UDC 621.391

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, Republic of Belarus

Received April 5, 2026

Abstract. Unmanned Aerial Vehicle (UAV)-satellite cross-view geolocalization is a core technology for autonomous navigation in GNSS-denied environments, which achieves precise pose estimation by matching low-altitude oblique UAV images with high-altitude orthophoto satellite images. However, the huge visual differences in perspective, scale and distortion between the two types of images significantly increase the difficulty of feature matching. Current mainstream methods are mostly based on CNN or Transformer architectures, while systematic comparative studies of pure MLP architectures in this task are still insufficient, and their performance differences and efficiency trade-offs remain unclear. To this end, this paper compares and analyzes the accuracy, robustness and computational efficiency of CNN and pure MLP architectures in UAV-satellite cross-view geolocalization on a unified benchmark dataset. The experimental results show that CNN has advantages in local feature extraction and is more robust to perspective distortion, while pure MLP architectures show potential in global feature modeling and inference speed, but their performance needs to be improved in small-sample and extreme perspective change scenarios. The comparative analysis in this paper provides an empirical basis for the architecture selection and optimization of cross-view geolocalization models.

Keywords: cross-view geolocalization; UAV-satellite images; CNN; pure MLP architecture; feature matching.

Introduction

Unmanned Aerial Vehicles (UAVs) have been widely applied in key scenarios such as autonomous navigation, emergency search and rescue, and land resource surveying due to their advantages of flexible deployment, low cost, and high-resolution imaging, among which UAV-satellite cross-view geolocalization serves as one of the core technologies supporting these tasks, achieving precise pose estimation of UAVs in the global geographic coordinate system by matching low-altitude oblique UAV images with high-altitude orthophoto satellite images, thus providing a reliable solution for positioning in GNSS-denied environments [1]. However, there exist huge visual differences between UAV and satellite perspectives, including perspective and scale differences where UAVs capture low-altitude oblique images with flexible and variable viewpoints while satellites take high altitude orthophotos with fixed perspectives, leading to distinct visual appearances of the same geographic entity due to significant disparities in imaging scale and projection mode, rotation and distortion differences caused by unstable UAV flight attitudes that introduce image rotation and distortion to further increase feature matching difficulty, as well as illumination and seasonal differences that alter lighting conditions and vegetation coverage to affect feature stability and repeatability, all of which pose severe challenges to cross-view feature matching [1]. At present, research on cross-view geolocalization mainly focuses on CNN and Transformer architectures, with the former performing well in feature extraction by virtue of local inductive bias and the latter excelling in global context modeling. Yet pure MLP architectures, as a simple model without inductive bias that has shown performance comparable to CNN in general computer vision tasks [5], have not been verified for their applicability in UAV-satellite cross-view geolocalization, and there remains a clear gap in systematic comparative studies with CNN, leaving their performance differences and efficiency trade-offs in this task unclear. Against this background, this

paper conducts a systematic comparative experiment between CNN and pure MLP architectures in UAV-satellite cross-view geolocalization on a unified benchmark dataset [4], analyzing their differences in positioning accuracy, robustness to visual variations, and computational efficiency, revealing the performance rules of different architectures under key challenges such as perspective distortion and scale change to provide empirical guidance for model selection, and discussing the optimization direction of pure MLP architectures in this scenario to offer a reference for subsequent research, with the rest of the paper organized as follows: Section 2 reviews related work on cross-view geolocalization and pure MLP architectures, Section 3 introduces the experimental dataset, comparative models, and evaluation metrics, Section 4 analyzes and discusses the experimental results, and Section 5 summarizes the full paper and prospects future work.

Related Work

Cross-view geolocalization focuses on matching low-altitude UAV oblique images with high-altitude satellite orthophotos to estimate precise UAV poses in GNSS-denied environments, serving as a core technology for autonomous navigation and emergency search and rescue [1]. Classic benchmark datasets including CVUSA [6], VIGOR [7], and University-1652 [3] have been established to evaluate task performance, among which University-1652 is widely adopted as the standard testbed for current UAV-satellite cross-view localization due to its diverse campus and urban scenes and well-aligned matching image pairs [3]. Early studies relied on hand-crafted features such as SIFT [8] and HOG [9] to capture invariant visual patterns for cross-view matching, but these methods struggled with severe perspective, scale, and distortion differences between UAV and satellite images, resulting in limited accuracy and robustness that failed to meet practical application demands. With the rise of deep learning, data-driven approaches gradually replaced hand-crafted feature methods, opening new avenues for addressing the challenges of cross-view geolocalization.

