Evaluating Domain Translation Approaches for Drone-Based Geo-Localization in Adverse Weather

Yiqing Li Shuke He
Zhejiang Scientific Research Institute of Transport
Hangzhou China
Zhejiang Jiande Institute of General Aviation
Hangzhou China
liyiqing_cs@163.com,he_shuke@163.com

Chen Jin*
College of Civil Aviation
Nanjing University of Aeronautics and Astronautics
Nanjing China
Zhejiang Scientific Research Institute of Transport
Hangzhou China
jin.chen@buaa.edu.cn

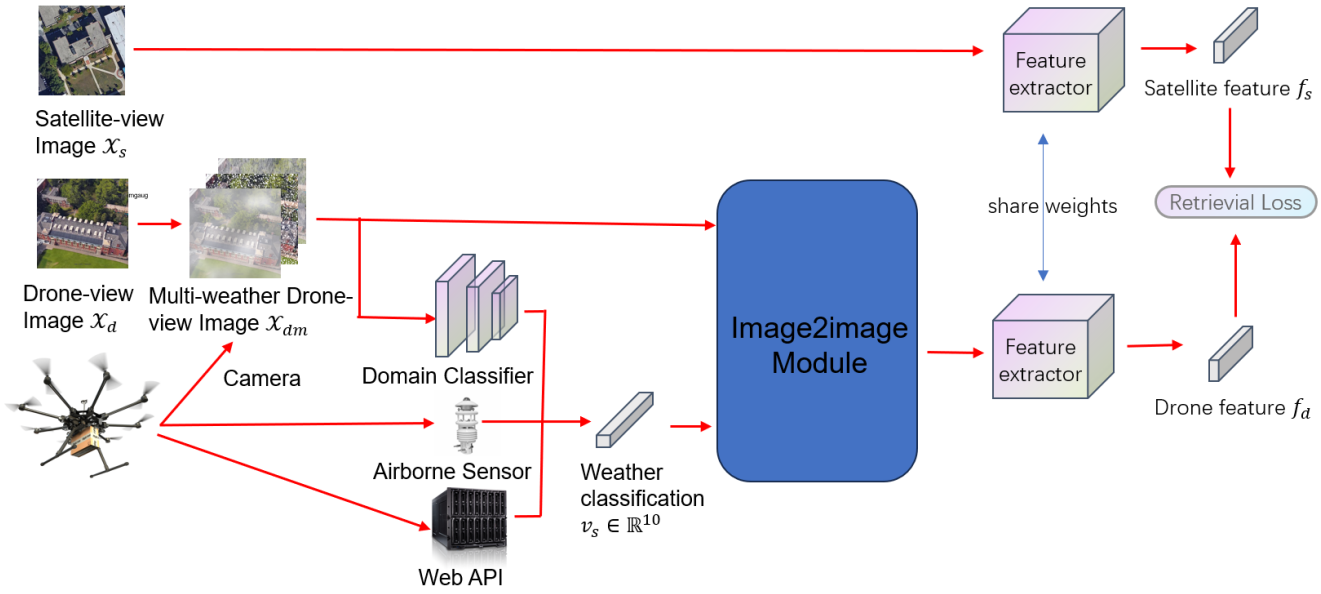

Figure 1: Overview of our proposed domain translation framework.

ABSTRACT

The UAVs in Multimedia (UAVM) 2024 competition aims to improve the performance of Drone-based Geo-localization task under extreme weather. On the task, we found the simple augmentation on training set can significant improve performance with zero-cost and won the 6th Place among 22 teams on this competition. Furthermore, based on our observations, we propose a domain translation framework to further enhance the performance of any model. We transfer images from a multi-weather domain to a normal domain using GAN and introduce FFM to further reduce computational costs. Our experiments revealed that while this strategy indeed improves image quality, it does not translate into improved recognition accuracy. Thus, while our strategy is supported by experimental evidence, practical methods for achieving performance improvement remain elusive. We hope this report provides valuable insights for developers and researchers in this field.

CCS CONCEPTS

• **Computing methodologies** → **Visual content-based indexing and retrieval**; **Image representations**.

