# Multi-weather Cross-view Geo-localization Using Denoising Diffusion Models


Tongtong Feng
fengtongtong@tsinghua.edu.cn
Tsinghua University
Beijing, China

Qing Li
soleilor@tsinghua.edu.cn
Tsinghua University
Beijing, China

Xin Wang*
xin_wang@tsinghua.edu.cn
Tsinghua University
Beijing, China

Mingzi Wang
wmz22@tsinghua.edu.cn
Tsinghua University
Beijing, China

Guangyao Li
guangyaoli@ruc.edu.cn
Renmin University of China
Beijing, China

Wenwu Zhu*
wwzhu@tsinghua.edu.cn
Tsinghua University
Beijing, China


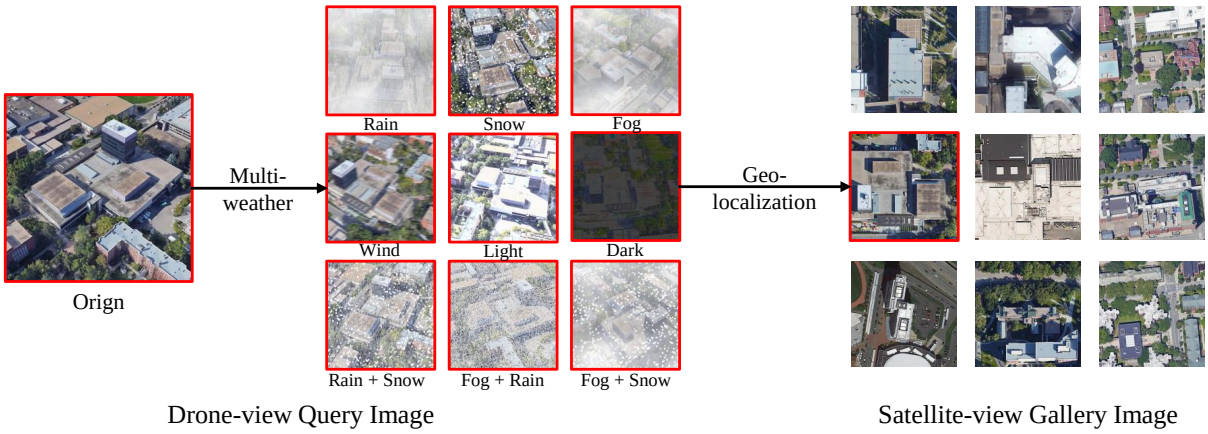


**Figure 1: Cross-view geo-localization depends on finding the correct location by matching drone-view images with satellite-view images. Weather variants, including fog, rain, snow, and multiple weather compositions, are randomly sampled to increase the difficulty of geo-localization. The red box represents the correct match we want to achieve regardless of weather conditions.**


## ABSTRACT

Cross-view geo-localization in GNSS-denied environments aims to determine an unknown location by matching drone-view images with the correct geo-tagged satellite-view images from a large gallery. Recent research shows that learning discriminative image representations under specific weather conditions can significantly enhance performance. However, the frequent occurrence of unseen extreme weather conditions hinders progress. This paper introduces MCGF, a Multi-weather Cross-view Geo-localization Framework designed to dynamically adapt to unseen weather conditions. MCGF establishes a joint optimization between image restoration and geo-localization using denoising diffusion models. For image restoration, MCGF incorporates a shared encoder and a lightweight restoration module to help the backbone eliminate weather-specific information. For geo-localization, MCGF uses EVA-02 as a backbone for feature extraction, with cross-entropy loss for training and cosine distance for testing. Extensive experiments on University160k-WX demonstrate that MCGF achieves competitive results for geo-localization in varying weather conditions.



## CCS CONCEPTS

• **Computing methodologies** → **Image representations**; • **Information systems** → *Top-k retrieval in databases*; Learning to rank.

## KEYWORDS

Cross-view Geo-localization, Multi-weather Restoration, Denoising Diffusion Model


## 1 INTRODUCTION

Cross-view geo-localization[1] aims to determine an unknown location by matching drone-view images with the correct geo-tagged satellite-view images from a large gallery, based on geographic features in the images, as shown in Figure 1. This task is crucial


*Corresponding author.




