} \sum_{i=1}^{4} \frac{1}{2^{6-i}} \cdot [\| \mathrm{V}^{(i)}(G(x)) - \mathrm{V}^{(i)}(y) \|_1] \\ + [\| \mathrm{V}^{(5)}(G(x)) - \mathrm{V}^{(5)}(y) \|_1] \tag{5}$$

The final optimization objective for pix2pixHD are as follows:

$$\min_{G} \left(\begin{array}{c} \left(\max\limits_{D_1,D_2,D_3} \sum\limits_{k=1,2,3} \mathcal{L}_{\text{GAN}}(G, D_k) \right) \\ +\lambda_1 \sum\limits_{k=1,2,3} \mathcal{L}_{\text{FM}}(G, D_k) + \lambda_2 \mathcal{L}_p(G) \end{array} \right) \tag{6}$$

where λ_1 and λ_2 control the importance of the three terms. It is worth noting that D_3 is not required, and if the resolution is 512 for image translation tasks, using D_1 and D_2 is sufficient.

[1] we denote $\mathbb{E}_x \triangleq \mathbb{E}_{x\sim p_{\text{data}}(x)}$ and $\mathbb{E}_{(x,y)} \triangleq \mathbb{E}_{(x,y)\sim p_{\text{data}}(x,y)}$ for simplicity.

3.2. MvAV-pix2pixHD

Since pix2pixHD [28] has demonstrated its superiority in the task of translating 512 × 512 as well as 1024 × 1024 higher resolution images. Wang et al. [28] proved the power of its generator to outperform common encoder-decoder architectures, Unet, etc., and its discriminators have been well proven to outperform individual discriminators. In addition, Yu et al. [7] also demonstrated its power in the aerial view image SAR2EO task compared to pix2pix. Therefore, we propose MvUAV-pix2pixHD based on pix2pixHD for the multi-view aerial view image translation, as shown in Fig. 2.

3.2.1 Multi-view dataset sampling

Random sampling. Unlike the Unpaired image dataset, the multi-view unpaired dataset does not directly randomly sample images from the complete set of the target domain for training but requires strategic sampling. We randomly sample y based on the set of target images Y$_x$ matched by input x, which is used to form the training pair $\{(x, y)\}$.

Time proximity sampling. In addition to the random sampling method, we have proposed a time proximity sampling method. This is motivated by the fact that different sensors, when collecting data from different modalities at the same location, also record the time of acquisition. Therefore, we can match the images in the source and target domains with respect to the acquisition time of each image to obtain a paired dataset. Certainly, due to the heterogeneity of sensors and variations in acquisition times, even for the same spatial location, there may exist differences.

3.2.2 Pipeline of MvAV-pix2pixHD

The pipeline of our proposed MvAV-pix2pixHD applied to the task of multi-view image translation is shown in Fig. 2. In addition to the common generator G and discriminator D, we introduce an inverse generator I. Due to the stochastic nature of the sampling process of the Multi-view dataset, which is not one-to-one pairing, the original adversarial loss alone does not guarantee that the learned function can map a single input x to the desired output y. To further narrow down the space of possible mapping functions, we argue that for each image x from domain x, the image transformation loop should be able to reduce $G(x)$ to the original image, i.e., $x \to G(x) \to I(G(x)) \approx x$. This also shows that the generated $G(x)$ has enough information to be reduced to x, guaranteeing the basic semantic information of the original image. This idea comes from cycleGAN [33] but is not exactly the same. CycleGAN [33] requires two discriminators to be involved so the model architecture is more complex. In contrast, we focus only on the task of unidirectional target translation in the aerial view scenario and only need one discriminator to ensure that the generated target image $G(x)$ is realistic.