KEYWORDS

Drone, Geo-localization, Domain Translation, Deep Learning, Image Retrieval

1 INTRODUCTION

Unmanned Aerial Vehicles (UAVs) fitted with cameras have found applications across various fields. High-precision positioning information is essential for these tasks, typically obtained through Real Time Kinematics (RTK) GPS technology. However, GPS-based positioning can suffer from high errors or even fail in certain challenging scenarios, such as navigating among tall buildings[1], passing through tunnels, or operating under bridges. In these instances, UAVs rely on camera-based visual positioning, also known as drone-based geo-localization, which aims to retrieve images of the same

geographic target from satellite image. This method becomes particularly critical in adverse weather conditions like rain or snow, where complex refractivity fields can disrupt GPS signals[6], necessitating the use of visual-based localization. However, such harsh weather conditions can also cause domain shifts that reduce the accuracy of visual positioning. Therefore, enhancing the generalization ability of drone-based geo-localization across diverse environmental scenarios is crucial.

To mitigate domain shift, a straightforward approach is to train directly on data from target domains. As demonstrated in the ACM MM 2024 competition, augmenting the University-1652 dataset significantly improved the performance of MLPN on the University-160k-wx dataset. Furthermore, we observed that models trained on augmented datasets perform better under normal weather conditions compared to adverse weather conditions such as rain or fog. This observation led us to consider whether a simple weather classification method combined with appropriate preprocessing could transform input images from adverse weather domains to normal weather domains, thereby improving model performance under harsh weather conditions.

In this technical report, we explore both frequency-domain and time-domain preprocessing operations. Contrary to our initial expectations, experimental results indicate that neither approach improved model performance. This finding suggests that straightforward preprocessing methods may not be sufficient to enhance performance under adverse weather conditions, highlighting the need for further research into image restoration techniques to improve recognition capabilities in challenging environments.

2 RELATED WORK

2.1 Geo-localization Methods

To address the limitations of traditional GPS-based localization methods, Geo-localization Methods have been developed as an image retrieval task. The key to these methods is to learn an image representation that is discriminative and robust to visual appearance changes due to different viewpoints[22]. These discriminative features can be obtained either through hand-crafted techniques or deep learning approaches. While earlier works [9, 10, 19] were restricted to considering only two viewpoints, Zhang et al.[22] developed a cross-view Geo-localization method that trains on data from three different viewpoints—street-view, drone-view, and satellite-view—resulting in more robust features and enhanced functionality for the task. To further enhance the performance of cross-view Geo-localization methods, one approach is to incorporate additional semantic information. For instance, Wang et al. [16] developed a hand-crafted pluggable module that leverages semantic associations to enhance the model's discriminative power. OGSample4Geo[3] exploits orientation information to improve localization accuracy. Another approach to improve performance is by advancing model architectures: TransGeo[23] and FSRA[2] replace the traditional CNN-based backbone with transformers[4], while GNN-Reranking[20] employs a Graph Neural Network as its backbone.

However, in multi-weather environments, the performance of these methods significantly decreases. To mitigate this domain gap, MuseNet[15] proposes two-branch neural network that dynamically adjusts to domain shifts caused by environmental changes, which can be viewed as encoding weather information implicitly. Given the challenge of multi-weather environments, which has not been sufficiently addressed, the ACM MM 2024 competition focused on this issue to spur advancements in this area.

2.2 Domain Translation

Domain translation refers to the method of transferring data from a source domain to a target domain while preserving the same semantic information[13]. In the ACM MM 2024 UAVM challenge, we observed that images in a normal weather target domain exhibit higher matching accuracy compared to those in adverse weather source domains, such as rain or fog. Therefore, techniques for deraining and dehazing images can be viewed as forms of domain translation. This domain translation can be achieved using image-to-image frameworks. For example, some works[17] have proposed using CNNs to separate foreground and background for this purpose. With the development of generative models, GANs have been applied to domain translation tasks. For instance, the Pix2Pix[7] framework can achieve various style transformations. Additionally, specific GAN-based methods[11, 12] have been introduced for targeted scenarios. Moreover, some studies[5, 18] have explored frequency domain learning. They utilize dual-guided designs that incorporate both frequency and spatial information to achieve their goals.

