3rd Multi-modal Aerial View Image Challenge: Sensor Domain Translation - PBVS 2025

Dylan Bowald[1]*, Justice Wheelwright[1]*, Oliver Nina[1], Angel Sappa[2,3],
Riad Hammoud[4], Erik Blasch[1], Nathan Inkawhich[1]
[1] Air Force Research Laboratory, USA
[2] ESPOL Polytechnic University, Ecuador
[3] Computer Vision Center, Spain
[4] PlusAI Inc., USA

Abstract

This paper highlights the objectives, metrics and top performers in the 3rd Multi/Cross Modal Aerial Imagery Translation Challenge (MAVIC-T) of the 21st CVPR PBVS workshop.The core goal of this competition remains the seeding of innovative model development for translating registered aerial images between diverse sensor modalities. Specifically, the challenge explores the transformation between synthetic aperture radar (SAR), electro-optical (EO), visible light (RGB), and infrared (IR) imagery in challenging real world conditions. The competition once again judges its entrants using a composite of the L1-norm, Learned Perceptual Image Patch Similarity (LPIPS), and the Fréchet Inception Distance (FID). An additional penalty for overfitting to one domain ups the challenge from 2024, while pushing for more generalizable solutions on the second year return of the Multi Modal Aerial Gathered Image Composite Stacks (MAGIC-STACKS) Dataset. Overall, this year saw 103 total participants. The top performers scored comparably overall to last year's winners, however, there was a notable improvement this year in the RGB→IR translation task. Interestingly, the winning team - up6 was not the best across all translation scenarios, signaling room for improvement in coming years.

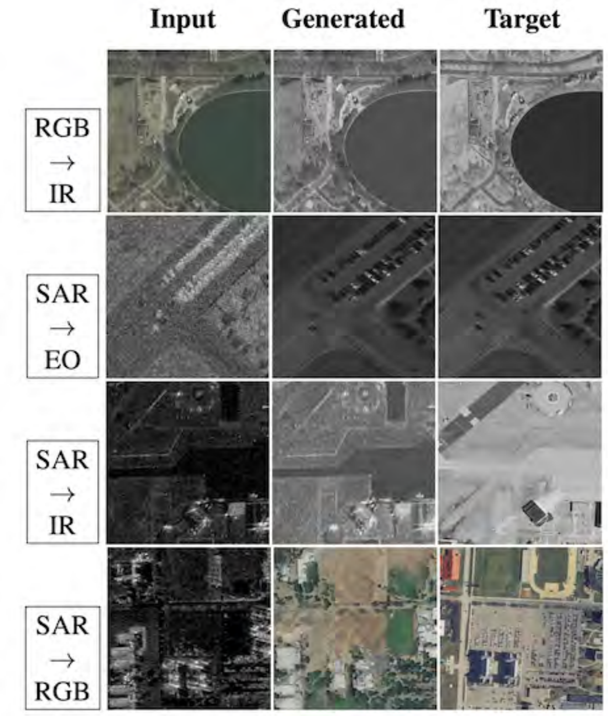


Figure 1. Examples of the four image translation tasks in the MAVIC-T competition. Input images are shown on the left, generated images in the middle, target images on the right. The goal of this competition is to make the generated image as close as possible to the target images.

1. Introduction

Data collected across multiple sensing modalities—such as Electro-Optical (EO), Infrared (IR), and Synthetic Aperture Radar (SAR)—each capture distinct characteristics and measurements of objects, providing complementary perspectives on the same scene. However, data coverage limitations and modality-specific constraints often hinder the full utilization of these diverse sources. Domain translation seeks to bridge this gap by converting one modality into another, enabling cross-modal data synthesis and enhancing sensor interoperability. The 2025 Multi-modal Aerial View Image Challenge - Translation (MAVIC-T), held in

* equal contribution

conjunction with the Perception Beyond the Visible Spectrum (PBVS) workshop, focuses on advancing multi-modal image translation techniques. Participants are challenged to develop models capable of accurately translating between EO, IR, and SAR imagery using a paired dataset, with performance evaluated based on L1, LPIPS, and FID scores. This challenge emphasizes producing high-fidelity translations while minimizing artifacts and hallucinations, pushing the boundaries of cross-modality data synthesis.