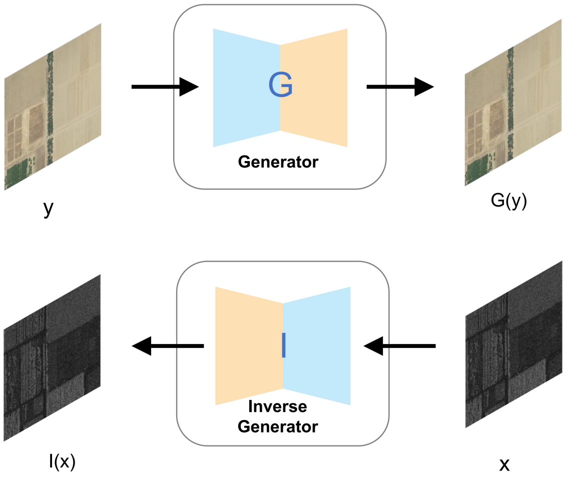


Figure 3. **An illustration of the Identity loss calculation.** G is usually used for X → Y and I for $G(X) \approx$ Y → X. Here the computation is G for Y → Y and I for X → X. This makes the inputs to G and I richer and makes it clearer that the outputs of G are in the Y domain and the outputs of I are in the X domain.

3.2.3 Loss function

Consistency Loss. For each image x from domain X, the image inverse translation should be able to bring $G(x)$ back to the original image, i.e., $x \to G(x) \to I(G(x)) \approx x$. We call this forward inverse consistency. We incentivize this behavior using a consistency loss:

$$\mathcal{L}_{\text{con}}(G, I) = \mathbb{E}_x[\|I(G(x)) - x\|_1] \tag{7}$$

Identity loss. For translation tasks in aerial view image scenarios, the dataset magnitude tends to be small due to the difficulty of data collection. We adapt the technique of Taigman et al. [26] and regularize the generator to be near an identity mapping when real samples of the target domain are provided as the input to the generator:i.e., $\mathcal{L}_{identity}(G, I) = \mathbb{E}_y(y)[\|G(y) - y\|_1] + \mathbb{E}_x(x)[\|I(x) - x\|_1]$. As shown in Fig. 3, this loss guarantees that both the generator and the inverse generator can firmly perform the task of mapping to the target domain when the amount of data is small.

High-level perceptual loss. In order to match the features of the generated $G(x)$ and the real target domain image y, the pix2pixHD method uses feature matching loss and perceptual loss but requires that x and y need to be highly aligned. However, in the multi-view image translation task, x and y are not fully aligned, and there are shooting angle deviations or some translations, so we propose high-level perceptual loss, which computes the similarity using only the outputs of the last two layers of the VGG model. high-level features can represent the deeper semantic features, and the captured perceptual field is larger. Therefore this loss is not limited by the need for height alignment requirement and is more suitable for multi-view aerial view image translation tasks. The high-level perceptual loss $\mathcal{L}_{hp}(G)$ is then calculated as:

$$\begin{aligned}\mathcal{L}_{hp}(G) = \mathbb{E}_{(x,y)}\frac{1}{4}\cdot[\|\mathrm{V}^{(4)}(G(x)) - \mathrm{V}^{(4)}(y)\|_1] \\ +[\|\mathrm{V}^{(5)}(G(x)) - \mathrm{V}^{(5)}(y)\|_1]\end{aligned} \tag{8}$$

Full Objective. The final optimization objective for MvAv-pix2pixHD are as follows:

$$\min_{G,I}\left(\begin{aligned}\left(\max_{D_1,D_2,D_3}\sum_{k=1,2,3}\mathcal{L}_{\text{GAN}}(G,D_k)\right) \\ +\lambda_1\mathcal{L}_{\text{con}}(G,I)+\lambda_2\mathcal{L}_{\text{identity}}(G,I)+\lambda_3\mathcal{L}_{hp}(G)\end{aligned}\right) \tag{9}$$

where λ_1, λ_2 and λ_3 control the importance of the four terms. It is worth noting that D_3 is not required, and if the resolution is 512 for image translation tasks, using D_1 and D_2 is sufficient.

4. Experiments

In this section, we first briefly discuss the dataset and evaluation metrics, followed by implementation details. In addition, we will quantitatively evaluate the performance of our approach on different types of datasets. Finally, we perform some ablation experiments to demonstrate the effectiveness of each component.

