UAV-TO-SATELLITE IMAGE GEO-LOCALIZATION VIA DUAL-BRANCH DEEP LEARNING: A COMPARATIVE STUDY OF RESNET-18 AND MLP-MIXER

Qinghan Yu

Belarusian State University of Informatics and Radioelectronics
Minsk, Republic of Belarus

Jun Ma – Assistant

Abstract. UAV-to-satellite cross-view geo-localization is an important technology for autonomous navigation in GNSS-denied environments. However, large differences in viewpoint, scale, and distortion between UAV and satellite images make feature matching difficult. This paper presents a dual-branch deep learning framework and compares two backbone architectures, ResNet-18 and MLP-Mixer, under the same bidirectional InfoNCE training objective. Experiments on the UAV-Visloc dataset show that MLP-Mixer achieves better retrieval performance than ResNet-18, reaching Recall@1 of 74.09% versus 72.40%, with consistent improvements at Recall@5 and Recall@10. Ablation results further show that independent branches are important for handling the domain gap between UAV and satellite imagery. The results indicate that pure MLP architectures have strong potential for cross-view geo-localization when combined with contrastive learning.

Keywords. Cross-view geo-localization, UAV-satellite images, ResNet-18, MLP-Mixer, contrastive learning.

Introduction

UAV-to-satellite cross-view geo-localization provides an effective solution for positioning in GNSS-denied environments by matching low-altitude UAV images with high-altitude satellite images. This task is challenging because the two views differ significantly in perspective, scale, rotation, and visual appearance. Although CNN-based and Transformer-based methods have been widely studied, pure MLP architectures remain less explored in this scenario.

In this paper, we adopt a dual-branch framework for UAV-satellite image matching and compare two representative backbone networks, ResNet-18 and MLP-Mixer, under the same contrastive learning setting. The goal is to evaluate their retrieval accuracy and analyze whether pure MLP architectures can serve as competitive alternatives to CNNs in cross-view geo-localization. Experimental results show that MLP-Mixer consistently outperforms ResNet-18 on the UAV-Visloc dataset.

Method

The proposed framework uses two independent branches to process UAV and satellite images separately, which helps address the large domain gap between the two views. Each branch employs either ResNet-18 or MLP-Mixer as the feature extractor. The extracted features are then passed through a projection head to generate 256-dimensional normalized embeddings.

$$\mathcal{L}_{\mathrm{d}\to\mathrm{s}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(\mathbf{z}_{\mathrm{d},i}^{\top}\mathbf{z}_{\mathrm{s},i}\ /\ \tau\right)}{\sum_{j=1}^{B}\exp\left(\mathbf{z}_{\mathrm{d},i}^{\top}\mathbf{z}_{\mathrm{s},j}\ /\ \tau\right)}$$

The model is trained using bidirectional InfoNCE loss, which pulls matched UAV-satellite pairs closer in the embedding space while pushing unmatched pairs apart. During inference, cosine similarity is computed between the UAV query embedding and all satellite gallery embeddings, and the gallery images are ranked according to similarity. Retrieval performance is evaluated by Recall@1, Recall@5, and Recall@10.

Experiments

Experiments are conducted on the UAV-Visloc dataset containing 768 UAV-satellite image pairs. The dataset is split into 80% training and 20% validation using a fixed random seed. All models are trained for 80 epochs with the AdamW optimizer. The embedding dimension is set to 256, and retrieval performance is measured by Recall@1, Recall@5, and Recall@10.

The main comparison results are shown in Table 1. MLP-Mixer achieves better performance than ResNet-18 across all evaluation metrics. In particular, MLP-Mixer improves Recall@1 from 72.40% to 74.09%, Recall@5 from 87.50% to 88.41%, and Recall@10 from 90.49% to 92.45%. These results suggest that the global token-mixing mechanism of MLP-Mixer is effective for modeling spatial correspondences between UAV and satellite views.

Table 1

Model	R@1	R@5	R@10
ResNet-18 (Baseline)	72,40%	87,50%	90,49%
MLP-Mixer (Baseline)	74,09%	88,41%	92,45%

To validate the importance of the dual-branch design, we further compare independent branches with shared weights. The results in Table 2 show that independent branches are clearly more effective, especially for ResNet-18. When shared weights are used, ResNet-18 drops from 72.40% to 49.35% in Recall@1, indicating that forcing UAV and satellite images into a single feature extractor significantly reduces retrieval performance. Although MLP-Mixer is less sensitive to weight sharing, the independent-branch setting still gives the best result. This confirms that separate feature extraction pathways are necessary for handling the domain gap between the two image modalities.

Table 2

Model	Independent Branches	Shared Weights	Δ R@1
ResNet-18	72,40%	49,35%	-23,05%
MLP-Mixer	74,09%	72,79%	-1,30%

Visualization

This advantage is also reflected in qualitative retrieval results. As shown in Figure 1, MLP-Mixer more consistently ranks the correct satellite image at the top positions, whereas ResNet-18 is more likely to confuse visually similar but geographically different locations. These examples demonstrate that MLP-Mixer learns more discriminative representations for practical cross-view retrieval.