The translation of data across diverse sensor modalities presents a significant avenue for leveraging the complementary strengths inherent in each sensing technology, mitigating issues related to data coverage deficiencies. The MAVIC-T is designed to facilitate the advancement of sensor data utility through modality conversion, promote data diversity across varying modalities and geographical regions, and serve as a benchmark platform for the progression of cross modality conditioned image synthesis techniques. By emphasizing inter-modal data translation, MAVIC-T seeks to expand the frontiers of multi-modal research, fostering the development of more adaptable and robust analytical models. By leveraging nascent machine learning techniques for multi-modal translation among heterogeneous sensor sources, we steadily progress towards solving the inherent limitations associated with sensor-specific constraints.

The inaugural 2023 Multi-modal Aerial View Image Challenge - Translation [13] challenge was focused on the translation of Synthetic Aperture Radar and Electro-Optical modalities, using the UNICORN aerial SAR and EO dataset. In 2024 [14], we utilized satellite imagery to augment the UNICORNS data and create the MAGIC-STACKS dataset. This allowed us to get a much higher volume of aligned data, including data that spanned across time, geographic area, and environmental conditions. This year, the dataset has largely remained the same, but the challenge has been tailored using the scoring to better encourage more generalizable solutions.

Following from last year's competition, this year focuses on four main translation tasks: RGB→IR; SAR→EO; SAR→IR; and SAR→RGB; an illustration is shown in Fig. 1.

This challenge is predicated on exploiting the complementary strengths of diverse sensor modalities to address inherent data availability limitations. Synthetic Aperture Radar (SAR) sensors, renowned for their all-weather operation and atmospheric penetration capabilities, provide distinct advantages over Electro-Optical (EO) sensors. However, SAR imagery is characterized by interpretive complexity and limited availability. Similarly, Infrared (IR) imagery, crucial for thermal analysis and nocturnal operations, faces availability constraints, albeit to a lesser extent than SAR [12]. Given the relative abundance of EO data, translating EO imagery to IR presents a viable strategy for enhancing data diversity and addressing coverage deficiencies for such operations. In short, this effort seeks to alleviate the scarcity of SAR and IR data in the context of a preponderance of EO imagery.

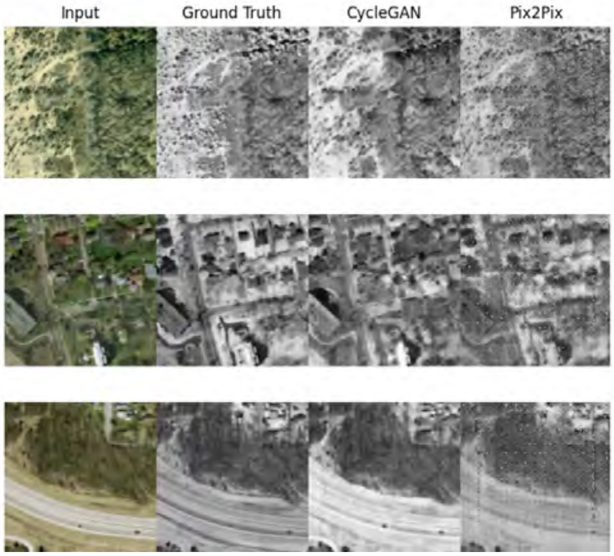


Figure 2. Example attempt of EO/IR translation using some out of the box deep learning paired image translation techniques.

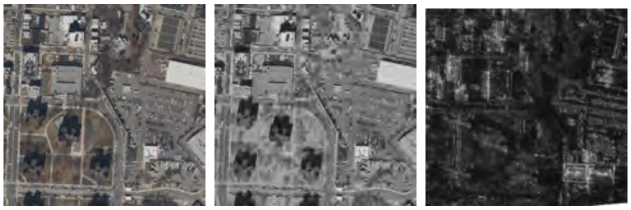


Figure 3. Example Stack (RGB, IR, SAR), from the MAGIC-STACKS dataset.

This translation is a key enabler of any task with limited modality data. One example area is automatic target recognition (ATR) tasks [9, 15]. The scarcity of labeled training data in SAR compared to EO [8, 10] makes EO-to-SAR translation a promising approach for enhancing SAR ATR model training by leveraging the abundance of labeled EO data. This principle extends to other translation tasks, as essential object features for training can be opportunistically captured across various sensor modalities.

This paper outlines the advancements enabled this year towards cross modal sensor translation facilitated by the 2025 MAVIC-T competition. We explore the winning solutions of this year's challenge and discuss their performance further in Section 5.