*ACM MM, 2024, Melbourne, Australia*



ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

for accurate navigation and safe planning[2–4] in GNSS-denied autonomous drone flights. Recent advances in vision transformer have led to significant breakthroughs in various cross-view geo-localization tasks, such as drone localization[5, 6] (matching drone-view query images with geo-tagged satellite-view images) and drone navigation[7, 8] (using satellite-view query images to guide drones to a target area). However, varying weather conditions, including fog, rain, snow, wind, light, dark, and combinations of multiple weather types, reduce visibility, corrupt the information captured by an image, significantly complicate image geographic representation, and lead to a sharp decline in task performance. The major challenge lies in adaptively achieving unbiased image geographic representation under diverse weather conditions.

A clean image without any weather degradation is desired in cross-view geo-localization. Early methods for weather removal using empirical observations [9], Convolutional Neural Networks (CNNs) based and transformer-based for deraining[10], dehazing[11], and desnowing[12]. Most of these methods achieve excellent performance, but these are not generic solutions for all adverse weather removal problems as the networks have to be trained separately for each weather[13]. The All-in-One Network[14] proposes a framework with separate encoders for each weather but a generic decoder and neural architecture search across weather-specific optimized encoders. The Transweather network[15] using vision transformer construction has a single encoder and a decoder and learns weather-type queries to solve all adverse weather removal efficiently. Wetherdiff[16] using diffusion models enables size-agnostic image restoration by using a guided denoising process. To our interest, these three studies focus on the inability of specific weather combinations to adapt to new weather types. Recently, MuSe-Net[17] employs a two-branch neural network containing one multiple-environment style extraction network and one self-adaptive feature extraction network to dynamically adjust the domain shift caused by environmental changes. However, this method does not perform well in some real-world high-intensity rains with a splattering effect.

To overcome these obstacles, this paper presents MCGF, a Multi-weather Cross-view Geo-localization Framework designed to dynamically adapt to unseen weather conditions, which establishes a joint optimization between image restoration and geo-localization using denoising diffusion models. In image restoration, MCGF includes a shared encoder and a lightweight restoration module that prompts the backbone to provide more beneficial information to eliminate the influence of weather-specific information. In geo-localization, MCGF uses EVA-02[18] as a backbone for feature extraction and uses cross-entropy loss for training and cosine distance for testing. EVA-02 is a ViT[19] model obtained using a series of stable optimization methods, which allows MCGF to extract more favorable information from drone and satellite images while using fewer parameters.

Diffusion models increasingly serve discriminative tasks such as classification and image segmentation. However, the geo-localization task with diffusion models under adverse weather conditions remains a challenging and under-explored area. Inspired by its powerful modeling capability and stable training process, we utilize the diffusion model to learn the denoising process from noisy images to clean images, facilitating robust matching in the presence of multi-weather.

Extensive experiments on University160k-WX demonstrate that MCGF achieves competitive results for geo-localization in varying weather conditions. The code will be released at https://github.com/fengtt42/ACMMM24-Solution-MCGF.

## 2 METHOD

MCGF establishes a joint optimization between image restoration and geo-localization using denoising diffusion models. The overview structure of MCGF is shown in Figure 2.

### 2.1 Denoising Diffusion Models

The diffusion model is a probabilistic model that has attracted considerable interest in the computer vision community. It can remarkably approximate the original data distribution by gradually adding Gaussian noise to the training data and learning to reverse this diffusion process.

The forward process is a fixed Markov Chain that sequentially corrupts the data $z_0 \sim q_\theta(z_0)$ at $T$ diffusion time steps, by injecting Gaussian noise according to a variance schedule $\beta_1, ..., \beta_T$. Given the clean drone-view images $z_0$, the forward process at step $t$ is defined as:

$$q_\theta(\mathbf{z}_t \mid \mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t; \sqrt{\alpha_t}\mathbf{z}_{t-1}, \beta_t\mathbf{I}) \tag{1}$$

$$q_\theta(\mathbf{z}_{1:T} \mid \mathbf{z}_0) = \prod_{t=1}^{T} q_\theta(z_t \mid z_{t-1}) \tag{2}$$

$$q_\theta(\mathbf{z}_t \mid \mathbf{z}_0) = \mathcal{N}(\mathbf{z}_t; \sqrt{\bar{\alpha}_t}\mathbf{z}_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{3}$$

where $\alpha_t$ and $\beta_t$ are noise schedule parameters, $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ and $\alpha_t = 1 - \beta_t$.