3 METHODOLOGY

The challenge of drone-based geo-localization is framed as an image retrieval task. Given an input drone-view image $\mathcal{X}_d \in \mathbb{R}^{H\times W\times C}$, where H, W, and C represent the image height, width, and number of channels, our objective is to determine the probability distribution $\mathcal{Y} \in [0,1]^{N_C}$ over N_C location categories for $\mathcal{X}_d$. This mapping is achieved using a feature extractor F and a classifier C. For a drone-view image $\mathcal{X}_d$ and a satellite-view image $\mathcal{X}_s$, the feature extractor F with shared weights generates the drone feature $f_d = F(\mathcal{X}_d)$ and the satellite feature $f_s = F(\mathcal{X}_s)$. These features, f_d and f_s, are then fed into the classifier C, resulting in the probability distribution $\mathcal{Y} = C(f_d, f_s)$. The entire network is trained by optimizing the loss function(1), where $\mathcal{Y}_{\text{label}}$ is the ground truth location category.

$$Loss = CrossEntropy(\mathcal{Y}_{label}, C(F(\mathcal{X}_d), F(\mathcal{X}_s))) \tag{1}$$

3.1 Data Augmentation

In the ACM MM 2024 competition, the task involves processing drone-view images enhanced under ten distinct weather conditions. As a preliminary approach, we experimented with weather-based data augmentation strategies on existing models, revealing a significant in performance across different weather scenarios. Two strategies were explored for weather simulation-based data augmentation. The first method integrated data augmentation within the PyTorch transforms pipeline, resulting in varying weather conditions for fixed images across iterations. The second method involved preprocessing the entire training dataset with data augmentation, ensuring fixed weather conditions for images throughout training. Notably, the latter approach demonstrated superior performance, prompting its exclusive adoption in this study. Remarkably, our experiments indicated that data augmentation even improved the MLPN model's

recognition capabilities in the normal domain. This observation motivates further exploration of diverse data augmentation techniques to enhance model performance in normal domain settings.

3.2 Domain Translation

The strategy of Domain Translation is based on our observation that models perform better under normal weather conditions than adverse weather conditions like rain or fog. Therefore, we aim to investigate whether applying Domain Translation to convert drone-view images from any domain into a normal domain as input for methods such as LPN, MLPN, and other related approaches to be presented at UAVM'24[21] can further enhance their recognition accuracy.

Considering the multi-environment scenario, we assume that $\mathcal{X}_d$ belongs to a domain $D = D_{\text{normal}}, D_{\text{rain}}, \cdots, D_{\text{snow}}$, which encompasses different environmental conditions. To translate $\mathcal{X}_d$ from $D \setminus D_{\text{normal}}$ to D_{normal}, we propose the framework shown in Figure 1. First, we use a classifier C_w that maps $\mathcal{X}_d$ to $\mathcal{W} \in [0,1]^{|D|}$ by training a neural network. It is noteworthy that C_w can also be implemented using airborne sensors or Web APIs to reduce unnecessary computational overhead. Based on the classification results of C_w, the input is directed to the corresponding Image2Image module to achieve domain translation, thereby aiming to improve detection performance.

3.2.1 Weather Classification. To facilitate the transformation of images from the source domain to the target domain, a weather classifier needs to be designed to identify the domain corresponding to each image, which will then be utilized by the subsequent Image2image module for conversion. For this weather classifier, we opted for independent training rather than end-to-end training, as we did not observe any significant benefits from end-to-end training in this context. Through experiments conducted on the ten domains included in the ACM MM 2024 competition, we found that ResNet-18 is sufficient for 256x256 input images, achieving a recognition accuracy exceeding 99%. In real-world scenarios involving drones, the size of the model could be further reduced to alleviate the computational burden on drone onboard systems. Additionally, Weather Classification can also be implemented through the drone's own equipment, such as onboard sensors and weather data obtained through networking.

3.2.2 Image2image Module. The Image2image module is designed to convert images of specific weather conditions into images of normal weather conditions. This is similar to tasks in computer vision such as de-raining and de-fogging, however, existing de-raining and de-fogging techniques often excel in handling individual weather conditions but lack generalizability across various scenarios. Therefore, we propose utilizing a general domain translation technique. In this technical report, we used both GAN based approach and Frequency Learning based approach.

Pix2pix With the advancement of generative methods, GAN has emerged as a primary tool for various de-raining and de-fogging tasks. We have explored the use of Pix2pix[7] as a general domain translation technique.