CNN-based architectures have dominated cross-view geolocalization research due to their inherent local inductive bias, which enables effective extraction of key local visual features such as building edges, road intersections, and texture patterns that are critical for matching. Representative works include NetVLAD [10] and CVM-Net [1]: NetVLAD introduces a vector of locally aggregated descriptors to aggregate convolutional features into compact global representations, laying the foundation for efficient large-scale cross-view image retrieval [10], while CVM-Net employs a siamese network structure with shared CNN weights to learn shared feature spaces between UAV and satellite images, and uses metric learning objectives to minimize the feature distance of matching image pairs and maximize that of non-matching pairs, significantly improving cross-view matching accuracy [1]. Despite their advantages in local feature robustness and parameter efficiency, CNN-based methods suffer from limited global context modeling ability, as their fixed-size receptive fields make it difficult to capture long-range dependencies between geographic entities, leading to potential feature misalignment under extreme perspective and scale variations.

In recent years, Transformer architectures have emerged as promising alternatives to CNNs in computer vision [11], offering new possibilities for cross-view geolocalization. TransGeo applies the self-attention mechanism of Transformers to cross-view localization, modeling global contextual relationships between image regions to alleviate the limitations of CNNs' local receptive fields, thus achieving better performance in scenarios with large-scale perspective differences [2]. As a representative pure MLP architecture, MLP-Mixer discards convolutional and attention operations entirely, relying solely on fully connected layers to interact across spatial and channel dimensions of feature maps, and has achieved performance comparable to state-of-the-art CNNs on general visual benchmarks like ImageNet, demonstrating its potential in simple structure and global feature modeling [5]. However, existing studies in cross-view geolocalization mainly focus on Transformer-based models, and the applicability of pure MLP architectures in UAV-satellite extreme perspective difference scenarios remains unvalidated, with a clear lack of systematic comparative analysis between pure MLPs and classic CNNs in this specific task.

Metric learning losses play a crucial role in optimizing cross-view geolocalization models, as the task essentially boils down to image retrieval where the goal is to rank matching image pairs higher than non-matching ones. Early works predominantly used Triplet Loss [12], which constructs triplets consisting of an anchor image, a positive matching image, and a negative non-matching image to pull

the anchor and positive features closer while pushing the anchor and negative features apart, effectively optimizing the feature embedding space for cross-view matching. Subsequent studies have shifted to contrastive learning losses such as InfoNCE and NT-Xent [13], which sample negative examples within a training batch to construct more efficient contrastive constraints, leading to faster convergence and better generalization performance under large-batch training settings. This loss has become a standard choice for training current cross-view geolocalization models [14]. Despite the maturity of metric learning loss designs, their compatibility and performance trade-offs with different backbone architectures in cross-view geolocalization have not been thoroughly explored.

Methodology

This work focuses on cross-view geo-localization, which aims to find the satellite image tile that matches a given UAV query image in geographic location. Each matching pair of UAV and satellite images shares the same GPS coordinates, but differs greatly in viewing angle, imaging scale, and data modality. This task is treated as an image retrieval problem, where the model learns an embedding function to make matching image pairs more similar and non-matching pairs more distinct in the feature embedding space [15]. For each UAV query image with a known correct satellite match, the model is trained to ensure that the similarity between the query and its ground-truth match is higher than the similarity between the query and any other non-matching satellite tile. The similarity measurement used in this work is cosine similarity, and all feature embeddings are normalized using L2 normalization to ensure stable and consistent similarity calculation [14].

The proposed framework uses a dual-branch structure with two independent encoders for UAV images and satellite images respectively. The design of independent branches is based on the large domain gap between UAV and satellite imagery. Using shared weights for both branches would force the two different types of images into a single feature space, which is proven to be ineffective through experiments. For each pair of matching UAV and satellite images, the two branches extract fixed-dimensional feature embeddings independently. These embeddings are then normalized and used for contrastive learning. The dual-branch structure allows each branch to learn specialized features suitable for its own image type: UAV images contain rich texture details from oblique perspectives, while satellite images show clear global spatial layouts from top-down views. Through contrastive learning, the two sets of features are aligned to form a unified embedding space where cross-view matching can be effectively realized [1].