The reverse process attempts to remove the noise added in the forward process. The reverse process defined by the joint distribution $p_\theta(z_{0:T})$ is a Markov Chain with learned Gaussian denoising transitions starting at a standard normal prior $p_\theta(z_T) = \mathcal{N}(z_T; \mathbf{0}; \mathbf{I})$. At step $t$, the reverse process is defined as:

$$p_\theta(\mathbf{z}_{0:T}) = p(z_T) \prod_{t=1}^{T} p_\theta(z_{t-1} \mid z_t) \tag{4}$$

$$p_\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t) = \mathcal{N}(\mathbf{z}_{t-1}; \mu_\theta(\mathbf{z}_t, t), \Sigma_\theta(\mathbf{z}_t, t)) \tag{5}$$

For simplicity, we assume $\Sigma_\theta$ is a known constant, thus the reverse process simplifies to:

$$p_\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t) = \mathcal{N}(\mathbf{z}_{t-1}; \mu_\theta(\mathbf{z}_t, t), \sigma^2\mathbf{I}) \tag{6}$$

Here the reverse process is parameterized by a neural network that estimates $\mu_\theta(\mathbf{z}_t, t)$ and $\Sigma_\theta(\mathbf{z}_t, t))$. The forward process variance schedule $\beta_t$ can be learned jointly with the model or kept constant, ensuring that $z_t$ approximately follows a standard normal distribution.

The training objective of the denoising diffusion model is to maximize the likelihood of the reverse process, which can be achieved by minimizing the variational lower bound (VLB) of the negative log-likelihood. The VLB is given by:

$$\mathcal{L}_{\text{VLB}} = \mathbb{E}_q \left[ -\log p_\theta(\mathbf{z}_0) + \right.$$

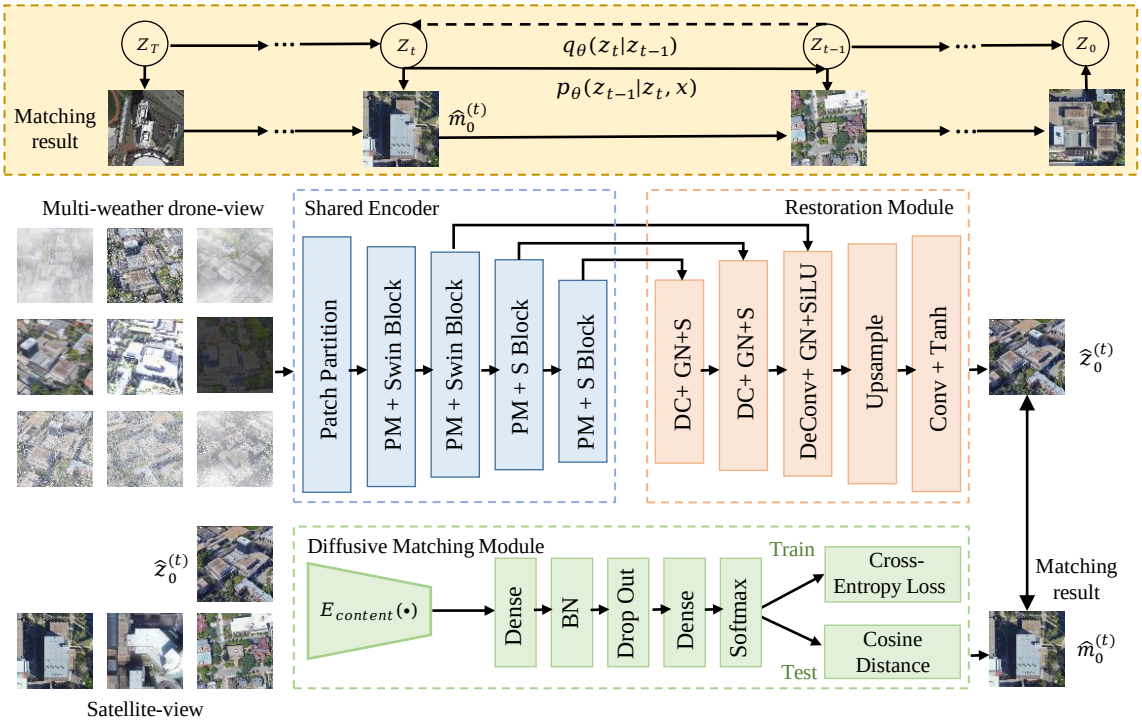


**Figure 2: The overview structure of MCGF. MCGF establishes a joint optimization between image restoration and geo-localization using denoising diffusion models.**

$$\sum_{t=1}^{T} \mathcal{D}_{KL}\left[q_\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t, \mathbf{z}_0) \,\|\, p_\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t)\right] \tag{7}$$