FFM However, as shown in the table 1, the issue with generative methods like Pix2pix is that they have high memory consumption and require significant computational resources, which may not be suitable for UAVs with limited onboard computing power.

Table 1: Results from two implementations of image2image module.

Method	PSNR↑		SSIM↑		Params↓	FLOPs↓
	Fog	Fog+Snow	Fog	Fog+Snow		
None	10.77	8.61	0.47	0.38	/	/
Pix2pix[7]	21.98	19.71	0.79	0.61	54.41M	36.32G
FFM	16.69	14.9	0.74	0.42	196.61K	63.31M

Inspired by [14], we conducted preliminary experiments on images enhanced under ten different weather conditions. The results, as shown in Figure 2, indicate that while the spatial differences between images before and after enhancement do not follow a fixed pattern due to the random spatial distribution of rain and fog, the differences in the frequency domain tend to exhibit a consistent pattern. Consequently, we propose using a Frequency Filtering Module to process images from each domain accordingly.

The design of the Frequency Filtering Module (FFM) is as Figure 4: Given an input drone-view image $\mathcal{X}_d$, the FFM first applies the Fourier transform $\mathcal{F}$ to obtain the amplitude spectrum $\mathcal{X}_{fa}$ and the phase spectrum $\mathcal{X}_{fp}$. Simultaneously, the input weather condition $\mathcal{W} \in [0,1]^{|D|}$ is used to weight the learnable filters $L \in \mathbb{R}^{|D| \times H \times W \times C}$, resulting in a weather-specific filter $L' \in \mathbb{R}^{H \times W \times C}$.

The filtered amplitude spectrum $\mathcal{X}'_{fa}$ is then computed as $L' \odot \mathcal{X}_{fa}$, where $\odot$ denotes the Hadamard product. Finally, the output image $\mathcal{X}'_d$ is obtained by applying the inverse Fourier transform $\mathcal{F}^{-1}$ to the filtered amplitude spectrum and the original phase spectrum, yielding $\mathcal{X}'_d = \mathcal{F}^{-1}(\mathcal{X}'_{fa}, \mathcal{X}_{fp})$. To train the learnable filters, the loss function is designed as follows:

$$Loss = (\mathcal{X}'_d - \mathcal{X}_o)^2 \tag{2}$$

where $\mathcal{X}_o$ is the original drone-view image. During training, $\mathcal{X}_d$ is generated from $\mathcal{X}_o$ using image augmentation techniques provided by imgaug.

As shown in Figure 3, Pix2pix demonstrates strong global domain translation capabilities, yet it introduces noticeable discrepancies in fine details compared to the original image. Conversely, Frequency Domain Mapping (FFM), as a method based on frequency domain learning, exhibits less proficiency in domain translation across highly distinct domains compared to generative approaches. However, FFM maintains fidelity in preserving details without introducing additional alterations.

Detailed comparisons of these methods across various datasets are provided in the accompanying table 1. Both methods effectively enhance the quality of input images, although Pix2pix achieves better enhancement, FFM requires significantly lower computational resources than Pix2pix.

4 EXPERIMENT

4.1 Implementation Details

The feature extractor F and classifier C used in our method are directly taken from MLPN[8]. Specifically, the model is retrained on the augmented university-1652 dataset. All parameters for this section use MLPN's default settings.