2. Previous Work

Historically, the general problem of image translation has been studied within the frameworks of paired and unpaired image translation, based on the availability of aligned images (i.e., geo-registered images) across source and target domains.

In the context of paired image translation, traditional methodologies, such as Pix2Pix [11], employ conditional Generative Adversarial Networks (GANs) [5] to facilitate the translation of aligned image pairs from a source domain to a target domain. Conversely, for unpaired image translation scenarios, conventional approaches, exemplified by CycleGAN [30], learn mappings between source and target domains in the absence of paired images [23]. Figure 2 shows examples of RGB to IR image translations produced by CycleGAN and Pix2Pix on the MAGIC 2024 dataset [14].

More recently, diffusion models have been proposed, demonstrating superior performance relative to contemporary generative models [4]. However, a primary limitation of diffusion models resides in the computational resources required to train a model. Furthermore, diffusion models tend to be susceptible to hallucinations or synthetically generated data not present in the ground truth. Hallucinations are particularly significant for sensor translation tasks, where accuracy of the model is paramount.

Other analogous cross-domain image translation problems for remote sensing applications and across diverse sensor modalities, include grayscale, near-infrared (NIR), thermal image colorization, and color transfer functions (e.g., [21], [16], [27], [3], etc.).

2024's Multi-modal Aerial View Image Challenge (MAVIC) for the Translation task demonstrated great improvements from previous baselines and promising results for the problem of electro-optical from (EO) to synthetic aperture radar (SAR) cross-domain translation [14]. MAVIC 2024 introduced the MAGIC dataset which is an extension of our previous translation dataset [13] that only contained EO and SAR data. MAVIC 2024 extended [13] by adding infrared (IR) sensor data to the collection, thus, being one of the first multi-modal datasets that provides 3 distinct and separate sensor channels, namely EO, SAR and IR.

This year's MAVIC challenge aims to continue the progress that has been achieved in this area of research in previous years with the purpose of improving the current state of the art methods and encourage innovation by our participants this year.

3. Challenge

Held in parallel with its sister competition the 2025 MAVIC-Classification challenge, the 2025 MAVIC-Translation contest is partnered with the Perception Beyond the Visible Spectrum (PBVS) workshop. Participants are evaluated on using a weighted average of the L1, LPIPS [28], and FID [6] score. Participants in this challenge are tasked with developing multi-modal image translation models using a provided dataset of paired EO, IR, and SAR imagery. The performance of these models, specifically their ability to accurately translate between modalities, is evaluated on an independent test set. The evaluation emphasizes the generation of high-quality translations, prioritizing the suppression of unexpected artifacts or hallucinations.

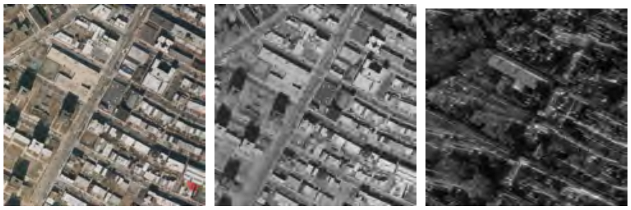


Figure 4. Sample (RGB, IR, SAR), from aligning UMBRA and HRO data.

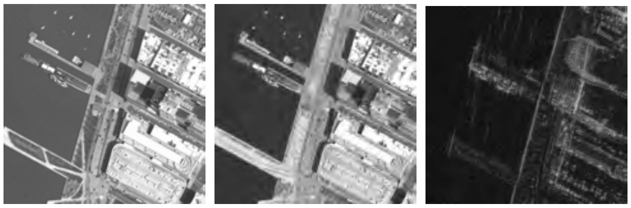


Figure 5. Sample (RGB, IR, SAR) from the Unicorns dataset.

3.1. Data

We reuse our MAGIC-STACKS data from last year [14]. This dataset incorporates diverse geospatial data sources for its analysis. The United States Geological Survey's (USGS) Earth Resources Observation and Science (EROS) program provides crucial land change imagery, including satellite and aerial data. Specifically, MAGIC leverages the EROS High Resolution Orthoimagery (HRO) dataset, which offers a vast collection of orthorectified aerial imagery with submeter resolution and consistent scale.

Hence, our dataset is a composite of three sources, processed and aligned to form uniform chipped stacks. The three source datasets used are UNICORN, USGS HRO [1], and the UMBRA [2] open data program. Note we have both an aerial source for SAR and a satellite source for SAR. Figure 5 illustrates a stack using our aerial source (Unicorns) while Figure 4 shows our satellite source (UMBRA).

