Multi-modal Aerial View Image Challenge: Sensor Domain Translation

Spencer Low
Brigham Young University
Provo, Utah
spencerlow@byu.edu

Oliver Nina
Air Force Research Laboratory
Dayton, OH
oliver.nina.1@afresearchlab.com

Dylan Bowald
Air Force Research Laboratory
Dayton, OH
dylan.bowald.1@us.af.mil

Angel D. Sappa
ESPOL Polytechnic University, Ecuador
Computer Vision Center, Spain
sappa@ieee.org

Nathan Inkawhich
Air Force Research Laboratory
Rome, NY
nathan.inkawhich@us.af.mil

Peter Bruns
University of Utah
Salt Lake, UT
brunsp10@gmail.com

Abstract

This paper describes the design, outcomes, and top methods of the 2nd annual Multi-modal Aerial View Image Challenge (MAVIC) aimed at cross modality aerial image translation. The primary objective of this competition is to stimulate research efforts towards the development of models capable of translating co-aligned images between multiple modalities. Specifically, the challenge centers on translation between synthetic aperture radar (SAR), electro-optical (EO), camera (RGB), and infrared (IR) sensor modalities, a budding area of research that has begun to garner attention. While last year's inaugural challenge demonstrated the feasibility of SAR→EO translation, this year's challenge made significant improvements in dataset coverage, sensor variation, experimental design, and methods covering the tasks of SAR→EO, SAR→RGB, SAR→IR, RGB→IR translation. By introducing a new dataset called Multi-modal Aerial Gathered Image Composites (MAGIC); multimodal image translation is available for different comparisons. With a more rigorous set of translation performance metrics, winners were determined from aggregation of L1-norm, LPIPS (Learned Perceptual Image Patch Similarity, and FID (Frechet Inception Distance) scores. The wining methods included the pix2pixHD and LPIPS metrics as loss functions with an aggregated score 5% better separated by the SAR→EO and RGB→IR translation scores.

1. Introduction

Last year's inaugural Multi-modal Aerial View Image Challenge - Translation (MAVIC-T) challenge was primarily engaged in the utilization of data from Synthetic Aperture Radar (SAR) and Electro-Optical (EO) imagery from one common collection [13]. Building on this foundation, the 2024 MAVIC-T shifts focus towards the transformation of data across multiple modalities (e.g., infrared (IR)) and different collections spanning different times, locations, and context. Data translations provides a unique opportunity to harness the distinctive benefits offered by various sensing technologies, while simultaneously minimizing coverage gaps. This process can be used to bolster existing data volume. The MAVIC-T challenge aims to enhance sensor data utility via modality conversion, increase data diversity across modalities and regions, and serve as a platform for advancing conditioned image synthesis techniques.

By prioritizing the conversion of data between modalities, MAVIC-T aims to broaden the scope of multi-modal research, fostering the development of more adaptable and robust models. The multi-modal translation among many sensor sources emphasizes the critical importance of diversifying data sources to overcome the challenges associated with sensor-specific limitations, thereby advancing the field of multi-modal image analysis. MAVIC-T focuses on four main translation tasks: RGB→IR; SAR→EO; SAR→IR; and SAR→RGB. As shown in Fig. 1.

The foundation of this challenge lies in leveraging the distinct benefits of various data modalities to overcome their

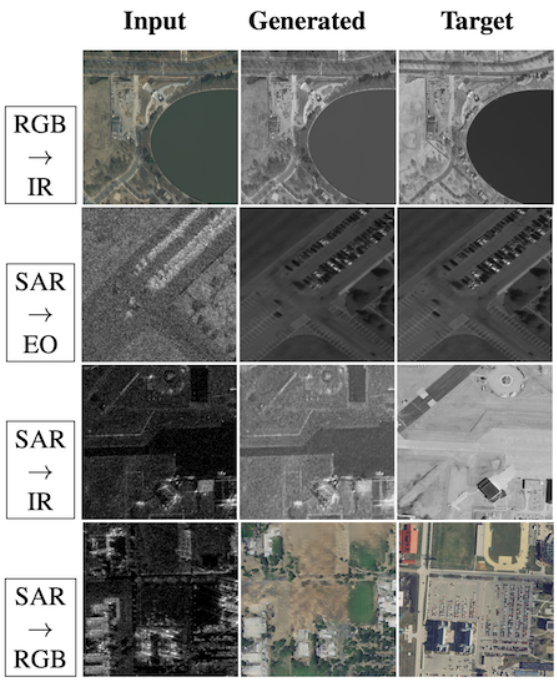


Figure 1. Example of the image translation task. Input images are shown on the left, generated images in the middle, target images on the right.