In practice, this can be decomposed into reconstruction error and KL divergence terms for each step, which are optimized accordingly.

## 2.2 Shared Encoder

To enhance feature representation and improve subsequent image restoration and geo-localization, we utilize the widely adopted state-of-the-art transformer-based model, Swin Transformer[20], as the shared encoder in our unified framework. The Swin Transformer is a hierarchical transformer that employs shifted windows, which restricts attention computation to non-overlapping local windows, making it adaptable for modeling at various scales. To balance computational overhead and inference speed, we select the tiny version of Swin Transformer as the default backbone.

## 2.3 Restoration Module

The restoration module utilizes a straightforward CNN-based encoder architecture, consisting of three deconvolutions, an upsampling, and a $Tanh$ activation function. It facilitates geo-localization by revealing clean features at multiple scales and produces weather-free images. We adopt a simple Mean Squared Error (MSE) as the loss function of the restoration subnetwork.

$$L_{res} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{8}$$

where $n$ denotes the patch size. It can minimize the pixel-wise difference between the clean image $Y_i$ and the estimated weather-free image $\hat{Y}_i$.

## 2.4 Diffusive Matching Module

*Feature extraction.* MCGF introduces the latest transformer-based visual representation, EVA-02, as the backbone of $E_{content}(.)$ in the network. In fact, EVA-02 has shown superior performance in most CV downstream tasks. EVA's architecture is a vanilla ViT encoder that can be regarded as a student model, with a shape following ViT giant and the vision encoder of BEiT-3. A big dataset, consisting of several typical and openly accessible datasets with 29.6 million images in total, is used as pre-training data. After pre-training, EVA is scaled up to 1.0B parameters compared to CLIP. Based on the theory of EVA, larger CLIP-like models will provide more robust target representations for masking image modeling.

*Loss calculation.* The feature map extracted by $E_{content}(.)$ encoder is fed into a multilayer perceptron (MLP) to calculate the cross-entropy loss for training or cosine distance for testing. MLP includes 2 dense layers, a Batch Normalization (BN) layer, a drop out layer, and a softmax activation function.

*Optimization.* MCGF contains two loss functions, one is image restoration loss $L_{res}$, and the other is matching loss. In the gradual denoising process of the diffusing model, the matching model can gradually obtain clearer drone-viewing images. This process enables the matching model to run at multiple granularities, resulting in more accurate matching results.

**Table 1: Matching results compared with SOTA methods.**

| Methods | R@1 | R@5 | R@10 | AP |
|---|---|---|---|---|
| LPN[22] | 7.98 | 10.25 | 11.21 | 8.49 |
| MBEG[23] | 26.17 | 32.84 | 35.32 | 29.32 |
| Muse-Net[17] | 50.48 | 63.19 | 67.34 | 53.27 |
| MCGF(ours) | 84.68 | 91.36 | 93.00 | 88.71 |

## 3 EXPERIMENT

*Dataset.* University160k-WX[21] is a multi-weather cross-view geo-localization dataset, which extends the University-1652 dataset with extra 167,486 satellite-view gallery distractors. University160k-WX further introduces weather variants on University160k, including fog, rain, snow and multiple weather compositions. These distractor satellite-view images have a size of 1024 × 1024 and are obtained by cutting orthophoto images of real urban and surrounding areas. Multiple weathers are randomly sampled to increase the difficulty of representation learning.

*Implement details.* We employed the EVA-02 model, which is based on the Vision Transformer, as the backbone for diffusive matching module. This model has been trained and fine-tuned on many large vision datasets. In our experiments, we resized each input image to a fixed size of 448 × 448 pixels. During training, we used SGD as the optimizer with a momentum of 0.9 and weight decay of 5 × 104, with a mini-batch size of 16. The initial learning rate was set to 0.01 for the backbone layer and 0.1 for the classification layer. Our model was built using Pytorch.