ResNet-18 is a lightweight convolutional neural network with residual connections, containing four residual stages with gradually increasing channel dimensions [16]. For an input image, the network extracts features layer by layer and outputs a fixed-length feature vector through global average pooling. Both the UAV branch and the satellite branch are initialized with weights pre-trained on the ImageNet dataset [17], which is critical for model performance as verified in the ablation study. Residual connections ensure smooth gradient propagation during training [16], and hierarchical feature extraction captures multi-scale spatial information. Each ResNet-18 branch has about 11,2 million trainable parameters. The choice of ResNet-18 instead of deeper networks is based on computational efficiency and better generalization to domain-shifted data such as satellite images.

MLP-Mixer is a pure MLP architecture that replaces convolution and self-attention with two alternating MLP operations applied to image patches [5]. The input image is divided into multiple non-overlapping patches, and each patch is projected into a feature token to form a token matrix. Each MLP-Mixer layer executes two operations in sequence: token-mixing MLP for spatial interaction between different patches, and channel-mixing MLP for feature interaction within each patch [5]. The activation function used is GELU, and layer normalization is applied before each MLP operation. The final feature representation is obtained by global average pooling on the token dimension. Unlike CNNs that capture local spatial features layer by layer, MLP-Mixer performs global spatial reasoning from the first layer through token-mixing, which is beneficial for cross-view matching that relies on global landmark layout [5]. The MLP-Mixer branch has about 59,9 million trainable parameters. Its pure MLP structure removes the inductive bias of convolution, forcing the model to learn spatial relationships explicitly, which helps adapt to large viewpoint changes in cross-view tasks.

Following the design of SimCLR [13], a two-layer MLP projection head is added after each backbone network to map backbone features into a compact 256-dimensional embedding space. The

projection head separates the backbone feature learning from the metric learning objective, avoiding over-specialization of backbone features to the contrastive loss [13]. All final embeddings are processed with L2 normalization, so that all embeddings are distributed on a unit hypersphere, making cosine similarity equivalent to dot product similarity [14]. This normalization step is essential for contrastive learning, as it eliminates the influence of embedding magnitude on similarity calculation. The 256-dimensional embedding space is a standard setting that balances feature expression ability and computational cost.

The InfoNCE loss, also known as NT-Xent loss, is used as the main training objective [13]. For a batch of matching UAV-satellite pairs, the loss calculates the contrastive constraint from UAV embeddings to satellite embeddings. A temperature hyperparameter is used to adjust the concentration of the similarity distribution, where a smaller temperature makes matching pairs cluster more tightly [13]. To utilize the symmetry of the cross-view matching task, the loss is applied in both directions: from UAV to satellite and from satellite to UAV. This bidirectional loss optimizes both matching directions simultaneously, improving model convergence and final performance. Compared with Triplet Loss [12], InfoNCE uses all sample pairs in the batch to calculate gradients, providing denser and more effective learning signals [14]. The model learns to maximize the similarity between each query and its correct match, while minimizing the similarity between the query and all other non-matching samples.

In the inference stage, all satellite tiles in the gallery are encoded by the satellite branch to build a feature bank. For each input UAV query image, the model extracts its feature embedding and computes the similarity with all embeddings in the satellite feature bank. The satellite tiles are sorted in descending order of similarity, and the GPS coordinates of the top-ranked tile are taken as the predicted location. The retrieval performance is evaluated using Recall@K metrics, which represent the percentage of queries whose correct satellite match appears in the top-K ranked results [3]. The main evaluation metrics are R@1, R@5, and R@10, among which R@1 is the most stringent and practical index for real geo-localization applications.

To verify the advantage of InfoNCE, a Triplet Loss version with hard negative mining is also implemented [12]. For each UAV query, the hardest negative sample is defined as the non-matching satellite tile with the highest similarity to the query. Triplet Loss only uses this single hardest negative for each anchor, resulting in sparse gradient signals [12]. In contrast, InfoNCE uses all samples in the batch as negatives, providing much denser gradients [14]. The ablation experiment shows that InfoNCE achieves significantly better performance than Triplet Loss, especially for MLP-Mixer, which confirms the rationality of choosing InfoNCE as the training loss.