availability constraints. Synthetic Aperture Radar (SAR) sensors, with their all-weather capabilities and ability to penetrate atmospheric obstructions, offer unique advantages over Electro-Optical (EO) sensors. However, SAR imagery is complex to interpret and less available. Infrared (IR) imagery, useful for thermal imaging and nighttime operations, similarly faces availability challenges, though to a lesser extent than SAR [12]. Given the widespread availability of EO imagery, translating EO data into the IR modality emerges as a strategic solution to improve data diversity and address coverage gaps. Conversely, due to the limitations of EO and IR data, translating SAR data into EO or IR data can enhance the availability of both modalities in adverse conditions. This initiative aims to mitigate the scarcity of SAR and IR data among a plethora of EO data.

The development of models that can translate between sensors of different modalities can enable the utilization of established algorithms. An example use case utilizing aerial imagery is in vision aided navigation, which compares live aircraft imagery to reference data in order to infer a position. As shown in Fig. 2, which has been proposed as a method to navigate in areas where Global Navigation Satellite Systems (GNSS) may not be available or reliable. This includes navigation in GNSS denied or untrusted environments [22] or in areas where GNSS is completely unavailable such as in lunar based navigation [6].

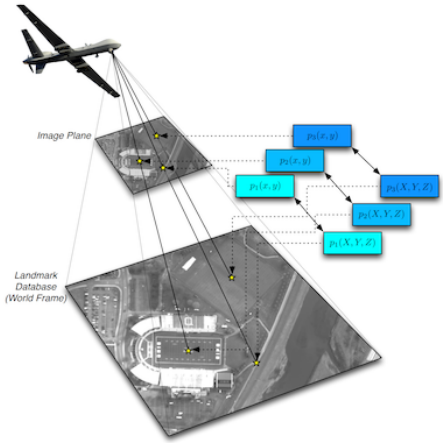


Figure 2. IR data from vision aided navigation [22]

Another application of image translation includes automatic target recognition (ATR) tasks [8, 15]. In modalities like SAR, it is often harder to find large amounts of labeled training data for objects of interest than it is to find training data in EO [7, 9]. Thus, translating EO→SAR may offer a way to train better SAR ATRs by using data collected and labeled in the EO modality. This case can be made for the translation tasks as important looks of objects (for training) may be opportunistically collected in any of the sensor modalities.

This paper outlines the advancements observed in sensor translation facilitated by the 2024 MAVIC-T. We also introduce the novel MAGIC dataset, which serves as the backbone to this challenge. Various methods and their performance are further detailed in Section 5.

The manuscript is organized as follows. Section 3 provides an introduction to the challenge dataset, evaluation metrics and competition phases. Section 4 summarizes the results obtained by different teams. Then, Section 5 presents a short description of the top approaches evaluated from submissions. The conclusion is presented in Section 6.

2. Previous Work

The problem of cross domain image translation addressed by the challenge is related with those approaches proposed in the literature for gray scale / near infrared (NIR) / thermal image colorization, or color transfer functions (e.g., [18], [16], [25], [3], just to mention a few). These problems are generally tackled through the use of Generative Adversarial Networks (GANs) [10], which allows the transformation of information between domains. Most GAN based approaches have focused on supervised contexts, where a pairing of correctly registered data are provided. The unpaired problem, which is more challenging, could be tackled by a GAN architecture in the unsupervised context under a cyclic structure (CycleGAN) [28]. CycleGAN learns to map images from one domain (source domain) onto another do-

main (target domain) when paired images are unavailable [21]. The domain adaptation makes models appropriate for image to image translation in the context of unsupervised learning. More recently, diffusion models have been proposed giving superior results than state-of-the-art generative models [4]. Unfortunately, the main limitation with diffusion models lie on the large amount of resources required for their training.

Moreover, it is worth noting that generative models are susceptible to producing hallucinations, which refer to outputs that deviate from the original source information. Such occurrences can be especially consequential in sensor translation tasks, where the preservation of information between modalities is crucial. Therefore, the central aim of the MAVIC challenge is to facilitate the development of reliable translation models that are capable of producing explainable, interpretable, and trustworthy outputs. Converting images across modalities is not a trivial task, as it poses several challenges related to non-collocated sensor collections, pixel intensity association, image size, ground sampling distance, and image noise differences [27].

The 2023 MAVIC challenge, we showed promising initial results on the EO to SAR cross domain translation [14]. We aim to expand on last year's challenge by adding in the IR domain and significantly increasing our data volume, culminating in our new and larger Multi-modal Aerial Gathered Image Composites (MAGIC) dataset. We observe an increase in sophistication from last year's methods when compared to this year's methods.