4.1. Dataset

The dataset used for SAR2EO consists of two types of small window areas (chips) taken from large images captured by multiple EO and SAR sensors mounted on the aircraft. The EO chips are 256 × 256 pix images and belong to targets taken from an airplane. The SAR chips contain roughly the same field of view as the corresponding EO images and are of matching resolution to the EO images. The dataset is divided into:

The SAR2IR, SAR2RGB, and RGB2IR tasks are derived from the same dataset. This dataset contains four separate locations with a stack of geo-spatially aligned images. The locations are UC Davis, Califonia; Manhattan, New York; Bingham Copper Mine, Utah; and Centerfield, Utah. The distribution of the four datasets varies greatly, and it is not suitable to use all of them together for the translation task; instead, the appropriate training set is selected for training according to the actual situation. Among them, the Bingham Copper Mine and Centerfield datasets are mostly suburbs and deserts, the Manhattan dataset is concentrated on the seashore, and the UC Davis dataset is mostly concentrated in the city.

The RGB2IR task is the paired dataset and the SAR2IR and SAR2RGB tasks are the multi-view datasets. Due to sensor variability, the resolutions vary widely for each chip. The resolution of the training data is mostly less than 656 × 656 pix, but validation and testing require the output image resolution of 1024 × 1024 pix.

4.2. Metrics

It is open and difficult to evaluate the quality of synthesized images [23]. L2 Norm, FID and LPIPS are used as the metrics.

L2 Norm, also known as Euclidean distance or L2 distance, is a commonly used distance metric to measure the difference between two vectors.

The L2 Norm measures the length of the vector, which is the distance from the origin to the point represented by the vector. In image processing, the L2 Norm is often used to calculate the pixel-wise difference between two images.

LPIPS (Learned Perceptual Image Patch Similarity) is a perceptual image quality metric that measures the similarity between two images based on the response of deep neural networks. LPIPS was introduced in a paper by [31] and has been shown to correlate well with human perception of image quality.

FID (Fréchet Inception Distance) [11] is a metric for evaluating the quality of generated images by measuring the similarity between the feature representations of generated and real images. It combines two key concepts: the feature representations from the intermediate layers of the Inception network and the Fréchet distance. A lower FID score indicates higher similarity between generated and real images, indicating better image quality produced by the generator. FID has been widely used for evaluating the quality of images generated by Generative Adversarial Networks (GANs).

The evaluation metric for the final ranking of the MAVIC-T Challenge is the average of the above three metrics for the four translation tasks.

4.3. Implementation Details

Because of the differences in the datasets and types of individual tasks, we discuss the implementation details of each of the four tasks separately.

SAR2EO. We use pix2pixHD to address this paired dataset translation task. We train our model for 200 epochs on 4 Nvidia A6000 GPUs with a batch size of 128 and input resolution of 256 × 256 pix. The other settings are consistent with the SAR2EO solution proposed by Yu et al. [7].

Task	Method	Validation phase				Test phase			
		LPIPS ↓	FID ↓	L2 ↓	Average ↓	LPIPS ↓	FID ↓	L2 ↓	Average ↓
SAR2EO	Pix2pix [14]	0.715	0.056	0.202	0.324	-	-	-	-
SAR2EO	Pix2pixHD [28]	0.609	0.041	0.118	**0.256**	0.234	0.020	0.033	**0.096**
RGB2IR	Pix2pixHD [28]	0.438	0.447	0.094	0.327	0.377	0.462	0.109	0.316
RGB2IR	Grayscale mapping	-	-	-	-	0.178	0.216	0.105	**0.166**
SAR2IR	MvAV-pix2pixHD(**R**)	0.632	0.835	0.141	**0.536**	0.665	0.715	0.176	**0.519**
SAR2IR	MvAV-pix2pixHD(**T**)	0.645	0.879	0.172	0.565	0.443	0.900	0.240	0.528
SAR2RGB	MvAV-pix2pixHD(**R**)	0.734	0.868	0.136	0.579	0.701	0.862	0.164	0.575
SAR2RGB	MvAV-pix2pixHD(**T**)	0.689	0.764	0.139	**0.530**	0.687	0.783	0.157	**0.542**

Table 1. **Validation and Test Metrics for Image Translation Tasks of MAVIC-T challenge. R** and **T** respectively represent random sampling and time proximity sampling. These metrics comprise LPIPS, FID, and L2, along with the average value of these three metrics.