Experiments and Results

We conduct experiments on the UAV-Visloc dataset comprising 768 image pairs, where each pair consists of a UAV query image and a corresponding satellite gallery tile [4]. The dataset is split into 80 % training and 20 % validation using a fixed random seed to ensure reproducibility. All models are trained for 80 epochs with batch size 32 using the AdamW optimizer [18] with learning rate 1e-4 and weight decay 1e-4. The embedding dimension is uniformly set to 256 across all experiments. We evaluate retrieval performance using Recall@K, which measures the percentage of queries for which the correct satellite tile appears in the top-K ranked results. All experiments are conducted on an NVIDIA RTX 5090 GPU with 30GB memory.

As shown in Table 1, we present a comprehensive quantitative performance comparison between the two backbone architectures ResNet-18, a classic convolutional neural network, and MLP-Mixer, a pure MLP-based model on the held-out validation dataset, which serves as a rigorous benchmark to evaluate their generalization and cross-view matching capabilities. Across all three core recall metrics, MLP-Mixer consistently outperforms ResNet-18, demonstrating the clear advantage of its architecture for cross-view geo-localization tasks. Most notably, MLP-Mixer achieves a top-1 recall of 74,09 %, which is 1,69 percentage points higher than ResNet-18's R@1 score of 72,40 %. This performance lead is consistent across all evaluated recall thresholds: at R@5, MLP-Mixer reaches 88,41 % compared to ResNet-18 87,50 %, marking a 0,91 percentage point improvement, while at R@10, MLP-Mixer's score of 92,45 % surpasses ResNet-18 90,49 % by a notable 1,96 percentage points. The consistent, across-the-board improvement in all recall metrics strongly indicates that MLP-Mixer's core token-mixing mechanism, which enables holistic global spatial reasoning from the very first network layer, delivers

far more effective cross-view feature alignment than ResNet-18 hierarchical convolutional design. Unlike CNNs, which rely on local receptive fields to extract features in a layer-wise manner, MLP-Mixer directly models long-range spatial relationships across the entire image, allowing it to capture the overall layout of geographic landmarks rather than focusing solely on local texture details sensitive to viewpoint and scale variations. Notably, the performance gap between the two models is most pronounced at R@1, the strictest and most practically relevant metric for real-world geo-localization applications. This significant improvement in top-1 recall demonstrates that MLP-Mixer learns far more discriminative feature embeddings, which are highly effective at distinguishing matching cross-view pairs from non-matching ones, a critical requirement for reliable UAV autonomous navigation systems.

Table 1. **Overall cross-view retrieval performance of ResNet-18 and MLP-Mixer on validation set**

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

To validate the necessity of independent branches for each view, we compare the dual-branch architecture against a variant with shared weights, where a single backbone processes both UAV and satellite inputs. This ablation tests whether the large domain gap between aerial and satellite imagery requires separate feature extraction pathways or if a unified representation space is sufficient. As reported in Table 2, the results reveal a dramatic performance degradation for the shared-weight variant. For ResNet-18, R@1 drops from 72,40 % to 49,35 %, a loss of 23,05 percentage points, indicating that forcing both modalities through a single backbone severely compromises the model ability to learn discriminative cross-view features. Interestingly, MLP-Mixer shared-weight variant shows only marginal degradation, suggesting that token-mixing operations may be more robust to domain shift than CNN local convolutions. However, the independent-branch configuration remains superior for MLP-Mixer as well, confirming that dual-branch architecture is essential for this cross-view task. The asymmetric sensitivity between the two architectures highlights that while MLP-Mixer global receptive field provides some inherent robustness to domain variations, explicit architectural separation for different modalities is still beneficial.

Table 2. **Ablation study: Effect of independent vs shared weight branches on R@1**

Model	Independent Branches	Shared Weights	Δ R@1
ResNet-18	72,40 %	49,35 %	23,0 %
MLP-Mixer	74,09 %	72,79 %	1,30 %

We compare InfoNCE loss with Triplet Loss using hard negative mining (margin=0,3) to verify contrastive learning effectiveness for cross-view geo-localization. InfoNCE computes gradients from all N×N batch pairs, while Triplet Loss uses only one hard negative per anchor, producing sparser gradients. The comparative results are listed in Table 3. For ResNet-18, both losses yield identical 72,40 % R@1, indicating architecture is the main constraint. For MLP-Mixer, InfoNCE reaches 74,09 % R@1, outperforming Triplet Loss by 5,34 percentage points. The temperature τ=0,07 in InfoNCE supports fine-grained similarity calibration, which well matches MLP-Mixer's global token-mixing mechanism. Models with global receptive fields better utilize dense contrastive gradients, making InfoNCE the optimal loss for this task, especially when combined with MLP-Mixer.