3. Challenge

The 2024 MAVIC-Translation challenge is held jointly with the Perception Beyond the Visible Spectrum (PBVS) workshop and is a complement to the MAVIC-Classification challenge. The MAVIC-T challenge is designed to facilitate innovative approaches in multi-modal sensor translation. Participants are evaluated on using a weighted average of the L1, LPIPS [26], and FID [5] score. The challenge centers on the advancement of multi-modal translation networks. Participating teams are provided with a collection of cross modality image pairs, consisting of EO, IR, and SAR modalities, and are tasked with performing image translation from one modality to the other. Upon completion, the teams' generated outputs are evaluated on a separate test set that was previously withheld. The performance of the teams is subsequently monitored and recorded. Emphasis is placed on generating high quality translations with an absence of hallucinations. Figure 3 illustrates an example of a failed translation that contains hallucinations. Since a hallucination is related to objects, it was not evaluated as part of the challenge problem tasks.

Figure 3. Example of failed translation from SAR to EO (order = SAR input, translation, ground truth). We draw attention to the aircraft in the image. The translated image illustrates an example of the generative network hallucinating.

3.1. The Multi-modal Aerial Gathered Image Composites Dataset

The current dataset consists of images pairs from the UNICORN dataset, and introduces newly collected, collocated MAGIC SAR, MAGIC IR and MAGIC RGB datasets. This new dataset is termed the Multi-modal Aerial Gathered Image Composites (MAGIC). This dataset is a custom processed, aligned, and transformed agglomeration of data from three sources: UNICORN, USGS HRO [1], and the Umbra [2] open data program. Figure 4 illustrates an example of a stack from the MAGIC dataset.

The UNICORN dataset features a curated SAR-EO dataset that is publicly accessible and aligned using advanced homography techniques. The United States Geological Survey (USGS) provides land change satellite and aerial imagery through its Earth Resources Observation and Science (EROS) program. MAGIC utilizes the High Resolution Orthoimagery (HRO) dataset provided through EROS. HRO contains a large volume of aerial imagery, all under one meter resolution, uniform and scale, and corrected for terrain relief, sensor geometry, and camera tilt (orthorectified).

Umbra is a space technology company that provides over $4 million worth of free SAR imagery through its UMBRA open data program. Through space based remote imaging, they provide up to 16cm remote imaging. This dataset provides MAGIC with additional SAR imagery.

Figure 4. Example stack (RGB, IR, SAR), made from aligning UMBRA and HRO data.

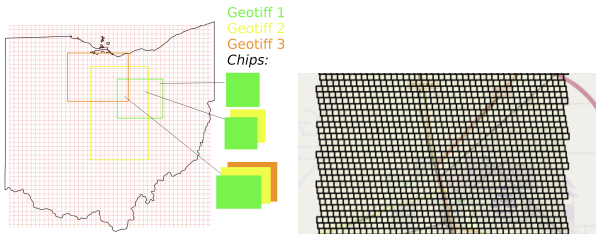


Figure 5. Examples of chipping. $(left)$ illustrates ideal chipping, and $(right)$ illustrates actual chipping used.

3.1.1 Cross Dataset Alignment & Chipping

To allow for cross dataset chipping, we take a "plug in" based approach. For each dataset, we process the images into a specific format and meet certain preconditions. Once the dataset has been processed, we chip the data into co-aligned stacks, covering a specific area. These stacks are then saved and sorted based on the WGS-84 location of its top left point.

To chip the dataset, we preprocess each dataset to ensure the following conditions are met:

1. The images are georectified (north-facing)
2. The images are in a GeoTiff format, where the GeoTiff format contains a transform from pixel space to local coordinate reference system
3. The resulting GeoTiffs are placed in a unified directory structure

Once these preconditions are met, we run our uniform chipping code over the datasets. First, a tile grid is established, such that each tile covers the same area in meters on the ground. To accomplish tiling, we take small arc lengths of the world's circumference, parallel to the latitude parallels and longitude meridians, and project onto the surface of the earth to get a specific size of each tile in meters. We assume that these tiles are approximately flat for sufficiently small tiles, in our case 200m×200m. Every tile is anchored to the 0, 0 latitude/longitude point (aka "null island"), so every tile is unique for a given size. Figure 5 illustrates both ideal and real chipping examples.

3.2. Dataset Split

The MAGIC dataset is split into train, validation, and test sets. In order to more rigorously test generalizability, the validation and test sets are withheld, and sampled from different geographical locations. The validation set is sampled from New Albany, Ohio, and the test set is sampled from Washington DC (see details in Table 1 and 2).