As mentioned previously, our dataset incorporates Synthetic Aperture Radar (SAR) data from two different sources. The publicly accessible UNICORN dataset, a curated SAR-EO collection aligned using advanced homogra-

Date	Event
2025.01.19	Train data (input/output) and validation data (inputs only) released
2025.01.21	Validation server online
2025.02.21	Final test data release (inputs only)
2025.03.11	Test output results submission deadline
2025.03.13	Fact sheets and code/executable submission
2025.03.13	Preliminary test results released

Table 1. Challenge Phases.

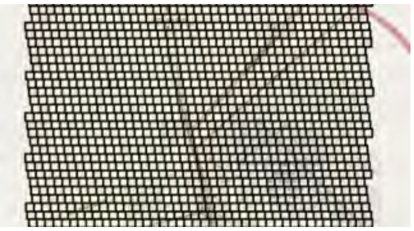


Figure 6. Example grid used to chip up the images.

phy techniques, contributes to this. Furthermore, UMBRA, a space technology company, provides high-resolution SAR imagery through its UMBRA open data program. This program grants access to over four million worth of free SAR data, offering up to 16cm resolution, further enhancing MAGIC's SAR capabilities.

A modular "plug-in" system prepares diverse image datasets for cross-dataset analysis. Each dataset undergoes preprocessing to ensure images are georectified, stored as GeoTIFFs with spatial referencing, and organized in a unified directory. Subsequently, a standardized chipping process divides the preprocessed data into spatially aligned 200m x 200m image stacks, indexed by their WGS-84 top-left coordinates. This process uses a grid anchored to a global origin, allowing for consistent and repeatable tiling across datasets, as illustrated in Figure 6.

We maintain our training, test, and validation split from last year. Our validation is sampled from New Albany, Ohio, and our test set is sampled over Washington DC. To ensure generalizability, the validation and test sets are not distributed, and sampled from different geographic locations.

3.2. Challenge Evaluation

The MAVIC-T challenge employs a rigorous evaluation framework designed to assess image translation submissions across four distinct tasks, aiming to achieve high-fidelity translations. The challenge focuses on the following translation tasks: SAR→EO, SAR→RGB, SAR→IR, and RGB→IR.

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

Table 2. The UNICORN dataset, employed exclusively in the SAR→EO translation task, is partitioned into training, validation, and testing sets.

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

Table 3. Details for he MAGIC dataset, used exclusively for the SAR→RGB, SAR→IR, and RGB→IR tasks.

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

1. **LPIPS (Learned Perceptual Image Patch Similarity):** This metric, based on the VGG-16 architecture [19], measures perceptual similarity using deep feature representations, aligning with human visual perception.
2. **FID (Fréchet Inception Distance):** Utilizing a pre-trained InceptionV3 network [22], FID quantifies the dissimilarity between the feature distributions of generated and target images, providing insights into visual fidelity and feature similarity.
3. **L1 Norm:** This metric calculates the pixel-wise absolute difference between target and generated images, ensuring structural integrity and content accuracy.

These metrics are chosen to comprehensively evaluate image translation quality, with LPIPS focusing on perceptual accuracy, FID on distributional similarity, and L1 on content and structural integrity. This strategy aims to minimize artifacts and ensure generated images exhibit high-resolution details and structural coherence, aligning with the target domain.

The evaluation process calculates each metric's score across all four tasks, followed by task-specific normalization to scale values between 0 and 1:

- **L1 Norm:** Pixel values are adjusted to fit within the desired range.
- **LPIPS:** Output weights are scaled for normalization.
- **FID:** A weighted $\arctan$ activation function is used to normalize scores, balancing each metric's influence.

