Multi-modal Aerial View Image Challenge: Translation from Synthetic Aperture Radar to Electro-Optical Domain Results - PBVS 2023

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

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

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

Erik Blasch
Air Force Research Laboratory
Arlington, VA
erik.blasch.1@us.af.mil

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

Abstract

This paper unveils the discoveries and outcomes of the inaugural iteration of the Multi-modal Aerial View Image Challenge (MAVIC) aimed at 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. To accomplish the task of image translation, the competition utilizes images obtained from both synthetic aperture radar (SAR) and electro-optical (EO) sources. Specifically, the challenge centers on the translation from the SAR modality to the EO modality, an area of research that has garnered attention. The inaugural challenge demonstrates the feasibility of the task. The dataset utilized in this challenge is derived from the UNIfied COincident Optical and Radar for recognitioN (UNICORN) dataset. We introduce an new version of the UNICORN dataset that is focused on enabling the sensor translation task. Performance evaluation is conducted using a combination of measures to ensure high fidelity and high accuracy translations.

1. Introduction

The practice of sensor fusion, which involves integrating data from multiple sensors to provide a more comprehensive understanding of an environment or system, has been a longstanding concept [2, 10]. The earliest known example of sensor fusion dates back to the 1930s, when Robert Watson-Watt pioneered the first practical radar system that fused acoustic and radar data to facilitate advanced decision-making [6].

While the Multi-modal Aerial View Object Classification (MAVOC) challenge primarily deals with fusing data from the Synthetic Aperture Radar (SAR) and Electro-Optical (EO) modalities [13], the Multi-modal Aerial View Image Challenge (MAVIC) focuses on the conversion of data from one modality to another. A data conversion approach aims to leverage the unique advantages of both SAR and EO sensors while mitigating their limitations. Notably, SAR sensors are capable of operating in any lighting conditions due to their self-illuminating nature and can even penetrate through atmospheric obstructions like clouds and vegetation, which EO sensors are subject to. However, interpreting SAR images can be challenging and requires expert knowledge and specialized algorithms [14]. Additionally, the scarcity of publicly available SAR data presents an obstacle to creating robust deep learning training sets. Conversion of SAR images into EO images thus presents an opportunity to address these issues.

The problem addressed by the challenge shares some similarities with those approaches proposed in the literature for gray scale / near infrared (NIR) / thermal image col-

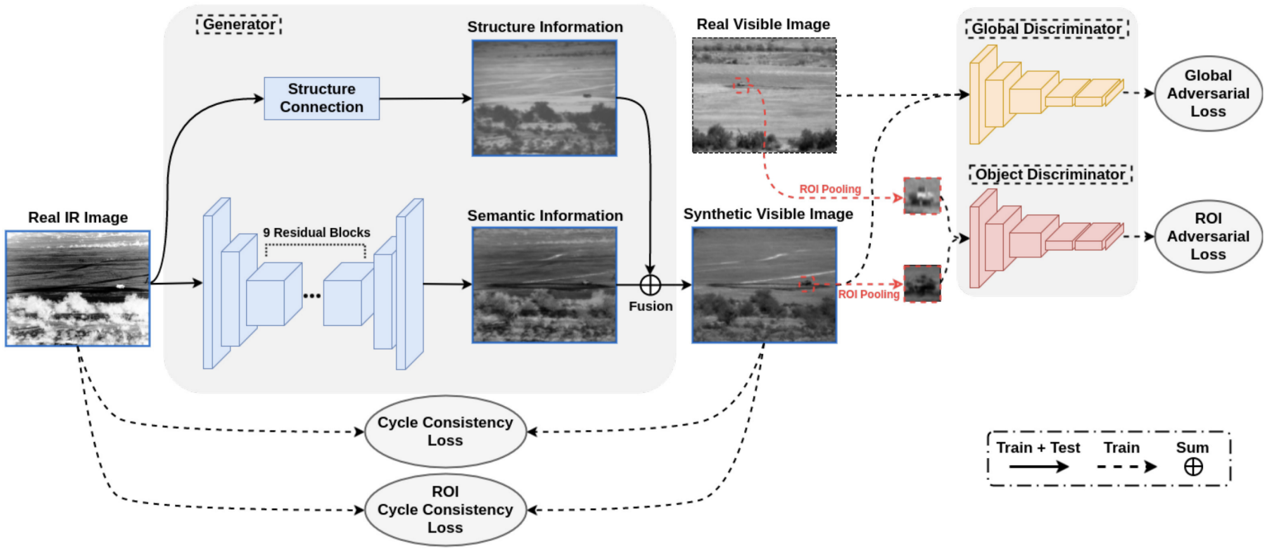


Figure 1. Overview of IR2VI architecture (illustration from [12]).