3.3. Evaluation

The MAVIC-T challenge's rigorous evaluation framework is designed to assess submissions across four distinct translation tasks, each contributing to the overarching goal of high-fidelity image translation. We evaluate image translation versus object recognition, from which future studies could evaluate image translation contribution to object recognition. We evaluate submissions on SAR→ EO, SAR→RGB, SAR→IR, and RGB→IR translations.

Table 1. Details of the UNICORN dataset used for training, validation, and testing. It is used exclusively in the SAR→EO task.

Modality	# Train	# Val	# Test
UNICORN SAR	68,151	80	3,586
UNICORN EO	68,151	80	3,586

Table 2. Details of the MAGIC dataset used for training, validation, and testing. This dataset is used exclusively in the SAR→RGB, SAR→IR, and RGB→IR tasks.

Modality	# Train	# Val	# Test
MAGIC SAR	10576	60	60
MAGIC RGB	2273	30	30
MAGIC IR	2273	30	30

Submissions are evaluated based on their performance in each task using a composite score derived from three meticulously chosen metrics:

1. LPIPS (Learned Perceptual Image Patch Similarity): This metric computes a similarity score using deep feature representations, reflecting human perceptual judgments, based on the VGG-16 architecture [17].
2. FID (Fréchet Inception Distance): Utilizing a pre-trained InceptionV3 network [19], the FID quantifies the dissimilarity between distributions of generated and target images, offering insights into visual fidelity and feature distribution similarity.
3. L1 Norm: Focuses on the pixel-wise absolute difference between target and generated images to ensure the structural integrity and overall content accuracy.

The selection of these metrics—LPIPS for perceptual accuracy, FID for distributional similarity, and L1 for content and structural integrity—aims to address the multifaceted aspects of image translation quality. This comprehensive evaluation strategy seeks to minimize generative artifacts while ensuring that generated images exhibit both high-resolution details and structural coherence, aligning closely with the target domain's characteristics.

The evaluation process entails calculating the score for each metric across all four translation tasks, followed by normalization to scale the values between 0 and 1. This normalization is task-specific and is performed as follows:

- L1 Norm, image normalization adjusts pixel values to fall within the desired range.
- LPIPS scores undergo scaling of the output weights to achieve normalization.

- FID scores are normalized employing a weighted arctan activation function to moderate each metric's influence evenly.

The final score for each task is derived by averaging these normalized metrics:

$$\text{Task Score} = \frac{\frac{2}{\pi}\arctan(\text{FID}) + \text{LPIPS} + \text{L1}}{3}. \quad (1)$$

The overall performance of a submission is then determined by averaging the Task Scores across the four translation tasks, ensuring a holistic assessment of the algorithm's capability to produce accurate and perceptually consistent translations across different imaging modalities:

$$\text{Overall Score} = \frac{\text{SAR2EO} + \text{SAR2RGB} + \text{SAR2IR} + \text{RGB2IR}}{4}. \quad (2)$$

This expanded evaluation methodology reflects the MAVIC-T challenge's commitment to fostering the development of versatile and robust image translation models, capable of handling a variety of source and target modalities with high fidelity.

3.4. Challenge Phases

The challenge began on January 19th, 2024. The test data was released on February 22nd, 2024, with the final day for submissions being March 5th.

4. Challenge Results

The challenge results are summarized in this section. This challenge had 95 teams participate. Results from the top ten teams are shown in Table 3, with samples from generated images from the top three teams shown in Figure 8. All the winning methods used pix2pixHD and LPIPS metrics as loss functions. While SAR-RGB and SAR-IR performance is generally the same for the top ten, the main differentiating tasks are SAR-EO and RGB-IR.

Table 3. Top Performing Teams in Competition

Rank	Team	Total ↓	SAR→EO	SAR→RGB	RGB→IR	SAR→IR
1	**NJUST-KMG**	**0.32**	**0.08**	**0.55**	**0.16**	**0.51**
2	USTC-IAT-United	0.33	0.10	0.54	0.17	0.52
3	wangzhiyu918	0.36	0.11	0.54	0.22	0.55
4	hsansui	0.40	0.10	0.57	0.36	0.58
5	lemonGJacky	0.40	0.33	0.57	0.19	0.53
6	Marry	0.42	0.25	0.59	0.30	0.52
7	Levin	0.43	0.25	0.56	0.40	0.52
8	yishifeng	0.35	0.13	0.55	0.55	0.51
9	wuxixian	0.44	0.13	0.55	0.54	0.52
10	xsd	0.44	0.13	0.55	0.55	0.53

5. Methods

This section briefly summarizes the approaches used by the top three participating teams.