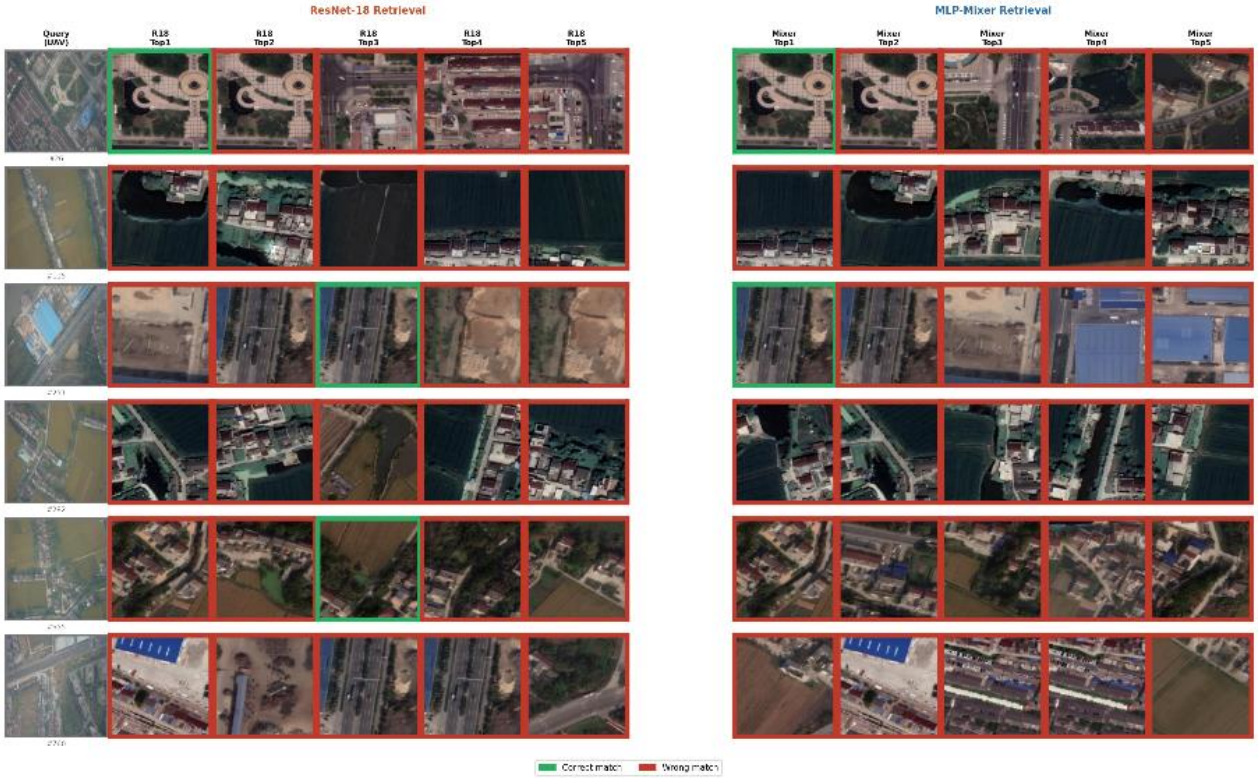

Figure 1 – The visualization results

Conclusion

This paper presents a compact comparative study of ResNet-18 and MLP-Mixer for UAV-to-satellite cross-view geo-localization under a dual-branch contrastive learning framework. Experimental results show that MLP-Mixer achieves consistently better retrieval performance than ResNet-18 on the UAV-Visloc dataset. In addition, the ablation study demonstrates that independent branches are essential for effective cross-view matching. These findings indicate that pure MLP architectures are promising for UAV-satellite geo-localization and deserve further investigation in larger-scale and more diverse datasets.

References:

1 K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770–778.

2 I. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, M. Lucic, and A. Dosovitskiy, "MLP-Mixer: An all-MLP architecture for vision," in Advances in Neural Information Processing Systems (NeurIPS), vol. 34, 2021, pp. 24261–24272.

3 T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in Proc. Int. Conf. Machine Learning (ICML), vol. 119, 2020, pp. 1597–1607.

4 Y. Shi, L. Liu, X. Yu, and H. Li, "Spatial-aware feature aggregation for image based cross-view geo-localization," in Advances in Neural Information Processing Systems (NeurIPS), vol. 32, 2019, pp. 10090–10100.

5 S. Zhu, M. Shah, and C. Chen, "TransGeo: Transformer is all you need for cross-view image geo-localization," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1162–1171.

6 Z. Zheng, Y. Wei, and Y. Yang, "University-1652: A multi-view multi-source benchmark for drone-based geo-localization," in Proc. ACM Int. Conf. Multimedia (ACM MM), 2020, pp. 1395–1403.