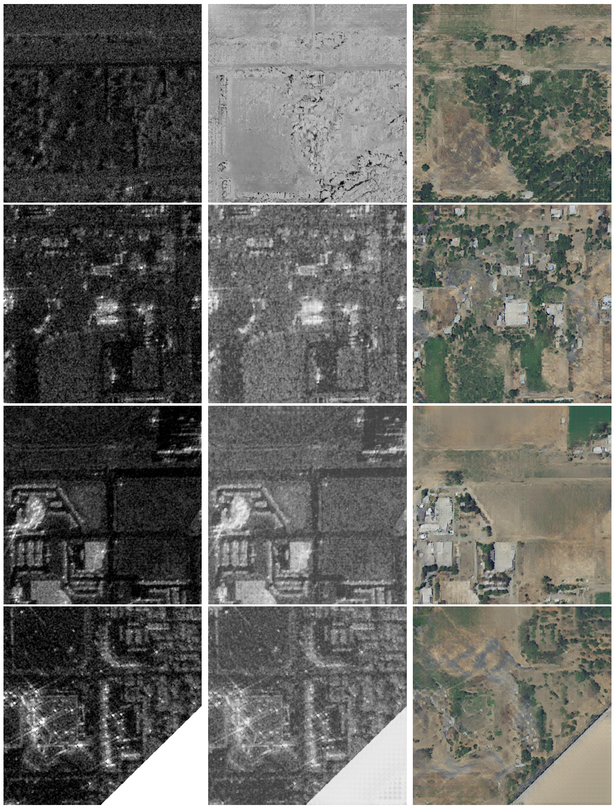


Figure 4. **The effect demonstration of MvAV-pix2pixHD Applied to Multi-View Image Translation Task(SAR2IR and SAR2RGB).**

RGB2IR. We use pix2pixHD to address this paired dataset translation task. We train our model for 200 epochs on 1 Nvidia A6000 GPU with a batch size of 1 and input resolution of 656 × 656 pix. In the development phase, we train with the UC Davis dataset, and in the testing phase, we train with the Manhattan dataset. In the testing stage, the image of 1024 × 1024 target resolution is obtained by resizing. Other settings are consistent with the SAR2EO solution described above.

Due to the nature of this task of RGB2IR, both are noise-free. We analyze the RGB and IR images of the training data, such as grayscale mapping and luminance extraction. We try grayscale mapping then adjusting the brightness of the grayscale maps by setting the adjustment factor, and finally outputting images with a similar distribution as IR.

SAR2RGB and SAR2IR. We use our proposed MvAV-pix2pixHD to address this multi-view dataset translation task. We use the UC Davis dataset to train our all models for 200 epochs on 2 Nvidia A6000 GPUs with a batch size of 4 and input resolution of 656 × 656 pix. The hyperparameters of the loss function are set $\lambda1 = \lambda2 = \lambda3 = 10$, and the other parameter settings are consistent with pix2pixHD. In the testing stage, the image of 1024 × 1024 target resolution is obtained by resizing.

4.4. Main results

As shown in Tab. 1, we present various metrics for four tasks during both the validation and testing phases. In the SAR2EO task, all metrics of pix2pixHD significantly outperform the baseline, the pix2pix method. Therefore, we applied pix2pixHD to the RGB2IR task, which also operates on a paired dataset. Additionally, for the RGB2IR task, we employ grayscale mapping during the testing phase. This proves to be a simple yet effective method, further enhancing the quality of generated images, specifically yielding an LPIPS score of 0.178, a FID score of 0.216, and an L2 score of 0.105.