Table 3. **Ablation study: InfoNCE vs Triplet Loss (hard negative mining, margin=0.3) for R@1**

Model	InfoNCE	Triplet Loss	Δ R@1
ResNet-18	72,40 %	72,40 %	0 %
MLP-Mixer	74,09 %	68,75 %	5,34 %

To quantify the contribution of ImageNet pretraining, we train both models from random initialization with all other hyperparameters kept identical. As summarized in Table 4, the results demonstrate that ImageNet pretraining is critical for cross-view geo-localization: ResNet-18 drops 13.81 percentage points in R@1, while MLP-Mixer suffers a catastrophic 54,04 percentage point decline. This dramatic performance degradation reveals that semantic features from ImageNet are directly transferable to UAV and satellite imagery. The asymmetric impact shows pure MLP architectures, lacking convolutional inductive biases, depend far more on transfer learning; CNNs' spatial locality and hierarchical feature learning offer better robustness to random initialization. These findings underscore

the necessity of ImageNet pretraining and transfer learning for this task, especially for MLP-based models.

Table 4. **Ablation study: Effect of ImageNet pretraining on R@1 performance**

Model	Pretrained	Random Init	Δ R@1
ResNet-18	72,40 %	58,59 %	13,81 %
MLP-Mixer	74,09 %	20,05 %	54,04 %

This subsection visualizes the top-5 retrieval results for both ResNet-18 and MLP-Mixer on representative queries. As illustrated in Figure 1, green boxes indicate correct matches, while red boxes indicate incorrect matches. This labeling convention makes it straightforward to visually assess each model's retrieval accuracy at a glance. The visualization reveals several key differences in retrieval behavior. MLP-Mixer consistently places the correct satellite tile at rank 1 or 2, while ResNet-18 sometimes ranks it lower, particularly in challenging scenarios with significant viewpoint variations. Both models struggle with extreme viewpoint differences, but MLP-Mixer recovers better in the top-5 results, suggesting more robust feature representations. A critical observation is that MLP-Mixer token-mixing operations appear to capture global spatial layouts and landmark configurations more effectively than ResNet-18 local convolutional features. For instance, in cases where buildings or road networks form distinctive patterns, MLP-Mixer correctly matches these patterns across views, while ResNet-18 may be distracted by local texture variations or shadows. These qualitative results corroborate the quantitative findings: MLP-Mixer global receptive field from the first layer enables superior cross-view matching by focusing on semantically meaningful, viewpoint-invariant features rather than low-level texture details.