orization, or color transfer functions (e.g., [22], [16], [28], [3], just to mention a few). These problems are generally tackled through the use of Generative Adversarial Networks (GANs) [9], which allows the transformation of information between domains. Most GAN based approaches have focused on supervised contexts, where a paired 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) [32]. CycleGAN learns to map images from one domain (source domain) onto another domain (target domain) when paired images are unavailable [24]. This functionality 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 [5]. Unfortunately, the main limitation with these approaches 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.

The process of converting SAR images to EO images is not a trivial task, as it poses several challenges related to non-collocated sensor collections, pixel intensity association, image size, ground sampling distance (GSD), and image noise differences, as noted in [30]. Different approaches could be followed for converting SAR to EO modality, and although similarities exist with the methods mentioned above, the particularity of the problem warrants a deep study on the advantages and drawback of each of the many possibilities evaluated with quantitative and qualitative metrics.

The development of models that can translate between sensors of different modalities can enable the utilization of established algorithms. For example, in automatic target recognition (ATR) tasks, models are often trained on one modality and tested on other modalities. By translating SAR images to EO images, traditional EO models can be used on SAR data.

Image translation has been popular for image analysis covering iso-domain modalities (visual-to-visual), between-domain (infrared-to-visual) and across-domain (SAR-visual) imagery. The concept of image translation was popularized in 2017 as demonstrated by Isola et al. [9] for photo generation and object semantic segmentation over a variety of experiments using conditional generative adversarial networks (cGAN):

- Semantic labels ↔ photo
- Architectural labels → photo
- Map ↔ aerial photo
- Grayscale → color photos
- Edges → photo
- Sketch → photo
- Day → night

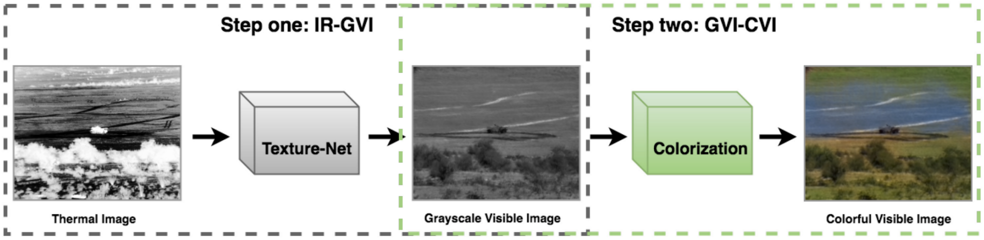


Figure 2. Example of colorization task (illustration from [11]).

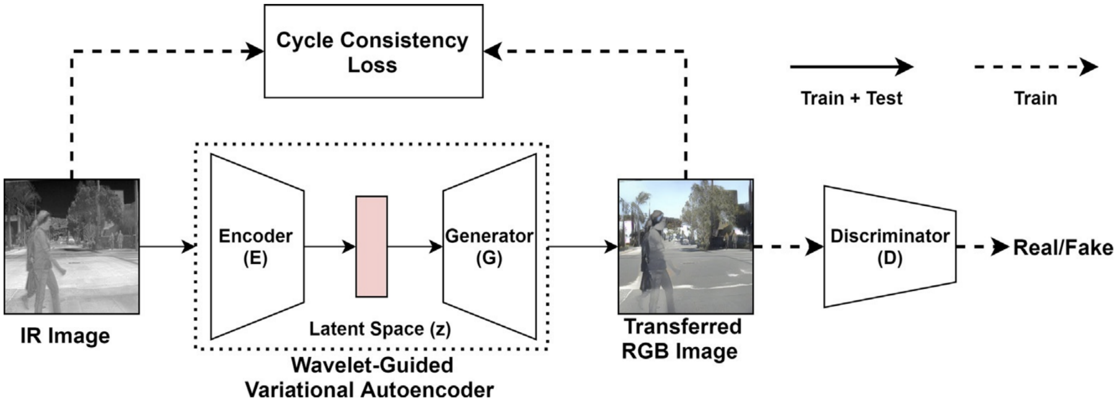


Figure 3. Overview of WGGAN architecture (illustration from [27]).