For the Multi-view dataset, as showmn in Tab. 1, we utilize our proposed MvAV-pix2pixHD method. We assess the impact of using two data sampling methods, namely random sampling and time proximity sampling, on the results. It can be observed that the SAR2IR task performs better with Random sampling, while the SAR2RGB task per-

Rank	Team Name	SAR2EO	RGB2IR	SAR2RGB	SAR2IR	Average
1	NJUST-KMG	0.08 (1)	0.16 (1)	0.55 (3)	0.51 (1)	0.32 (1)
2	**USTC-IAT-United**	0.10 (2)	0.17 (2)	**0.54 (1)**	**0.52 (2)**	0.33 (2)
3	up6	0.12 (5)	0.19 (3)	0.56 (4)	0.54 (3)	0.35 (3)
4	wangzhiyu918	0.11 (4)	0.22 (4)	0.54 (2)	0.55 (4)	0.36 (4)
5	hsansui	0.10 (3)	0.36 (5)	0.57 (5)	0.58 (5)	0.40 (5)

Table 2. **The final leaderboard of Multimodal Aerial View Imagery Challenge: Translation(listing the top five).**

I	loss	LPIPS ↓	FID ↓	L2 ↓
w/o	$\mathcal{L}_p$	0.70	0.97	0.30
w/o	$\mathcal{L}_{\text{identity}}, \mathcal{L}_p$	0.65	0.87	0.17
w/o	$\mathcal{L}_{\text{identity}}, \mathcal{L}_{hp}$	0.64	0.84	0.15
w	$\mathcal{L}_{\text{con}}, \mathcal{L}_{\text{identity}}, \mathcal{L}_{hp}$	**0.63**	**0.83**	**0.14**

Table 3. **Ablation study on the MvAV-pix2pixHD method.** Here, Inverse generator I: w/o and loss: $\mathcal{L}_p$ represent the basic pix2pixHD architecture, indicating only the use of the initial adversarial loss and perceptual loss. The data sampling method for the aforementioned experiments adopts random sampling, and evaluation is conducted specifically on the SAR2IR task.

forms better with Time proximity sampling. Furthermore, as shown in Fig. 4, we demonstrate the generation of high-quality images for the SAR2RGB and SAR2IR tasks.

In the MAVIC-T competition, our USTC-IAT-United team achieve an excellent second place in the final test set. This achievement is attributed to its outstanding performance metrics, notably with SAR2EO task scoring 0.10, RGB2IR task scoring 0.17, SAR2RGB task scoring 0.54, and SAR2IR task scoring 0.52. Our team's composite score is calculated as 0.33, representing the average of the aforementioned performance metrics. These results strongly demonstrate the effectiveness of the team's competition strategy. The final test set leaderboard is provided in Tab. 2, listing the top five teams' scores. It is evident that our proposed MvAV-pix2pixHD application achieved Top-1 and Top-2 rankings for the SAR2RGB and SAR2IR tasks, respectively, showcasing remarkable performance.

4.5. Ablation studies

In this section, we conduct ablation experiments to demonstrate the effectiveness of our framework. The validation dataset was utilized for the experimentation.

Effectiveness of identity loss. Identity loss can be applied in unpaired image translation tasks and similarly enhances multi-view tasks significantly. As shown in Table Tab. 3, it reduces LPIPS and L2 metrics by 10% and 13%, respectively.

Effectiveness of high-level perceptual loss. The perceptual loss simultaneously incorporates low-level and high-level features for computation, which is unreasonable for multi-view image pairs. Therefore, we only utilize high-level features for calculation. As shown in Table Tab. 3, compared to $\mathcal{L}_p$, the introduction of $\mathcal{L}_{hp}$ results in higher quality image generation.

Effectiveness of inverse generator. The inverse generator ensures that the generated image $G(x)$ can be restored to the original image x and largely preserves the semantic content of the original image. In our experiments, we find that adding the inverse generator and the corresponding consistency loss yields better results.