The final score for each task is the average of these normalized metrics:

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

Rank	Team	Overall ↓	SAR→EO	SAR→RGB	RGB→IR	SAR→IR
1	**up6**	**0.33**	**0.08**	0.56	**0.13**	**0.53**
2	wangzhiyu918	0.35	0.11	**0.54**	0.23	**0.53**
3	rs6	0.37	**0.08**	0.58	0.27	0.55
4	xiaojie163	0.39	0.13	0.60	0.25	0.58
5	wsqmyself_1	0.47	0.25	0.60	0.47	0.54
6	donghongwei	0.55	0.09	0.57	1.0	0.55
7	HuangRoman	0.58	0.48	0.62	0.58	0.62
8	IOSB-VCA	0.59	0.35	0.68	0.64	0.68
9	kjhfkjn	0.61	0.25	0.60	1.0	0.57
10	egshklm	0.69	1.0	0.62	0.54	0.59
(2024) 1	**NJUST-KMG**	**0.32**	**0.08**	0.55	**0.16**	**0.51**
(2024) 2	USTC-IAT-United	0.33	0.10	**0.54**	0.17	0.52
(2024) 3	wangzhiyu918	0.36	0.11	**0.54**	0.22	0.55

Table 4. Top Performing Teams in MAVIC-T Competition (for all measures, lower numbers indicate better performance)

The overall submission performance is then determined by averaging the Task Scores across the four translation tasks:

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

Finally, a score penalty of 1 is added for each unattempted domain. This is done to encourage generalizability, without being too suppressing any models or techniques that may have exceptional results focusing on a smaller number of domains.

4. Challenge Results

We received 103 valid submissions to this year's MAVIC-T competition. Table 4 shows the results of the top 10 performers, along with results from the top-3 teams from last year's competition. Generated image samples from the top-3 teams in 2025 are shown in Figure 10 for qualitative inspection. Comparing results from the top-10 teams in 2025, we notice that there was a particularly large range in performance in the SAR→EO and RGB→IR tasks, while the range for SAR→RGB and SAR→IR is notably less. This variability can be attributed to differences in methodology:

SAR→EO translation remains challenging due to the inherent complexity of converting SAR to optical imagery. Top-performing teams such as up6 and wsqmyself 1 used advanced techniques, including diffusion models (E3Diff) and multi-stage training strategies, respectively, to improve translation quality. However, the performance range suggests that approaches varied significantly in effectiveness.

RGB→IR saw notable variation, with some teams (e.g., up6 and wangzhiyu918) achieving strong results using simple grayscale conversions rather than deep learning-based methods. This highlights the modality's inherent simplicity and the potential limitations of data-driven approaches when training data is scarce.

In contrast, the SAR→RGB and SAR→IR tasks exhibited more consistent performance across teams. The smaller performance range suggests that methods such as Pix2PixHD (used by wangzhiyu918 and wsqmyself 1) and CycleGAN (used by up6) provided stable and reliable results. The use of transfer learning in wsqmyself 1's SAR→RGB and SAR→IR models, leveraging knowledge from SAR→EO, may have contributed to this consistency.

Finally, comparing the top results from 2025 to 2024, we note that there was a slight decrease in overall performance by 0.01. Examining individual tasks, 2025 competitors improved in RGB→IR, remained stable in SAR→RGB and SAR→EO, and performed worse in SAR→IR. This decline in SAR→IR could be due to differences in dataset characteristics or methodological choices, such as the reliance on pre-trained models rather than specialized architectures for this task.

5. Methods

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

5.1. Rank 1: up6

The *up6* team develop different models for each task. For the SAR2EO task, they train a one-step conditional diffusion model (E3Diff) [18] to translate SAR images into EO images. The input consists of denoised SAR images and Canny edges. The model is trained in two stages. The first stage follows the standard 1000-step DDPM training [7, 17]. In the second stage, they employ a one-step DDIM [20] to convert DDPM into a one-step diffusion

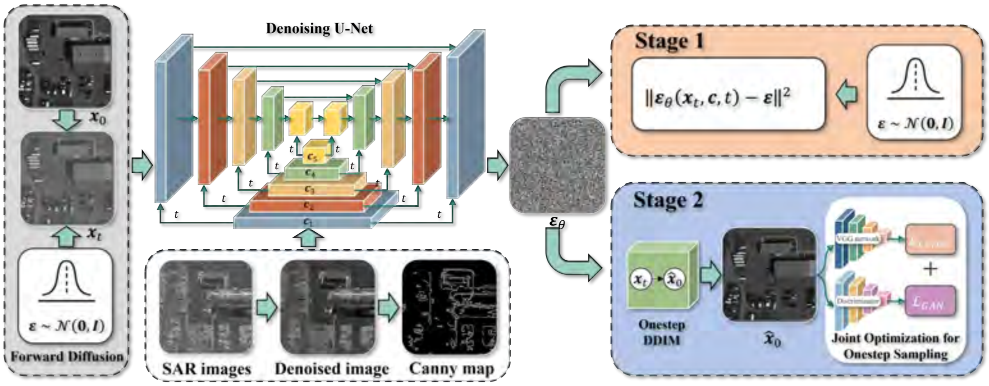


Figure 7. Architecture proposed by the *up6* team for SAR2EO translation.