Many researchers extended theses types of methods on the same data sets such as a combination of the conditional variational autoencoder GAN and conditional latent regressor GAN called the BicyleGAN [31]. Additionally, using the same datasets, another popular method is the DualGAN that focuses on training over two sets of unlabeled images from two domains [26]. Building on these ideas included within-domain and cross-domain methods for seasonal changes such as colorization and object variation using the Multimodal Unsupervised Image-to-image Translation (MUNIT) framework [8]. Using only a single domain model, the STARGAN can do image transition to many different domains [4].

A common use of the image-to-image translation supports a method of transfer learning conducting domain adaptation. The goal is to extract domain agnostic features, then reconstruct a domain specific image with cycle consistency, and predict labels from the agnostic features to simultaneously learn from the source domain and adapt to the target domain [15]. Additional methods have been applied to medical imagery as MEDGAN [1]. Review papers highlight the different image translation methods developed in recent years [17].

Among these developments, there was consideration for image-to-image translation for EO and infrared data. Liu et al. [12] utilized the above methods to develop the IR2VI method and compared to the CycleGAN, DUALGAN, and STARGAN (Figure 1). Results demonstrate that the IR2VI adds semantic visible information and object shape information to the original thermal images while CycleGAN and UNIT were not able to do cross-modality image translation. The StarGAN translated images lack texture information, however, the blur shape information did support the VI object detector.

Following up with mage colorization using the No-reference Image Quality Evaluation metric, a texture net was shown superior [11] (Figure 2) and a WGGAN: A Wavelet-Guided Generative Adversarial Network for Thermal Image Translation [27] (Figure 3) further improved the analysis. Key to these approaches was the features used in the GAN to support multimodal image translation.

2. Challenge

The 2023 MAVIC challenge is held jointly with the Perception Beyond the Visible Spectrum (PBVS) workshop and is a complement to the MAVOC challenge. The MAVIC challenge is designed to facilitate innovate approaches in multi-modal sensor translation. Participants are evaluated on using a weighted average of the L2, LPIPS [29], and FID [7] score. The challenge centers on the advancement of multi-modality translation networks. Participating teams are provided with a collection of image pairs, consisting of SAR and EO modalities, and are tasked with performing image translation from one modality to the other. Upon completion, the teams' generated outputs are evaluated on

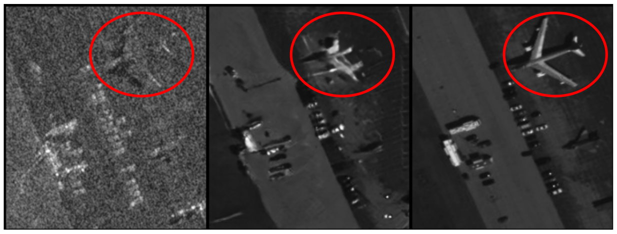


Figure 4. Example of failed translation from SAR to EO (order = SAR input, translation, ground truth). We draw attention to the aircraft in the image. The translated image illustrates an example of the generative network hallucinating.

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 4 illustrates an example of a failed translation that contains hallucinations.

The manuscript is organized as follows. Section 2 provides an introduction to the challenge dataset, evaluation metrics and competition phases. Section 3 summarizes the results obtained by different teams. Then, Section 4 presents a short description of the top approaches evaluated from submissions. The conclusion is presented in Section 5, followed by an appendix containing information on the teams.

2.1. Dataset

The present dataset is derived from the UNICORN dataset. Unlike the existing UNICORN dataset, which was annotated with respect to target chips, the current dataset is constructed by dividing co-aligned scenes into 256 × 256 patches. Figure 5 provides examples of these image pairs. These patches form image pairs that are used for training. The allocation of training and validation sets is performed according to the protocol presented in Table 1.