5. Conclusion

This paper validates the effectiveness of pix2pixHD on SAR2EO and RGB2IR paired translation tasks. Additionally, we propose MvAV-pix2pixHD based on pix2pixHD, suitable for translation tasks on multi-view high-definition datasets. We utilize the coarse-to-fine generator and multi-scale discriminators from pix2pixHD to construct our model architecture, ensuring the resolution and quality of the generated images. Furthermore, we introduce the reverse generator and consistency loss to further improve the conversion quality. Additionally, we propose high-level perceptual loss and introduce identity loss to constrain the authenticity of the generated images. We evaluate the performance of our proposed method in the MAVIC-T competition, where all our pix2pixHD-based models achieved an average score of 0.33 across four tasks, ranking second place in the competition. In the future, we will continue to delve into this field, aiming to devise model architectures that are better suited for multi-view aerial view image translation.

Acknowledgments

This work was supported by the Natural Science Foundation of China (62276242), National Aviation Science Foundation (2022Z071078001), CAAI-Huawei MindSpore Open Fund (CAAIXSJLJJ-2021-016B, CAAIXSJLJJ-2022-001A), Anhui Province Key Research and Development Program (202104a05020007), Dreams Foundation of Jianghuai Advance Technology Center (2023-ZM01Z001).

References

[1] Asha Anoosheh, Eirikur Agustsson, Radu Timofte, and Luc Van Gool. Combogan: Unrestrained scalability for image domain translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 783–790, 2018. 3

[2] Yusuf Aytar, Lluis Castrejon, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Cross-modal scene networks. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2303–2314, 2017. 3

[3] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Cvae-gan: fine-grained image generation through asymmetric training. In *Proceedings of the IEEE international conference on computer vision*, pages 2745–2754, 2017. 3

[4] Sagie Benaim and Lior Wolf. One-sided unsupervised domain mapping. *Advances in neural information processing systems*, 30, 2017. 3

[5] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3722–3731, 2017. 3

[6] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. *Neural Information Processing Systems,Neural Information Processing Systems*, 2016. 4

[7] Shenshen Du, Jun Yu, Guochen Xie, Renjie Lu, Pengwei Li, Zhongpeng Cai, and Keda Lu. SAR2EO: A high-resolution image translation framework with denoising enhancement. In *AI 2023: Advances in Artificial Intelligence - 36th Australasian Joint Conference on Artificial Intelligence, AI 2023, Brisbane, QLD, Australia, November 28 - December 1, 2023, Proceedings, Part I*, pages 91–102. Springer, 2023. 2, 5, 6

[8] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR2016*, 2016. 4

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1, 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[12] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. Pmlr, 2018. 2

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2, 3, 7

[15] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. *Perceptual Losses for Real-Time Style Transfer and Super-Resolution*, page 694–711. 2016. 4

[16] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International conference on machine learning*, pages 1857–1865. PMLR, 2017. 3

[17] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. 2, 3

[18] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 105–114, 2017. 4

[19] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. *Advances in neural information processing systems*, 29, 2016. 3

[20] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30, 2017. 3

[21] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019. 2

[22] Resales, Achan, and Frey. Unsupervised image translation. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 472–478. IEEE, 2003. 3

[23] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 6

[24] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2107–2116, 2017. 3

[25] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015. 3

[26] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. 3, 5

[27] Hao Tang, Dan Xu, Wei Wang, Yan Yan, and Nicu Sebe. Dual generator generative adversarial networks for multi-domain image-to-image translation. In *Asian Conference on Computer Vision*, pages 3–21. Springer, 2018. 3

[28] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 2, 3, 4, 5, 7

[29] Yaxing Wang, Joost van de Weijer, and Luis Herranz. Mix and match networks: encoder-decoder alignment for zero-pair image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5467–5476, 2018. 3

[30] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2868–2876, 2017. 3

[31] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[32] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 286–301. Springer, 2016. 3

[33] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 5

[34] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2, 3

[35] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. *Advances in neural information processing systems*, 30, 2017. 2