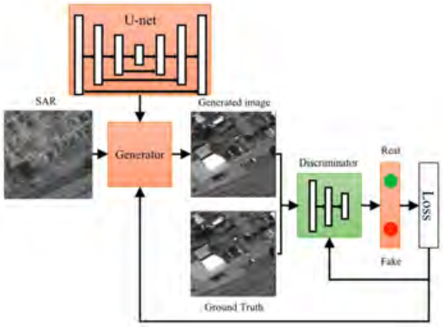


Figure 8. An illustration of Pix2PixHD network.

model, optimizing it with a combination of L1 loss, LPIPS loss [29], and GAN loss [18].

For the SAR2IR and SAR2RGB tasks, they use a vanilla CycleGAN model [30] and a Pix2PixHD model [24], respectively. During training, images are resized to 512×512. Each model is trained for 200 epochs, with a learning rate of 2e-4 for the first 100 epochs, which then linearly decays to zero over the remaining 100 epochs. In the testing phase, images are resized to 1024×1024 before being fed into the models. Figure 7 presents an illustration of the architecture.

For the RGB2IR task, considering the unique characteristics of IR and RGB images, they found that grayscale conversion outperforms other approaches. Hence, they first convert RGB images to grayscale and then enhance waterbody regions by darkening them based on pixel-value distribution analysis. All methods were implemented using Python and PyTorch on a system equipped with two A6000 48GB GPUs. The SAR2EO model requires approximately one week of training, while the SAR2IR and SAR2RGB models require less than a day.

5.2. Rank 2: wangzhiyu918

The *wangzhiyu918* team utilizes the Pix2PixHD model (see illustration in Fig. 8) for remote sensing image translation to effectively handle high-resolution image translation (1024×1024). Pix2PixHD enhances image quality by incorporating a coarse-to-fine generator, multi-scale discriminators, and an improved adversarial loss, enabling it to generate more realistic and visually coherent results compared to conventional methods. Given the nature of the dataset, which includes images captured at different times, temporal misalignment presents a challenge. To mitigate this, they carefully curate the training data by filtering out misaligned pairs and standardizing all images to a uniform resolution of 1024×1024, ensuring consistency during both training and inference. Training is conducted on an RTX 4090 GPU, with an efficient training pipeline that limits the maximum training time to half a day per task. The models are trained for 200 epochs with a batch size of 32, using an initial learning rate of 2e-4, which linearly decays after 100 epochs to facilitate stable convergence. To further refine the model's ability to capture fine-grained details, they employ a combination of L1 loss, Binary Cross-Entropy (BCE) loss, and Learned Perceptual Image Patch Similarity (LPIPS) loss. Notably, a significant performance boost is observed when incorporating LPIPS loss, improving translation quality from 0.15 to 0.11. An interesting observation arises in the RGB-to-IR translation task, where a simple conversion of RGB images to grayscale using OpenCV consistently outperforms neural network-based approaches. This suggests that the intrinsic relationship between RGB and IR modalities is straightforward enough that direct grayscale transformation is sufficient, likely due to the limited availability of training data for this specific task. These insights highlight the importance of selecting appropriate methodologies based on data characteristics and computational efficiency.

5.3. Rank 5: wsqmyself_1

The *wsqmyself_1* solution builds upon the Pix2PixHD framework [25], enhanced with tailored data augmentation and multi-stage training strategies to address multi-

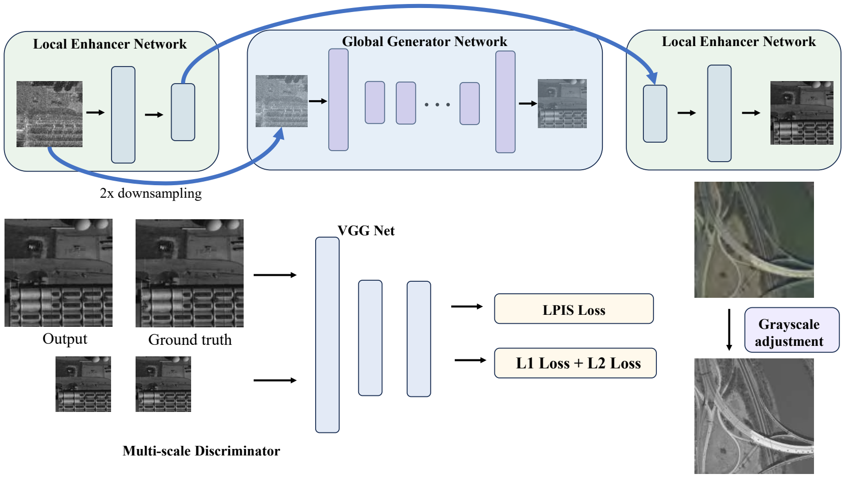


Figure 9. Architecture proposed by the wsqmyself_1 team.