*Evaluation metrics.* The performance of our method is evaluated by the Recall@K (R@K) and the average precision (AP). R@K denotes the proportion of correctly localized images in the top-K list, and R@1 is an important indicator. AP is equal to the area under the Precision-Recall curve. Higher scores of R@K and AP indicate better performance of the network.

### 3.1 Geo-localization results

We train MCGF with outstanding algorithms (including LPN[22], MBEG[23], and Muse-Net[17]) on the University-160k-WX train set until convergence and obtain optimal results. We test all trained models on the official unified test set provided by the competition organizer. All test results can be displayed and downloaded on the competition result submission platform. Table 1 shows that MCGF is significantly better than existing methods in all evaluation metrics. Especially compared with the latest research Muse-Net, MCGF can achieve a 67.75% performance improvement in the Recall@1 indicator. MCGF shows considerable potential for geo-localization as a general framework.

### 3.2 Visualization

As shown in Figure 3, we visualise heatmaps and Top-5 matching results generated by our method in 10 different weather conditions. Since the drone is flying around, the drone images is not only interfered by weather but also by rotational posture. Therefore, we also show the impact of drone posture changes on geo-localization in Figure 3. The heatmap shows that our method can accurately

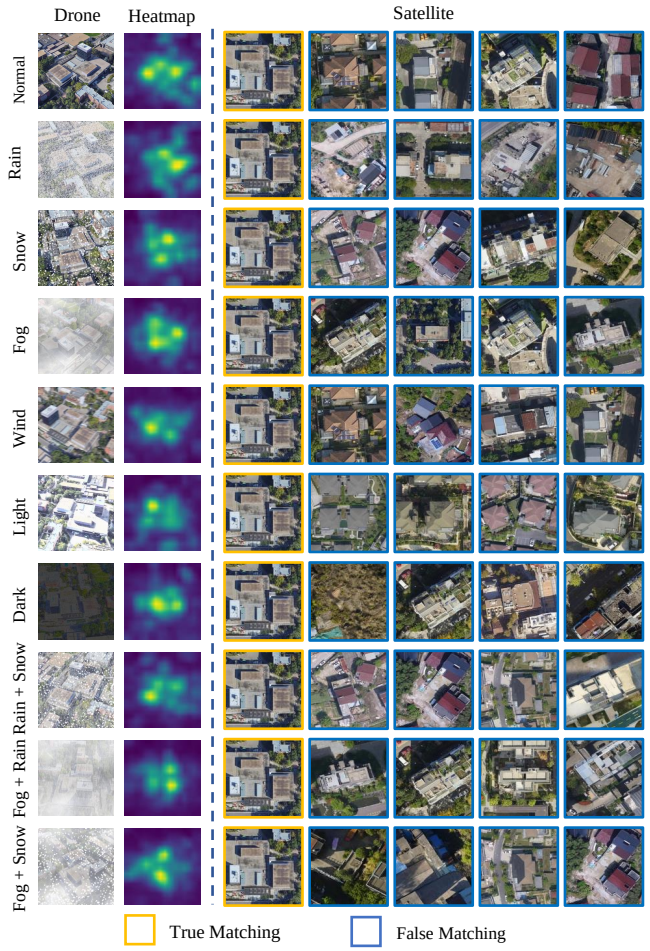


**Figure 3: Visualization of heatmaps generated by our method and Top-5 matching results for a drone-view image in different conditions.**

extract the shape and relative position of geographic targets under weather and pose interference. From the matching results shown, we observe that our model obtains the true match in the Top-1 yet the remaining matching results are inconsistent under 10 different conditions, which also indicates that the adjusted features still contain a few discrepancies.

### 3.3 Conclusion

In this paper, we presents MCGF, a Multi-weather Cross-view Geo-localization Framework designed to dynamically adapt to unseen weather conditions, which establishes a joint optimization between image restoration and geo-localization using denoising diffusion models. In image restoration, MCGF includes a shared encoder and a lightweight restoration module. In geo-localization, MCGF uses EVA-02 as a backbone for feature extraction and uses cross-entropy loss for training and cosine distance for testing. Extensive experiments on University160k-WX demonstrate that MCGF achieves competitive results for geo-localization in varying weather.