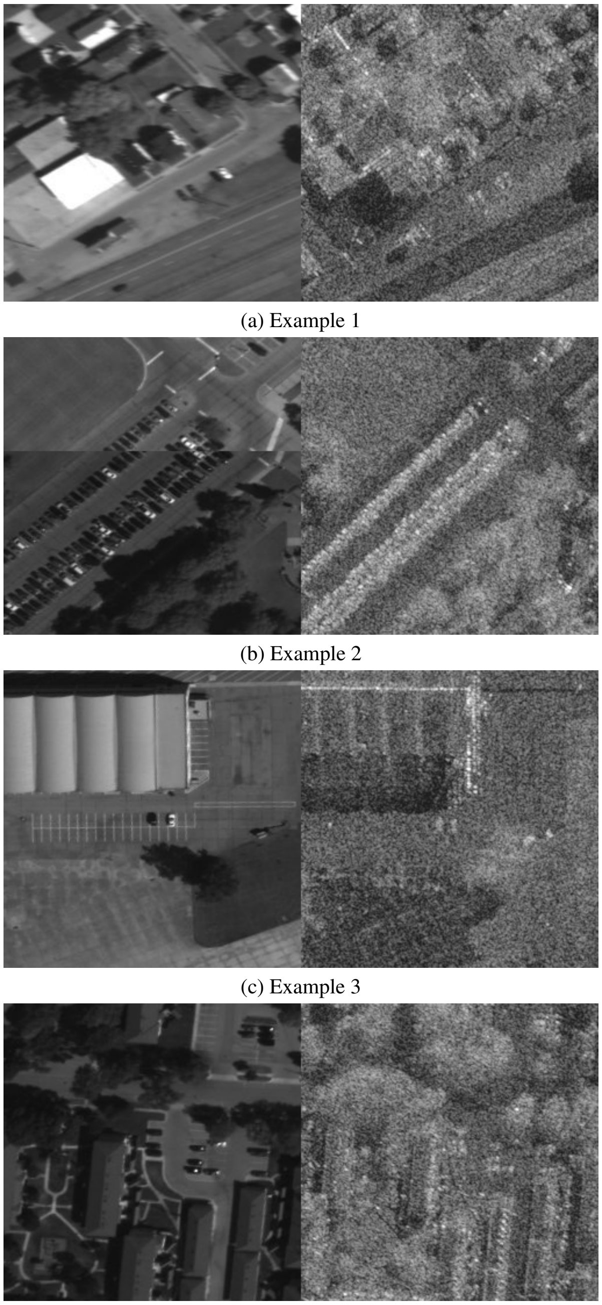
Table 1. Details of the UNICORN dataset used for training, validation, and testing.

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

2.2. Evaluation

The submissions are evaluated using an average of three metrics:

1. LPIPS: Learned Perceptual Image Patch Similarity
2. FID: Frechet Inception Distance
3. L2: Pixel-wise L2 norm



(a) Example 1

(b) Example 2

(c) Example 3

(d) Example 4

Figure 5. Examples of (SAR, EO) image pairs pulled from the training set. The EO image is on the left while the SAR image is on the right. These are 256 × 256 pixels. This shows examples of buildings, cars, and trees. These have been aligned and re-sampled to ensure consistent size.

The selection of the three evaluation metrics serves the purpose of promoting the production of high-fidelity translations. Specifically, the L2 metric is incorporated into the evaluation process to mitigate against the occurrence of generative hallucinations. By enforcing pixel-wise L2 norms, the L2 metric effectively preserves the overall structure of the target image. However, the L2 metric has the drawback of prioritizing low-frequency details while eliminating high-frequency ones. The combined utilization of the LPIPS and FID metrics complements the L2 metric by ensuring the presence of high-resolution details in the generated images. Additionally, these metrics function to ensure that the generated images are well-aligned with the target image domain, thus contributing to the overall accuracy and effectiveness of the translation process.

The LPIPS metric is constructed based on the VGG-16 architecture [21]. LPIPS is a technique that leverages deep feature representations of two images to compute a similarity metric that closely approximates human perceptual judgments. The FID metric, on the other hand, is established using a pre-trained InceptionV3 [23] network, and its computation employs the feature layer 64. The FID metric provides a measure of the dissimilarity between two probability distributions. When applied to an activation layer of a neural network, it enables an effective comparison of the feature spaces of two images. The FID is calculated as:

$$\text{FID} = |\mu - \mu_\omega| + \text{tr}\left(\Sigma + \Sigma_\omega - 2\left(\Sigma\Sigma_\omega\right)\right). \qquad (1)$$

where μ corresponds to the mean, "tr" is the trace, and Σ corresponds to the covariances. The subscript ω indicates a fake or generated image [20].