modal aerial image translation tasks, including SAR2EO, SAR2RGB, SAR2IR and RGB2IR (see Fig. 9).

For the **SAR2EO task**, the proposed method expands the training dataset by stitching four adjacent SAR images into a single composite image, improving the model's capacity to capture fine-grained spatial details. The proposed method increased the dataset size fivefold, mitigating overfitting and enhancing generalization. The proposed method employed a three-stage training process: (1) 200 epochs with GAN loss and LPIS loss to establish structural coherence; (2) 50 epochs with L1 loss to refine pixel-level accuracy; and (3) 50 additional epochs with L2 loss and LPIPS loss to optimize texture and perceptual quality.

For the **RGB2IR task**, the proposed method designed a color-space transformation algorithm to convert RGB images to grayscale by calculating per-pixel color differences relative to black [26]. Specifically, the mean difference across RGB channels is computed for each pixel and scaled this grayscale array using an adaptive intensity factor to enhance contrast and detail visibility. The proposed method effectively bypassed the need for extensive training data while generating high-quality IR images efficiently.

For the **SAR2RGB and SAR2IR tasks**, the proposed method utilized a transfer learning strategy, initializing the model with weights pre-trained on the SAR2EO task. This initialization enabled the model to leverage prior knowledge from SAR-to-optical image translation, effectively compensating for limited dataset sizes. During training, input images across all tasks were standardized to a resolution of 512×512 pixels to balance computational efficiency and detail preservation. For RGB2IR inference, a resolution of 1024×1024 pixels was used to maximize output fidelity [14].

The proposed method emphasizes seamless integration of domain-specific adaptations. The staged training strategy, which progressively introduced loss functions, enabled the model to prioritize global structure before fine-tuning local details—critical for achieving high visual realism.

By combining data augmentation, adaptive loss functions, and task-specific preprocessing, the proposed method demonstrated competitive performance across modalities.