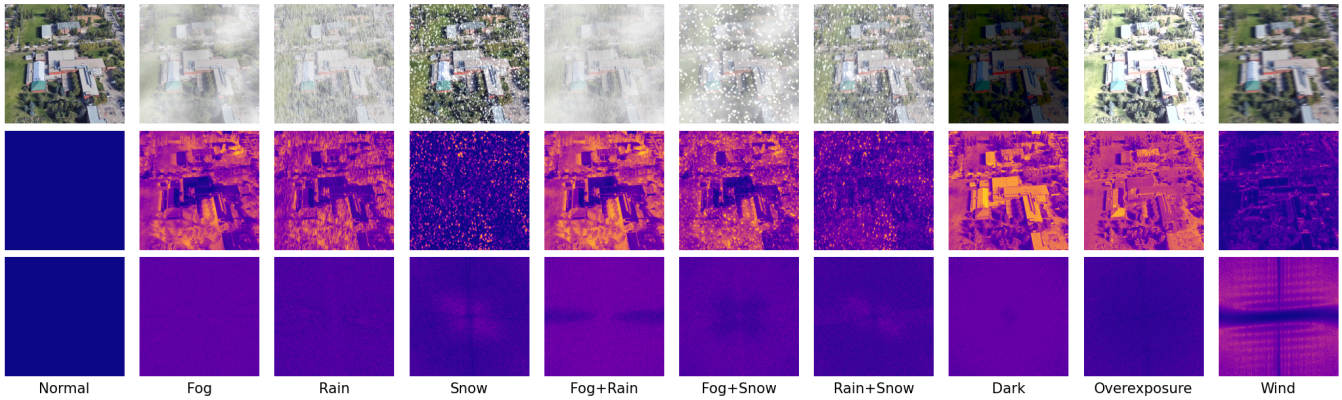

Figure 2: The top row exhibits the enhanced drone-view image. The second row delineates the spatial domain disparities between the enhanced and original images. Notably, these disparities are contingent upon both the structural nuances of the original image and the applied synthesized environments. The third row portrays discrepancies in the frequency domain.

(a) Adverse Image (b) Pix2pix[7] (c) FFM (d) Ground Truth

Figure 3: Visual results from two implementations of image2image module. The top row shows performance under foggy weather conditions, while the bottom row includes fog and snow.

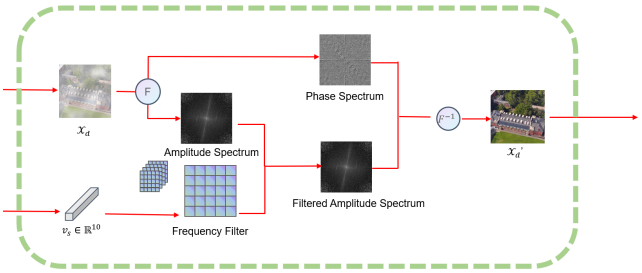

Figure 4: The design of proposed Frequency Filtering Module.

For the FFM module, we trained it on the entire university-1652 training set, with drone-view images augmented for the corresponding weather conditions, totaling 4,732 images. We used Adam as the optimizer with a learning rate of 0.001. The frequency filters converged after approximately two epochs. The implementation was done using PyTorch, and on an RTX 3080 Laptop GPU, each frequency filter required about half an hour to complete training.

4.2 Dataset

This technical report utilizes the training set of the University-1652 dataset for training and evaluates the model using the test set of University-1652 and the University-160k-wx dataset. The University-1652 dataset is the first to introduce the drone perspective for image retrieval tasks. It comprises images from the surroundings of 1,652 universities, based on Google Maps, including multiple street images, one satellite-view image, and 54 drone images per area. The dataset supports cross-view training methods.

The University-160k-wx dataset is an enhanced version of the University-160k dataset, where drone images have been augmented with various weather conditions, such as rain and snow, to simulate different weather scenarios. The University-160k dataset itself extends the satellite-view image count, incorporating a portion of satellite-view images as distractors. To better evaluate the matching accuracy under specific weather conditions, we further augmented the test set of the University-1652 dataset to include various weather effects, as the multi-weather images in University-160k-wx are fixed.

By incorporating these datasets, we aim to thoroughly assess the performance and robustness of our geo-localization methods under diverse environmental conditions.

4.3 Results

We conducted a series of experiments on the augmented university-1652 dataset. First, as shown in Table 2, we compared the performance of existing methods LPN and MLPN trained on the original university-1652 dataset and the augmented university-1652 dataset on the augmented university-1652 test set. We also compared these results with the MuseNet model specifically designed for multi-weather conditions. The results indicate that the augmented MLPN model outperforms all others, achieving the best performance. This model achieves a Recall1 of 85.08 on university160k-wx dataset, secured sixth place in the ACM MM 2024 Competition.

Table 2: Evaluation of Recall@1 and Average Precision (AP) accuracy for the drone-view target localization task using the University-1652 dataset[22] , with data augmented by 11 different environmental styles. Method denoted with * indicate training on the dataset with augmentation.