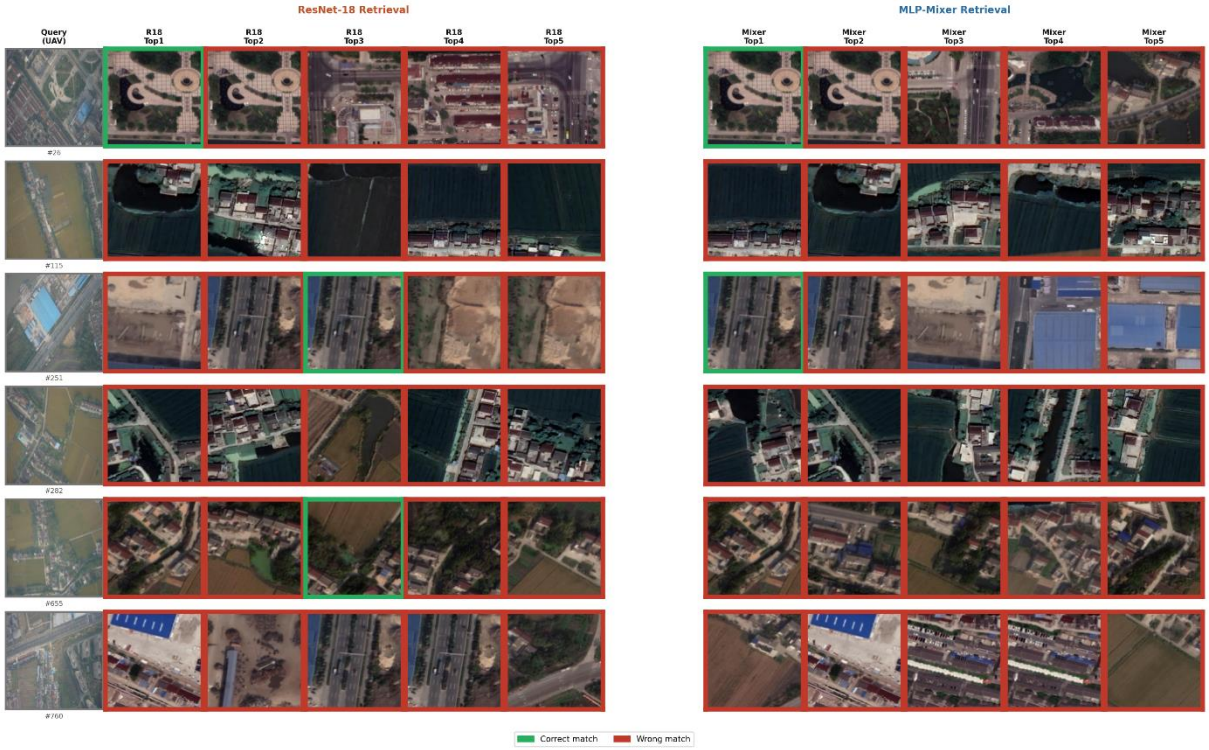

Figure 1. Top-5 UAV-satellite cross-view retrieval results for ResNet-18 and MLP-Mixer. Green boxes denote correct matches, red boxes denote incorrect matches

This subsection visualizes the learned embedding space using t-SNE dimensionality reduction [19]. As displayed in Figure 2, each color represents a unique GPS location. Circles denote UAV query embeddings, and stars denote satellite gallery embeddings. The visualization reveals striking differences in cross-view feature alignment between the two models. For ResNet-18, UAV and satellite embeddings for the same location are scattered across the embedding space, indicating weak cross-view alignment. Many locations show significant spatial separation between circles and stars, suggesting that the model has not learned a unified representation space where matching pairs are close. In contrast, MLP-Mixer demonstrates superior cross-view alignment, with UAV and satellite embeddings clustering tightly

together by location. The clear color-based clustering indicates that the model has learned a unified representation space where matching pairs are proximal in the embedding space. This tight clustering is particularly evident for locations with distinctive landmark configurations, where both UAV and satellite views converge to similar embedding regions. The superior alignment in MLP-Mixer embedding space directly translates to better retrieval performance, as queries and gallery items from the same location are naturally closer in the learned metric space. This visualization provides compelling evidence that MLP-Mixer token-mixing mechanism produces more aligned cross-view embeddings through its global spatial reasoning capabilities.

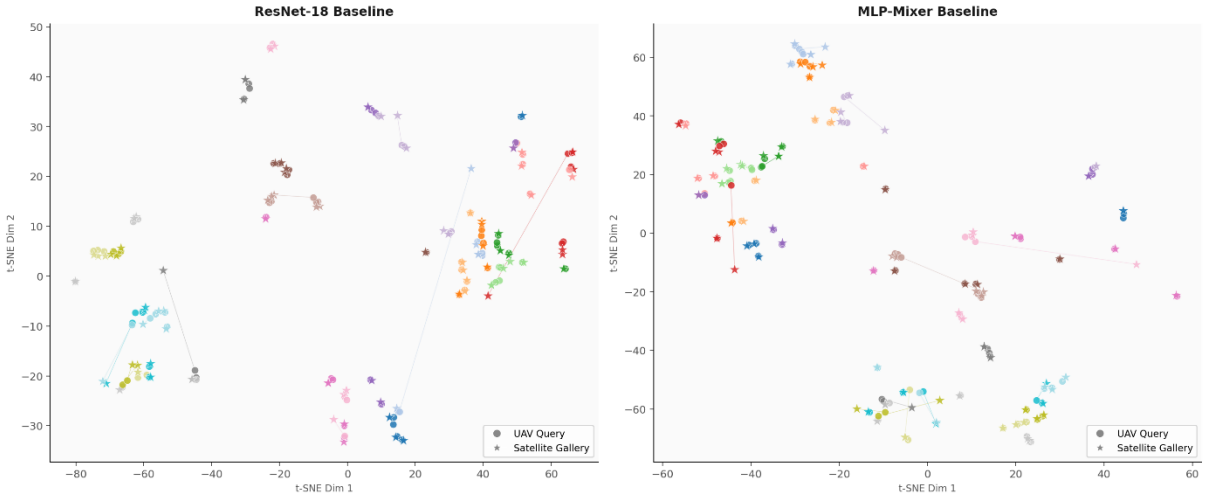

Figure 2. t-SNE visualization of learned cross-view embedding spaces for ResNet-18 and MLP-Mixer. Each color represents a unique GPS location