In order to consolidate the three metrics, a normalization procedure is executed to ensure that each metric is scaled between the range of 0 and 1. The normalization methodology differs among the metrics. In the case of the L2 norm, an image normalization technique is employed to restrict the values of the input images being compared between the range of 0 and 1. For the LPIPS metric, the normalization process necessitates the scaling of the output weights. Finally, the FID metric is normalized via the application of a weighted $\arctan$ activation function.

$$\text{Final Score} = \frac{\frac{2}{\pi}\arctan(\text{FID}) + \text{LPIPS} + \text{L2}}{3} \qquad (2)$$

2.3. Challenge Phases

The challenge began January 11, 2023, and the test data was released March 1, 2023. The testing phase ended on March 7, 2023 with team submissions finalized.

3. Challenge Results

The challenge results are summarized in this section. This challenge had 52 teams participate with 5 teams submitting results during the development phase and 8 teams submitting results in the testing phase, see Table 2. Resulting submissions from top three teams can be seen in Figure 8.

Table 2. Top Performing Teams in Competition

Rank	Team	Total ↓	LPIPS ↓	FID ↓	L2 ↓
1	**USTC-IAT-United**	**0.09**	**0.25**	**0.02**	**0.01**
2	pokemon	0.14	0.35	0.04	0.01
3	wangzhiyu918	0.14	0.38	0.02	0.01
4	ngthien	0.18	0.43	0.10	0.01
5	Wizard001	0.26	0.50	0.27	0.02
6	hanhai	0.30	0.30	0.59	0.01
7	u7355608	0.33	0.54	0.43	0.02
8	jsyoon	0.33	0.46	0.53	0.01

4. Methods

This section briefly summarizes the approaches used by the teams that submitted their models and documentation for prize consideration. Not all teams submitted their methods and are subsequently absent from this paper. We examine the submitted methods from the top teams. This section consists of edited summaries submitted by each team.

4.1. Rank 1: USTC-IAT-United

Team USTC-IAT-United proposed a SAR2EO framework, which is capable of converting SAR images into EO images. For the characteristics of SAR images, a SAR image pre-processing module is proposed to process SAR images. This module proves effective in experiments. They also compared pix2pix [9], pix2pixHD [25], SPADE [18], UNIT DDPM [19] and other methods, and finally, chose pix2pixHD as their base model through experiments. Figure 6 shows the flow chart of the proposed final solution for the competition.

In the training phase, they first pass the SAR images through a data pre-processing module, and then generate the low-precision output after going through the low-precision generative model model, and then discriminate the training data and the low-precision model through the discriminator. The low-precision image and training data are fed into the high-precision generator together to generate the high-precision output, which is the final result. To make the final result more high-definition, the output and training data continue to be supervised by the discriminator. Similarly the test set part goes through the same pre-processing module and goes through two low-precision and high-precision generators to get the final EO results. Finally their proposed SAR2EO framework achieved better results and won the first place in the competition.

On the whole, their proposed methods is made of four optimized parts to generate high-quality high-definition large images:

1. The generator is upgraded from U-Net to a multi-level generator (coarse-to-fine generator)

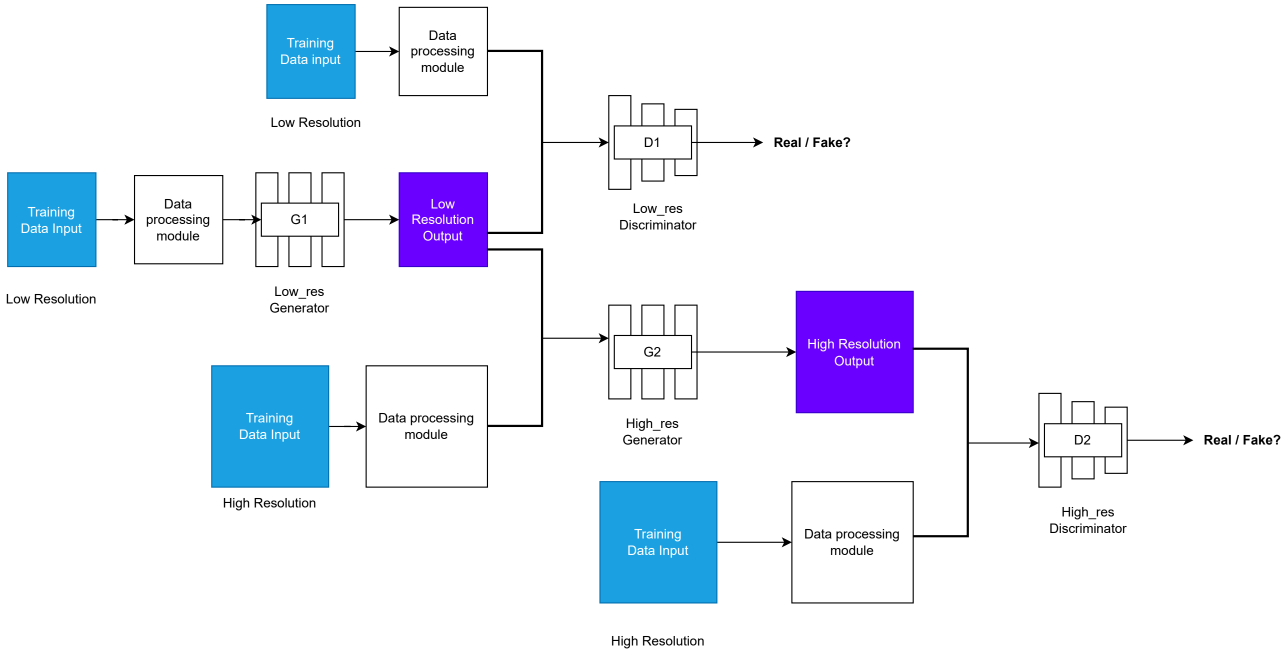


Figure 6. Overview of team USTC-IAT-United's model. It utilizes a four stage process to produce high quality translated images.

2. The discriminator is upgraded from patch GAN to a multi-scale discriminator (multi-scale discriminator)
3. The matching loss and content loss based on discriminator features are added to the optimization objective
4. The proposed data pre-processing module is effective for SAR images.

Based on the experimental results, the proposed scheme demonstrates the ability to generate images of comparatively high quality on the test set. Furthermore, its performance is favorably rated upon human visual inspection. Additionally, the USTC-IAT-United's image generation model exhibits a relatively fast processing speed.

4.2. Rank 3: wangzhiyu918

Team wangzhiyu918 developed a SAR-to-Optical image translation network based on the *pix2pix* [9], which is a milestone method for applying GAN to image-to-image translation. Figure 7 overviews their proposed method. They adopt a U-Net-based network as their generator, which takes 256 × 256 SAR images as inputs. The discriminator uses the patchGAN structure to distinguish whether an input image is real or fake. It can classify whether overlapping patches are real or fake. Such a patch-level discriminator architecture has fewer parameters than a full-image discriminator and can work on arbitrarily-sized images in a fully convolutional fashion. They train their model with vanilla L1-Norm loss and Binary Cross Entropy (BCE) classification loss.

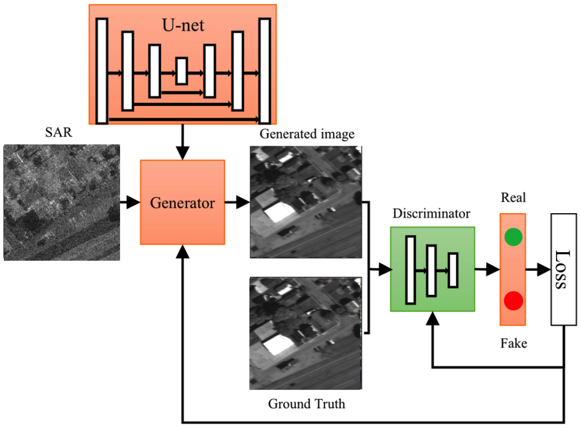


Figure 7. Illustration of the architecture proposed by the wangzhiyu91 team.