Method	Normal		Fog		Rain		Snow		Fog+Rain		Fog+Snow		Rain+Snow		Dark		Overexposure		Wind		Mixed	
	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP
LPN[16]	77.05	80.07	5.96	9.56	4.84	6.72	3.03	4.75	0.77	1.49	0.26	0.97	3.53	5.56	9.45	12.87	35.55	40.35	5.03	8.59	14.69	17.20
LPN*[16]	74.93	78.04	69.09	72.69	64.68	68.45	64.39	68.19	60.47	64.50	53.93	58.35	62.13	66.02	72.74	76.06	61.10	65.18	58.55	62.58	64.06	67.85
MLPN[8]	93.71	94.69	80.77	83.45	57.52	61.56	40.22	45.19	40.30	44.51	22.02	25.85	37.43	41.99	92.21	93.45	85.42	87.59	78.20	81.21	63.14	66.21
MLPN*[8]	94.07	95.03	93.38	94.45	92.05	93.31	90.97	92.41	91.27	92.64	89.94	91.52	91.23	92.61	93.92	94.91	90.40	91.96	89.58	91.17	91.75	93.05
MuseNet[15]	73.24	76.73	67.07	71.10	62.05	66.29	57.98	62.49	58.87	63.38	48.50	53.35	59.56	63.97	68.96	72.81	57.17	61.83	57.63	62.25	61.36	65.64
FFM (Ours)	94.07	95.03	92.67	93.83	90.87	92.26	84.85	87.08	88.73	90.42	80.36	83.04	86.79	88.77	93.71	94.72	89.50	91.19	89.30	90.94	89.09	90.72

Table 3: Comparisons with methods on University-160k-wx[21]. Method denoted with * indicate training on the original dataset with augmentation.

Method	Image Size	Backbone	University-160k-wx(R@1)
LPN[16]	256x256	Resnet-50	8.28
LPN*[16]	256x256	Resnet-50	47.56
MLPN[8]	256x256	Swin ViT v2	50.36
MLPN*[8]	256x256	Swin ViT v2	85.02
GAN[7]+MLPN*	256x256	Swin ViT v2	84.51
FFM+MLPN*	256x256	Swin ViT v2	81.07

Next, we explored the impact of additional Image2Image domain translation on detection results. As shown in Table 3, both the FFM-based domain translation module and the pix2pix-trained domain translation module resulted in decreased recognition performance. This decline in performance suggests that while domain translation methods like FFM and pix2pix are promising, they may introduce artifacts or alter important image features in ways that negatively impact the recognition accuracy in this specific task.

The negative results highlight the challenges of implementing domain translation for multi-weather conditions and suggest that further research is needed to refine these techniques. Despite the drop in performance, our findings contribute valuable insights for future research, particularly for engineering applications where robust performance across varying environmental conditions is critical.

5 CONCLUSION

In this paper, we present our solution to the ACM MM 2024 Multimedia Drone Satellite Matching Challenge. Firstly, we have found that data augmentation significantly enhances performance. Secondly, we propose a domain translation framework that is theoretically capable of enhancing performance and simple to implement. In implementing this framework, we introduce Frequency Filtering, which significantly reduces computational costs compared to GAN-based methods. Although we observed improvements in image evaluation metrics such as PSNR and SSIM after this processing, it did not enhance overall task performance, indicating the need for further exploration of this approach. We hope that our technical report will benefit future researchers and prove particularly useful for engineering applications in this field.

ACKNOWLEDGEMENT

This research was funded by the Zhejiang 'JIANBING' R&D Project (No. 2022C01055), R&D Project of Department of Transport of Zhejiang Province (No.2024011), Independent Research and Development Project of Zhejiang Scientific Research Institute of Transport (No.ZK202302, ZK202405 and ZK202409).

REFERENCES

[1] Eli Brosh, Matan Friedmann, Ilan Kadar, Lev Lavy, Elad Levi, Shmuel Rippa, Yair Lempert, Bruno Fernandez-Ruiz, Roei Herzig, and Trevor Darrell. 2019. Accurate Visual Localization for Automotive Applications. 1307–1316. https://doi.org/10.1109/CVPRW.2019.00170

[2] Ming Dai, Jianhong Hu, Jiedong Zhuang, and Enhui Zheng. 2021. A transformer-based feature segmentation and region alignment method for uav-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 7 (2021), 4376–4389.