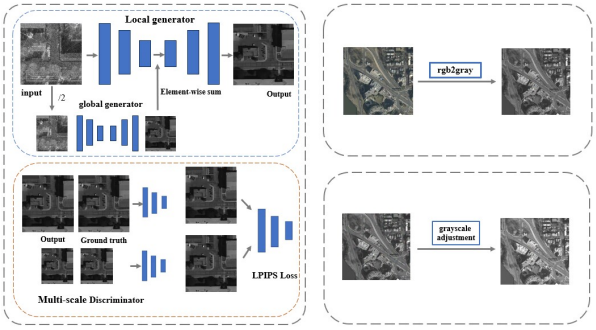


Figure 6. NJUST-KMG's solution comprises three modules: Pix2PixHD model training module, RGB to grayscale image conversion module, and grayscale adjustment module.

5.1. Rank 1: NJUST-KMG

NJUST-KMG developed an innovative method to transform RGB and SAR images into IR images by first converting RGB images to grayscale and then applying an intensity adjustment. This significantly improved the IR image quality.

For the SAR2EO, SAR2RGB, and SAR2IR tasks, they utilized the Pix2PixHD model for both training and testing. Given the SAR2EO dataset's size, they adjusted the learning rate towards the end of the training phase to ensure comprehensive training. For SAR2RGB and SAR2IR, due to their smaller dataset sizes, they employed the Pix2PixHD model pre-trained on the SAR2EO dataset, conducting further training on SAR2RGB and SAR2IR images. This strategy allowed for the generation of more detailed images, enhancing task performance (see model design in Fig. 6).

Furthermore, to boost the Pix2PixHD model's effectiveness on the competition dataset, they integrated the L2 norm and LPIPS metrics as loss functions during training. This adaptation significantly improved the model's performance, as reflected in the final evaluation scores.

For efficiency in training across all tasks, NJUST-KMG standardized the image size to 512×512 pixels. This resizing facilitates faster training without compromising the integrity of the data being processed. During testing, images for the SAR2EO, SAR2RGB, and SAR2IR tasks were also resized to 512×512 pixels to maintain consistency with the training phase. However, for the RGB2IR task, images were resized to 1024×1024 pixels to capture more detailed information. The gains of the Pix2PixHD model with various methods are compared in Table 4.

5.2. Rank 2: USTC-IAT-United

Team USTC-IAT-United proposes a unique pipeline termed: MvAV-pix2pixHD. They employed strategic methodologies across different tasks in the competition, leveraging advanced image translation techniques to

Model	Combined	SAR2EO	SAR2RGB	RGB2IR	SAR2IR
Pix2PixHD	0.44	0.08	0.56	0.55	0.58
Grayscale	/	/	/	0.16	0.53
Pretrained	/	/	0.55	/	0.55
All Combined	0.32	0.08	0.55	0.16	0.51

Table 4. **Pix2PixHD** denotes the scores on the test set for the Pix2PixHD model with modified loss. **Grayscale** represents the scores after two-stage grayscale adjustments. **Pretrained** indicates the use of weights from the Pix2PixHD model trained on the SAR2EO dataset. **All Combined** signifies the scores obtained by combining these methods.

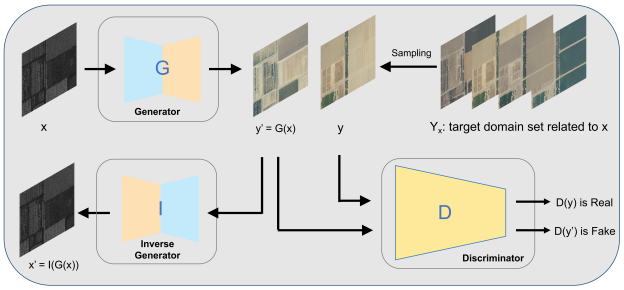


Figure 7. The MvAV-pix2pixHD method, designed for multi-view image translation tasks, features a generator (G) and discriminator (D) based on the coarse-to-fine and multi-scale principles of pix2pixHD, alongside an inverse generator (I) mirroring G's structure but with independent training parameters. Using SAR2RGB as an example, SAR images are transformed into the RGB.

achieve notable success.

- SAR2EO Task: The team utilized their previous year's championship-winning solution, pix2pixHD, capitalizing on its proven capabilities for the SAR to EO translation.
- SAR2IR and SAR2RGB Tasks: For these MAGIC datasets, the team introduced the MvAV-pix2pixHD method. This innovative approach enhances the realism of generated images by incorporating an inverse generator along with three robust loss functions, demonstrating the team's commitment to pushing the boundaries of image synthesis quality.
- RGB2IR Task: An in-depth analysis of IR training images led to the adoption of grayscale mapping and luminance extraction techniques. By applying grayscale mapping and adjusting the brightness of RGB images, the team was able to closely mimic the characteristics of IR images, showcasing their analytical approach to overcoming the challenges of this specific task.