5. Conclusion

The 2023 edition of the MAVIC competition was organized with the goal of promoting research in sensor translation across multiple modalities. The inaugural competition successfully demonstrated the feasibility of devising effective translation models, with participating teams submitting methods that achieved notable success. The development of such models holds significant potential for facilitating the integration of SAR imagery into traditional EO algorithms, thereby enhancing their overall utility. Furthermore, these models have the potential to enable the accomplishment of a range of important tasks, including automatic target recognition (ATR), among others.

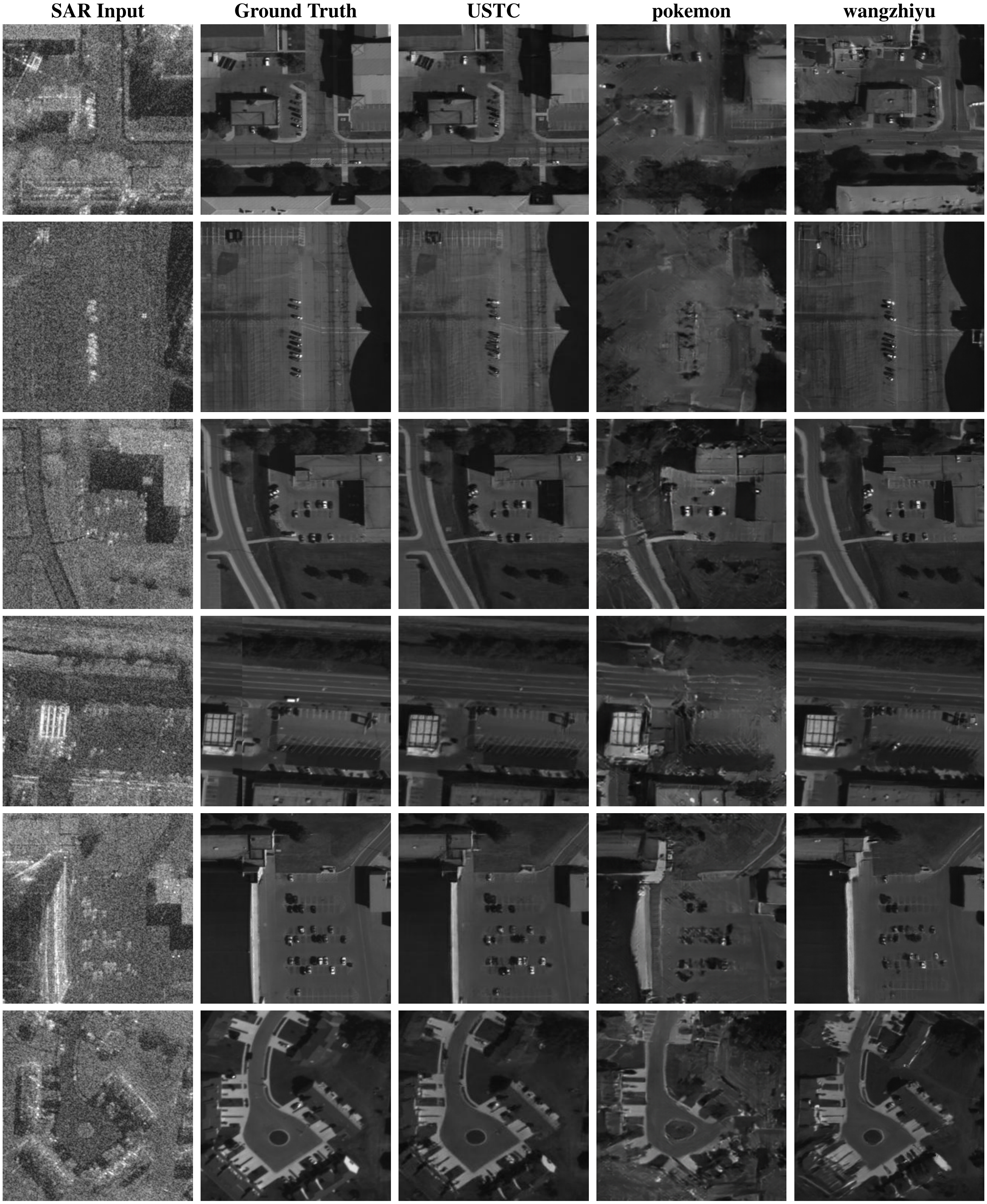


Figure 8. Comparison of model inputs, ground truth, and outputs from top-three teams.