This subsection presents visualizations of the attention regions for both models using the Grad-CAM method [12], aiming to reveal which image regions contribute most to the generation of the final feature embedding and to deeply analyze the feature extraction mechanism of each network. As demonstrated in Figure 3, the generated attention heatmaps offer clear and intuitive insights into the interpretability of the model decision-making process and the robustness of feature representation, which is crucial for understanding the performance differences between the two architectures in cross-view geo-localization tasks. For ResNet-18, the attention distribution is scattered across multiple irrelevant regions of the image. Instead of focusing on core semantic landmarks that are stable across views, the model often concentrates on low-level textures, local shadows, noise, or non-essential background elements that are highly sensitive to perspective changes. This scattered and unstable attention pattern strongly indicates that ResNet-18 tends to rely on texture-based shallow features, which are vulnerable to variations in viewing angle, imaging scale, and illumination, thus leading to poor feature stability in cross-view matching. In contrast, MLP-Mixer presents highly concentrated attention on high-level semantic structural features, including typical buildings, complete road networks, regular vegetation distribution patterns, and uniquely identifiable landmark configurations. These semantic features possess strong invariance across oblique UAV views and orthogonal satellite views, making them ideal cues for cross-view matching. More importantly, MLP-Mixer attention maps maintain highly consistent focus regions for corresponding UAV and satellite image pairs, which directly proves that the model has successfully learned viewpoint-invariant semantic features during the training process. For instance, when a prominent building or a critical road intersection exists in both the UAV image and the satellite image, MLP-Mixer stably focuses on these key regions in both views, while ResNet-18 attention shifts randomly and fails to align across views. This excellent cross-view attention consistency is the key reason for MLP-Mixer superior retrieval performance: by capturing semantically meaningful and perspective-insensitive features, the model constructs highly robust feature representations that can adapt to the drastic perspective differences inherent in UAV-satellite cross-view geo-localization. Benefiting from the global token-mixing operations inside the architecture, MLP-Mixer is capable of reasoning about global spatial relationships and overall landmark configurations in a holistic manner. Unlike ResNet-18, which is easily disturbed by local texture fluctuations and view-specific imaging

artifacts, MLP-Mixer maintains stable and focused feature perception, further enhancing its reliability and accuracy in cross-view matching tasks.