5.2.1 MvAV-pix2pixHD

The MvAV-pix2pixHD approach, built upon the foundational strengths of pix2pixHD [23], is tailored for high-resolution multi-view aerial image translation. Pix2pixHD has been validated by Wang et al. [23] for its capability to surpass traditional encoder-decoder and Unet architectures in handling 512×512 and 1024×1024 resolutions. Its discriminators also outperform standard single-discriminator setups. Additionally, the efficacy of pix2pixHD in aerial view image translation, particularly for the SAR2EO task, was further confirmed by Yu et al. [24], showcasing superior performance over the original pix2pix [11] framework. Given these advancements, team USTC-IAT-United proposes the MvAV-pix2pixHD method, an adaptation of pix2pixHD for enhanced multi-view aerial view image translation, as depicted in Fig. 7.

Contrasting with unpaired image datasets, multi-view unpaired datasets necessitate a more deliberate sampling strategy. Instead of random sampling, their method selects target images from the dataset that correspond to the input based on location, sensor type, viewing angle, and time, ensuring a match within the target domain.

Beyond traditional random sampling, they introduce a time proximity sampling technique. This method leverages the temporal data recorded by different sensors during image acquisition, allowing for the pairing of source and target domain images based on their acquisition times. This approach acknowledges the inherent discrepancies caused by sensor heterogeneity and time variances, even within images captured at identical spatial locations.

5.2.2 Pipeline of MvAV-pix2pixHD

The pipeline of USTC-IAT-United's proposed MvAV-pix2pixHD framework applied to the task of multi-view image translation is shown in Fig. 7. In addition to the common generator G and discriminator D, they introduce an inverse generator I. Given the stochastic nature of the sampling process in the MAGIC dataset, which does not involve one-to-one pairing, relying solely on the original adversarial loss does not provide sufficient assurance that the model can effectively map a single input x to the desired output y. To further narrow down the space of possible mapping functions, they 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., $x \rightarrow G(x) \rightarrow I(G(x)) \approx x$. This also shows that the generated $G(x)$ has enough information to be reduced to x, guaranteeing the basic semantic informa-

tion of the original image. In order to achieve the aforementioned process, they introduced three loss functions detailed below.

5.2.3 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., $x \rightarrow G(x) \rightarrow I(G(x)) \approx x$. They call this forward inverse consistency, and incentive this behavior using a consistency loss:

$$\mathcal{L}_{\text{con}}(G, I) = \mathbb{E}_x[\|I(G(x)) - x\|_1] \tag{3}$$

Identity loss. For translation tasks in aerial view image scenarios, the dataset magnitude tends to be small due to the difficulty of data collection. They adapt the technique of Taigman et al. [20] 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., $\mathcal{L}_{identity}(G, I) = \mathbb{E}_y(y)[\|G(y) - y\|_1] + \mathbb{E}_x(x)[\|I(x) - x\|_1]$.

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 V. 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:

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

5.3. Rank 3: wangzhiyu918

Team wangzhiyu918 improves the baseline model (i.e., Pix2Pix) from three aspects:

- **Model Architecture**: They observe that although the baseline model performs well in the SAR-to-EO task, it performs poorly in the other three tasks (i.e., SAR-to-IR, SAR-to-RGB, and RGB-to-IR). This is primarily due to the variations in image resolutions across the four tasks. The image resolution for SAR-to-EO is 256x256, while the resolutions for the other three tasks are 1024x1024. Pix2pix does not perform well for high-resolution image translation. Therefore, they utilize the pix2pix model for the SAR-to-EO task due to its efficient training (requiring only a single 4090 GPU for one day of training), while employing the pix2pixHD model for the other three high-resolution image conversion tasks. Pix2pixHD proposes coarse-to-fine generator, multi-scale discriminators and improved adversarial loss to generate more realistic high resolution images. They discover that in the three high-resolution image translation tasks, the performance of pix2pixHD significantly surpasses that of pix2pix.
- **Training Data**: The training dataset includes images taken at a different time, thus some images are not temporally aligned, which increases the difficulty of image translation. They carefully selected training data pairs and manually verified them, discarding some that were not aligned. Moreover, since variability in the sensors, the resolutions vary widely for each image. For training, they uniformly resize each image to 1024x1024 to maintain consistency with the inference stage.
- **Training Strategies**: They use a linearly decaying learning rate. The images are normalized from [-1, 1]. They train their models with vanilla L1-Norm loss, Binary Cross Entropy (BCE) classification loss, and Learned Perceptual Image Patch Similarity (LPIPS) loss. The addition of LPIPS improves the performance.