Acknowledgements

We would like to thank Bob Lee for providing logistic support for the 2023 PBVS MAVOC Challenge. We would also like to thank Angel Wheelwright for helping run the competition.

References

[1] Karim Armanious, Chenming Jiang, Marc Fischer, Thomas Küstner, Tobias Hepp, Konstantin Nikolaou, Sergios Gatidis, and Bin Yang. Medgan: Medical image translation using gans. *Computerized Medical Imaging and Graphics*, 79, Jan. 2020. 3

[2] Erik Blasch, Eloi Bosse, and Dale A Lambert. *High-level Information Fusion Management and System Design*. Artech House, 2012. 1

[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] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *CoRR*, abs/1711.09020, 2017. 3

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 2

[6] Ethw. Robert watson-watt, Feb 2016. 1

[7] 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

[8] Xun Huang, Ming-Yu Liu, Serge J. Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. *CoRR*, abs/1804.04732, 2018. 3

[9] 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, 5, 6

[10] Martin II Liggins, David Hall, and James Llinas. *Handbook of Multisensor Data Fusion: theory and practice*. CRC Press, 2007. 1

[11] Shuo Liu, Mingliang Gao, Vijay John, Zheng Liu, and Erik Blasch. Deep learning thermal image translation for night vision perception. *ACM Transactions on Intelligent Systems and Technology*, 12:1–18, 12 2020. 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, 06 2018. 2, 3

[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] Uttam K. Majumder, Erik P. Blasch, and David A. Garren. Deep learning for radar and communications automatic target recognition. *Microwave Journal*, 63(10):118, 10 2020. Copyright - Copyright Horizon House Publications, Inc. Oct 2020; Last updated - 2022-10-19. 1

[15] Zak Murez, Soheil Kolouri, David J. Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. *CoRR*, abs/1712.00479, 2017. 3

[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] Yingxue Pang, Jianxin Lin, Tao Qin, and Zhibo Chen. Image-to-image translation: Methods and applications. *IEEE Transactions on Multimedia*, 24:3859–3881, 2022. 3

[18] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 5

[19] Hiroshi Sasaki, Chris G. Willcocks, and Toby P. Breckon. UNIT-DDPM: unpaired image translation with denoising diffusion probabilistic models. *CoRR*, abs/2104.05358, 2021. 5

[20] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, August 2020. Version 0.3.0. 5

[21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 5

[22] 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

[23] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015. 5

[24] Harrish Thasarathan and Mehran Ebrahimi. Artist-guided semiautomatic animation colorization. *CoRR*, abs/2006.13717, 2020. 2

[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 Conference on Computer Vision and Pattern Recognition*, 2018. 5

[26] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. *CoRR*, abs/1704.02510, 2017. 3

[27] Ran Zhang, Junchi Bin, Zheng Liu, and Erik Blasch. Wggan: A wavelet-guided generative adversarial network for thermal image translation. In *Generative Adversarial Networks for Image-to-Image Translation*, pages 313–327. Elsevier, 2021. 3

[28] 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

[29] 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

[30] Y. Zheng, E. Blasch, and Z. Liu. *Multispectral Image Fusion and Colorization*. Press Monographs. SPIE Press, 2018. 2

[31] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A. Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. *CoRR*, abs/1711.11586, 2017. 3

[32] 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

Appendix A. Teams Information

We acknowledge the participants. We used edited versions of team submissions for method explanations.

MAVIC 2023 organization team:

Members: Spencer Low, Dr. Oliver Nina, Dr. Angel Sappa, Dr. Nathan Inkawhich

Affiliation: BYU, AFRL/RY, AFRL/RI, ESPOL

USTC-IAT-United:

Members: Jun Yu, Shenshen Du, Renjie Lu, Pengwei Li, Guochen Xie, Zhongpeng Cai, Keda Lu, Qing Ling, Cong Wang, Luyu Qiu, Wei Zheng

Affiliation: University of Science and Technology of China, Huawei Technologies

wangzhiyu918:

Members: Zhiyu Wang, Xudong Kang, Shutao Li

Affiliation: College of Electrical and Information Engineering, School of Robotics, Hunan University