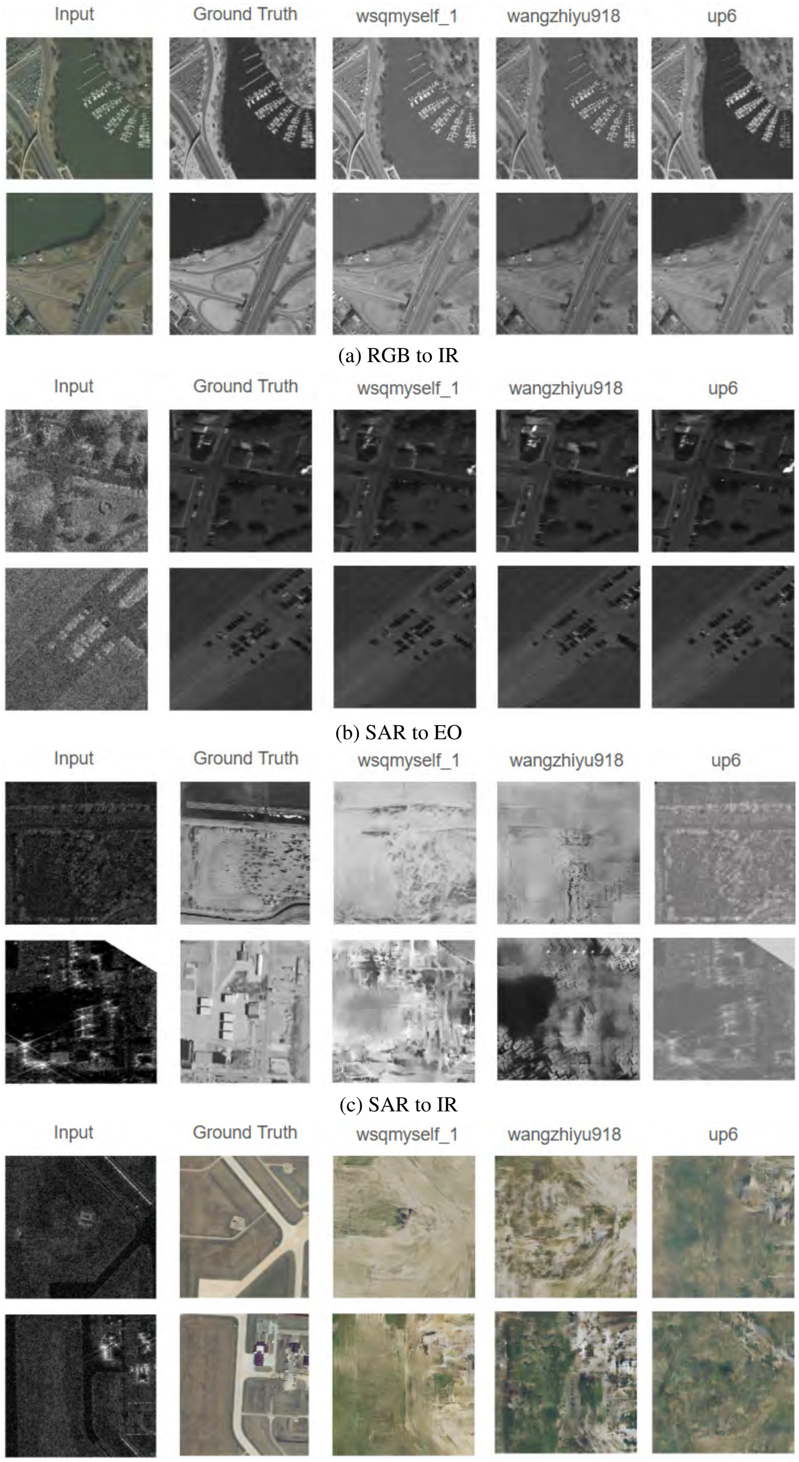
Implementation Details. All experiments were conducted on a server with 2 NVIDIA A100 GPUs (40GB VRAM), an Intel Xeon Platinum 8362 CPU, and 512GB system memory. The proposed method implemented the models in Python 3.8 using Pix2PixHD framework with 458.5 MB parameters. Training each task required approximately 24 hours to converge. The code and pre-trained models are available at https://github.com/wsqmyself/PBVS2025_translation.

6. Conclusion

The 2025 MAVIC-T challenge showcases the expanded MAGIC dataset, emphasizing the diverse range of methods employed in multi-modal image translation. While overall performance improvements from last year are incremental, the real value lies in the varied approaches participants have taken. From Pix2PixHD and CycleGAN to conditional diffusion models, the variety in methods reflects different strategies for addressing the challenge's complex tasks, such as SAR-to-EO, RGB-to-IR, and SAR-to-RGB translation. This diversity not only underscores the flexibility of the multi-modal image translation domain but also opens avenues for exploring new techniques, cross-method comparisons, and future refinements to the field.

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 con-



(a) RGB to IR

(b) SAR to EO

(c) SAR to IR

(d) SAR to RGB

Figure 10. 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.

tained 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.

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. 3

[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] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 3

[6] 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

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 5

[8] 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

[9] 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

[10] 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

[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. 3

[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 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. 2, 3

[14] Spencer Low, Oliver Nina, Dylan Bowald, Angel D Sappa, Nathan Inkawhich, and Peter Bruns. Multi-modal aerial view image challenge: Sensor domain translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3096–3104, 2024. 2, 3, 7

[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. 3

[17] Jiang Qin, Kai Wang, Bin Zou, Lamei Zhang, and Joost van de Weijer. Conditional diffusion model with spatial-frequency refinement for sar-to-optical image translation. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 5

[18] Jiang Qin, Bin Zou, Haolin Li, and Lamei Zhang. Efficient end-to-end diffusion model for one-step sar-to-optical translation. *IEEE Geoscience and Remote Sensing Letters*, 2024. 5, 6

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

[20] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*. 5

[21] 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. 3

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

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

[24] 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

[25] 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/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 6

[26] Xixian Wu, Dian Chao, and Yang Yang. High-resolution image translation model based on grayscale redefinition. *arXiv preprint arXiv:2403.17639*, 2024. 7

[27] 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. 3

[28] 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

[29] 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

[30] 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. 3, 6