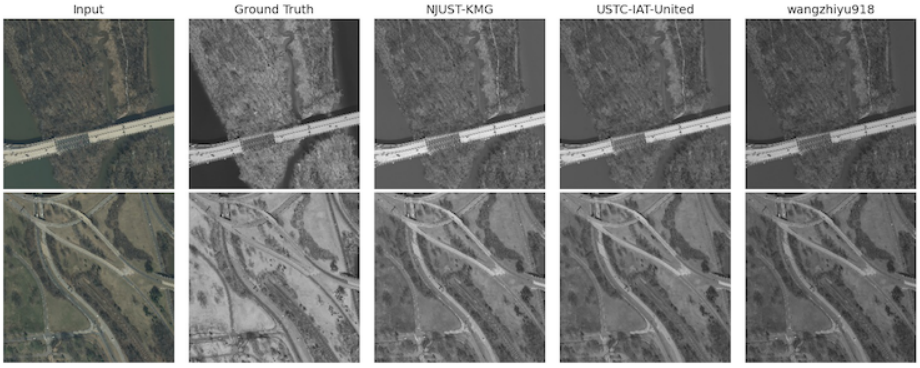
Additionally, they have observed that for the RGB-to-IR translation task, simply transforming RGB images to grayscale using OpenCV yields superior outcomes compared to training neural networks. This could be attributed to the limited availability of training data for the RGB-to-IR translation task.

6. Conclusion

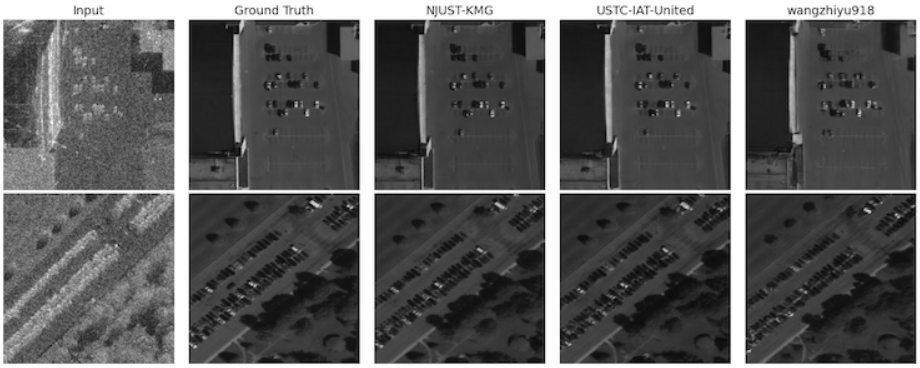
The 2024 MAVIC-T challenge presents the enhanced image-to-image MAGIC dataset, a comprehensive exploration of multi-modal image translation methodologies. This year's advancements in the challenge underscore the field's progress and open avenues for deeper analysis, presenting promising opportunities for future enhancements and functionalities.

Acknowledgements

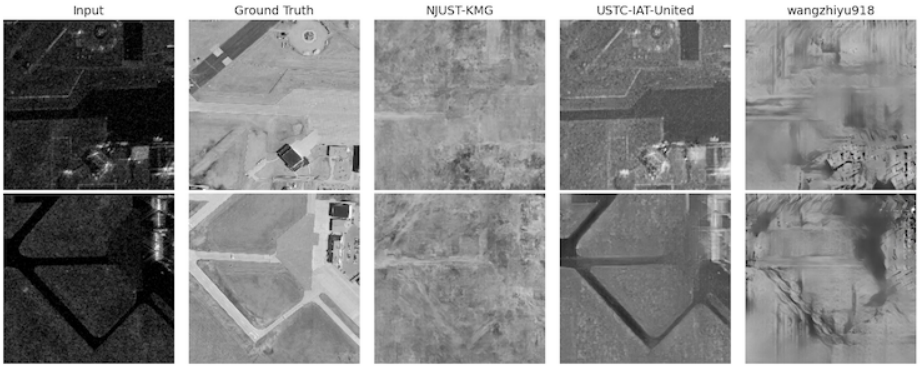
High Resolution Orthoimagery is available from the U.S. Geological Survey. See USGS Visual Identity System Guidance for further details. Funding for the project was provided through AFRL. The views and conclusions contained herein are those of the author and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Author Universities, Industries, or the U.S. Government.