# REFERENCES


[1] Yujiao Shi, Xin Yu, Liu Liu, Dylan Campbell, Piotr Koniusz, and Hongdong Li. Accurate 3-dof camera geo-localization via ground-to-satellite image matching. *IEEE transactions on pattern analysis and machine intelligence*, pages 2682–2697, 2022.

[2] Sizhe Wei, Yuxi Wei, Yue Hu, Yifan Lu, Yiqi Zhong, Siheng Chen, and Ya Zhang. Asynchrony-Robust Collaborative Perception via Bird's Eye View Flow. *Advances in Neural Information Processing Systems*, 36, 2024.

[3] Yifan Lu, Yue Hu, Yiqi Zhong, Dequan Wang, Siheng Chen, and Yanfeng Wang. An Extensible Framework for Open Heterogeneous Collaborative Perception. *The Twelfth International Conference on Learning Representations*, 2024.

[4] Binyu Zhao, Wei Zhang, and Zhaonian Zou. BM2CP: Efficient Collaborative Perception with LiDAR-Camera Modalities. *arXiv preprint arXiv:2310.14702*, 2023.

[5] Zhenbo Song, Jianfeng Lu, Yujiao Shi, et al. Learning dense flow field for highly-accurate cross-view camera localization. 2024.

[6] Yujiao Shi, Xin Yu, Shan Wang, and Hongdong Li. Cvlnet: Cross-view semantic correspondence learning for video-based camera localization. In *Asian Conference on Computer Vision*, pages 123–141, 2022.

[7] Yujiao Shi, Fei Wu, Akhil Perincherry, Ankit Vora, and Hongdong Li. Boosting 3-dof ground-to-satellite camera localization accuracy via geometry-guided cross-view transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21516–21526, 2023.

[8] Yujiao Shi and Hongdong Li. Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17010–17020, 2022.

[9] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, pages 2341–2353, 2010.

[10] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, pages 2341–2353, 2010.

[11] Jingang Zhang, Wenqi Ren, Shengdong Zhang, He Zhang, Yunfeng Nie, Zhe Xue, and Xiaochun Cao. Hierarchical density-aware dehazing network. *IEEE Transactions on Cybernetics*, pages 11187–11199, 2021.

[12] Kaihao Zhang, Rongqing Li, Yanjiang Yu, Wenhan Luo, and Changsheng Li. Deep dense multi-scale network for snow removal using semantic and depth priors. *IEEE Transactions on Image Processing*, pages 7419–7431, 2021.

[13] Dong Zhao, Jia Li, Hongyu Li, and Long Xu. Hybrid local-global transformer for image dehazing. *arXiv preprint arXiv:2109.07100*, 2(3), 2021.

[14] Ruoteng Li, Robby T Tan, and Loong-Fah Cheong. All in one bad weather removal using architectural search. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3175–3185, 2020.

[15] Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M Patel. Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2353–2363, 2022.

[16] Ozan Özdenizci and Robert Legenstein. Restoring vision in adverse weather conditions with patch-based denoising diffusion models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10346–10357, 2023.

[17] Tingyu Wang, Zhedong Zheng, Yaoqi Sun, Chenggang Yan, Yi Yang, and Tat-Seng Chua. Multiple-environment self-adaptive network for aerial-view geo-localization. *Pattern Recognition*, 152:110363, 2024.

[18] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *arXiv preprint arXiv:2303.11331*, 2023.

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

[20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[21] Zhedong Zheng, Yujiao Shi, Tingyu Wang, Chen Chen, Pengfei Zhu, and Richard Hartley. 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*, 2024.

[22] Tingyu Wang, Zhedong Zheng, Chenggang Yan, Jiyong Zhang, Yaoqi Sun, Bolun Zheng, and Yi Yang. Each part matters: Local patterns facilitate cross-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(2):867–879, 2021.

[23] Runzhe Zhu, Mingze Yang, Kaiyu Zhang, Fei Wu, Ling Yin, and Yujin Zhang. Modern backbone for efficient geo-localization. In *Proceedings of the 2023 Workshop on UAVs in Multimedia: Capturing the World from a New Perspective*, pages 31–37, 2023.