[3] Fabian Deuser, Konrad Habel, Martin Werner, and Norbert Oswald. 2023. Orientation-Guided Contrastive Learning for UAV-View Geo-Localisation. In *Proceedings of the 2023 Workshop on UAVs in Multimedia: Capturing the World from a New Perspective*. 7–11.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[5] Xin Guo, Xueyang Fu, Man Zhou, Zhen Huang, Jialun Peng, and Zheng-Jun Zha. 2022. Exploring Fourier Prior for Single Image Rain Removal.. In *IJCAI*. 935–941.

[6] Thomas Hobiger, Seiichi Shimada, Shingo Shimizu, Ryuichi Ichikawa, Yasuhiro Koyama, and Tetsuro Kondo. 2010. Improving GPS positioning estimates during extreme weather situations by the help of fine-mesh numerical weather models. *Journal of atmospheric and solar-terrestrial physics* 72, 2-3 (2010), 262–270.

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

[8] Haoran Li, Quan Chen, Zhiwen Yang, and Jiong Yin. 2023. Drone Satellite Matching based on Multi-scale Local Pattern Network. In *Proceedings of the 2023 Workshop on UAVs in Multimedia: Capturing the World from a New Perspective*. 51–55.

[9] Tsung-Yi Lin, Yin Cui, Serge Belongie, and James Hays. 2015. Learning deep representations for ground-to-aerial geolocalization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5007–5015.

[10] Liu Liu and Hongdong Li. 2019. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5624–5633.

[11] Rui Qian, Robby T Tan, Wenhan Yang, Jiajun Su, and Jiaying Liu. 2018. Attentive generative adversarial network for raindrop removal from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2482–2491.

[12] Yanyun Qu, Yizi Chen, Jingying Huang, and Yuan Xie. 2019. Enhanced pix2pix dehazing network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8160–8168.

[13] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. 2021. Dino: A conditional energy-based gan for domain translation. *arXiv preprint arXiv:2102.09281* (2021).

[14] Kunyu Wang, Xueyang Fu, Yukun Huang, Chengzhi Cao, Gege Shi, and Zheng-Jun Zha. 2023. Generalized uav object detection via frequency domain disentanglement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1064–1073.

[15] Tingyu Wang, Zhedong Zheng, Yaoqi Sun, Chenggang Yan, Yi Yang, and Tat-Seng Chua. 2024. Multiple-environment Self-adaptive Network for Aerial-view Geo-localization. *Pattern Recognition* 152 (2024), 110363.

[16] Tingyu Wang, Zhedong Zheng, Chenggang Yan, Jiyong Zhang, Yaoqi Sun, Bolun Zheng, and Yi Yang. 2021. Each part matters: Local patterns facilitate cross-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*

32, 2 (2021), 867–879.

[17] Wenhan Yang, Robby T Tan, Jiashi Feng, Jiaying Liu, Zongming Guo, and Shuicheng Yan. 2017. Deep joint rain detection and removal from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1357–1366.

[18] Hu Yu, Naishan Zheng, Man Zhou, Jie Huang, Zeyu Xiao, and Feng Zhao. 2022. Frequency and spatial dual guidance for image dehazing. In *European Conference on Computer Vision*. Springer, 181–198.

[19] Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. 2017. Predicting ground-level scene layout from aerial imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 867–875.

[20] Xuanmeng Zhang, Minyue Jiang, Zhedong Zheng, Xiao Tan, Errui Ding, and Yi Yang. 2020. Understanding image retrieval re-ranking: A graph neural network perspective. *arXiv preprint arXiv:2012.07620* (2020).

[21] Zhedong Zheng, Yujiao Shi, Tingyu Wang, Jun Liu, Jianwu Fang, Yunchao Wei, and Tat-seng Chua. 2024. The 2nd Workshop on UAVs in Multimedia: Capturing the World from a New Perspective. In *Proceedings of the 32nd ACM International Conference on Multimedia Workshop*.

[22] Zhedong Zheng, Yunchao Wei, and Yi Yang. 2020. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In *Proceedings of the 28th ACM international conference on Multimedia*. 1395–1403.

[23] Sijie Zhu, Mubarak Shah, and Chen Chen. 2022. TransGeo: Transformer Is All You Need for Cross-view Image Geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1162–1171.