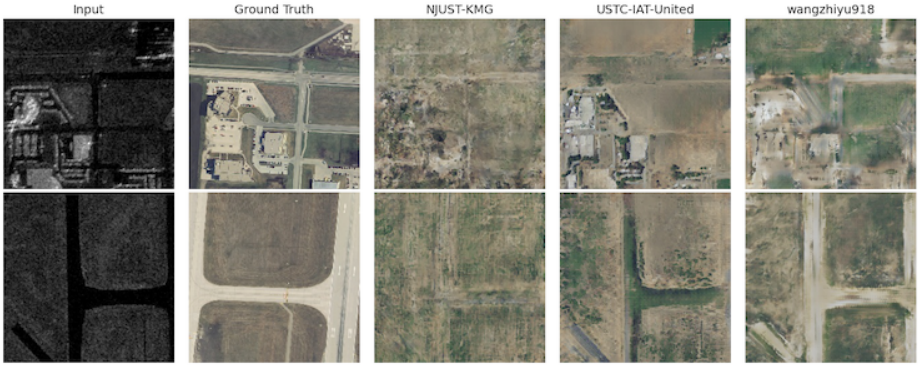
(a) RGB to IR



(b) SAR to EO



(c) SAR to IR



(d) SAR to RGB

Figure 8. Comparison of the four translation task for the top three teams. The input and ground truth are given as well as the generated outputs for each of the to three teams.

References

[1] Eros archive - aerial photography - high resolution orthoimagery (hro) — u.s. geological survey. 3

[2] Synthetic aperture radar (sar) open data - registry of open data on aws. 3

[3] Zezhou Cheng, Qingxiong Yang, and Bin Sheng. Deep colorization. In *Proceedings of the IEEE international conference on computer vision*, pages 415–423, 2015. 2

[4] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 3

[5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a nash equilibrium. *CoRR*, abs/1706.08500, 2017. 3

[6] Karl B Hille. Nasa developing ai to steer using landmarks – on the moon, 2022. 2

[7] Nathan Inkawhich. A global model approach to robust few-shot sar automatic target recognition. *IEEE Geoscience and Remote Sensing Letters*, 20:1–5, 2023. 2

[8] Nathan Inkawhich, Eric K. Davis, Matthew Inkawhich, Uttam K. Majumder, and Yiran Chen. Training sar-atr models for reliable operation in open-world environments. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:3954–3966, 2021. 2

[9] Nathan Inkawhich, Matthew J. Inkawhich, Eric K. Davis, Uttam K. Majumder, Erin Tripp, Chris Capraro, and Yiran Chen. Bridging a gap in sar-atr: Training on fully synthetic and testing on measured data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14: 2942–2955, 2021. 2

[10] 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

[11] 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. 6

[12] Shuo Liu, Vijay John, Erik Blasch, Zheng Liu, and Ying Huang. Ir2vi: Enhanced night environmental perception by unsupervised thermal image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1234–12347, 2018. 2

[13] Spencer Low, Oliver Nina, Angel D. Sappa, Erik Blasch, and Nathan Inkawhich. Multi-modal aerial view object classification challenge results - pbvs 2022. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 349–357, 2022. 1

[14] Spencer Low, Oliver Nina, Angel D. Sappa, Erik Blasch, and Nathan Inkawhich. Multi-modal aerial view image challenge: Translation from synthetic aperture radar to electro-optical domain results - pbvs 2023. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 515–523, 2023. 3

[15] Uttam K. Majumder, Erik P. Blasch, and David A. Garren. Deep learning for radar and communications automatic target recognition. *Artech House*, 2020. 2

[16] Miguel Oliveira, Angel Domingo Sappa, and Vitor Santos. A probabilistic approach for color correction in image mosaicking applications. *IEEE Transactions on image Processing*, 24(2):508–523, 2014. 2

[17] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 4

[18] Patricia L Suárez, Angel D Sappa, Boris X Vintimilla, and Riad I Hammoud. Near infrared imagery colorization. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2237–2241. IEEE, 2018. 2

[19] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015. 4

[20] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016. 7

[21] Harrish Thasarathan and Mehran Ebrahimi. Artist-guided semiautomatic animation colorization. *CoRR*, abs/2006.13717, 2020. 3

[22] Donald Venable. Improving real-world performance of vision aided navigation in a flight environment. In *Air Force Institute of Technology - Theses and Dissertations*, 2016. 2

[23] 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. 6

[24] Jun Yu, Shenshen Du, Renjie Lu, Pengwei Li, Guochen Xie, Zhongpeng Cai, Keda Lu, Qing Ling, Cong Wang, Luyu Qiu, et al. Sar2eo: A high-resolution image translation framework with denoising enhancement. *arXiv preprint arXiv:2304.04760*, 2023. 6

[25] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 649–666. Springer, 2016. 2

[26] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 3

[27] Y. Zheng, E. Blasch, and Z. Liu. *Multispectral Image Fusion and Colorization*. SPIE Press, 2018. 3

[28] 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