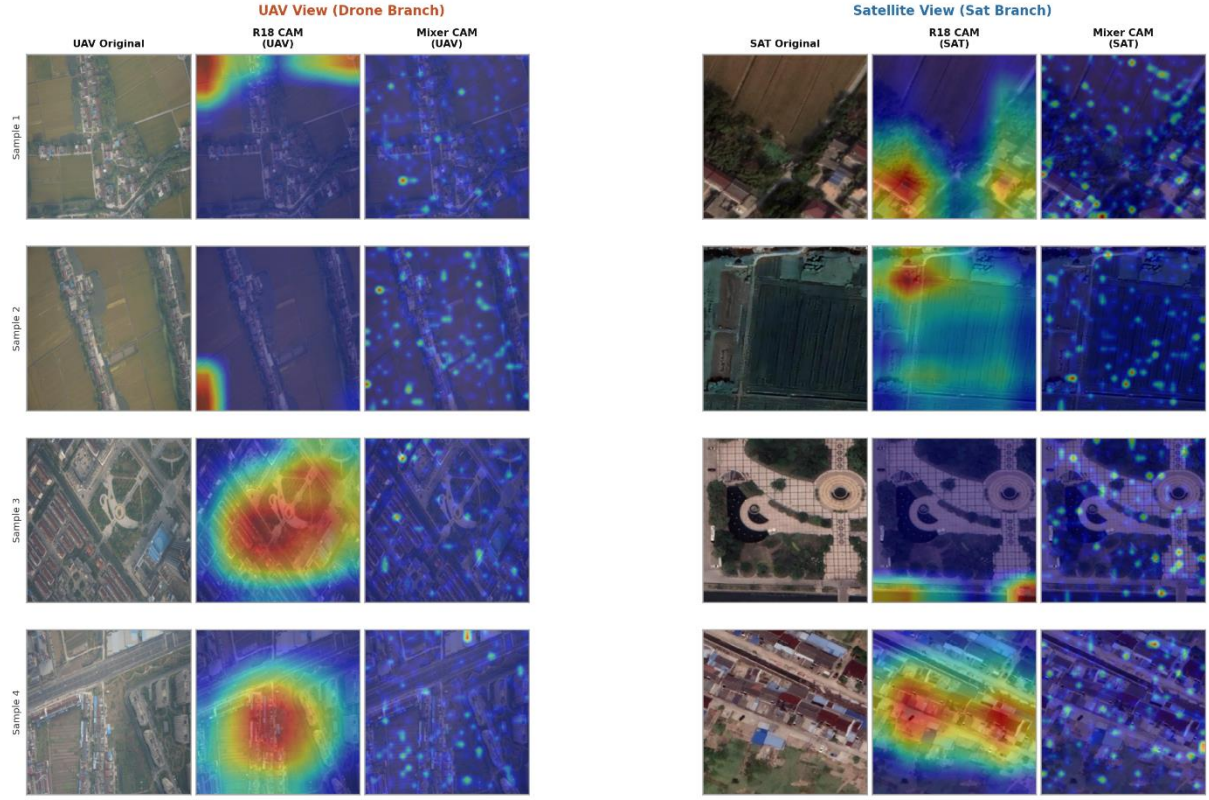

Figure 3. Grad-CAM attention heatmap visualization of ResNet-18 and MLP-Mixer. Warm colors indicate high-attention regions, cool colors indicate low-attention regions

Conclusion

This paper presents a systematic comparative study of dual-branch deep learning architectures for cross-view geo-localization, evaluating ResNet-18 and MLP-Mixer as backbones under a unified contrastive learning framework. Trained on UAV-satellite pairs with bidirectional InfoNCE loss and a shared 256-dimensional embedding space, we show that backbone architecture, loss function, weight-sharing, and pretraining each play distinct, measurable roles in retrieval performance.

Our experiments confirm MLP-Mixer [5] consistently outperforms ResNet-18 [16] across all metrics: R@1 of 74,09 % vs. 72,40 %, R@5 of 88,41 % vs. 87,50 %, and R@10 of 92,45 % vs. 90,49 %. This advantage stems from MLP-Mixer's global token-mixing mechanism, which enables full-image spatial reasoning from the first layer critical for cross-view matching, where landmark layouts, not local textures, are the primary matching cue.

The ablation study reveals three key design principles: First, independent UAV satellite branches are essential, as shared weights cause a catastrophic 23,05 percentage point R@1 drop for ResNet-18, confirming the large domain gap between aerial and satellite imagery requires dedicated feature pathways. Second, InfoNCE loss significantly outperforms Triplet Loss for MLP-Mixe, as dense contrastive gradients from all batch pairs provide richer learning signals than single hard-negative mining [4]. Third, ImageNet pretraining is critical, especially for MLP-Mixer, whose R@1 collapses from 74,09 % to 20,05 % without initialization, a 54,04 percentage point drop, highlighting transfer learning's necessity for inductive bias-free pure MLP architectures.

Qualitative analyses validate these findings: t-SNE visualizations [19] confirm tighter cross-view alignment with compact same-location embedding clusters, Grad-CAM maps [20] show MLP-Mixer focuses on semantic, viewpoint-invariant regions while ResNet-18 attends to view-sensitive low-level textures, and retrieval visualizations demonstrate higher-ranked correct matches for MLP-Mixer, especially in extreme viewpoint scale scenarios.

Despite these results, limitations remain: the 768-image dataset is small, performance on larger diverse datasets requires validation, the framework lacks GPS and temporal metadata for real-world deployment, and MLP-Mixer 59,9M parameter count is far higher than ResNet-18's 11,2M, posing constraints for resource-limited UAV systems.

Future work will focus on three directions: scaling to larger cross-view datasets to test generalizability, integrating lightweight attention/knowledge distillation to reduce MLP-Mixer computational cost while preserving performance, and extending to multi-modal fusion with GPS priors and flight trajectories for robust real-time UAV geo-localization.

In summary, this work provides a rigorous empirical foundation for cross-view geo-localization performance factors, establishing MLP-Mixer with bidirectional InfoNCE loss and independent dual branches as a strong baseline for future research.

References

1.°Shi Y., Liu L., Yu X., Li H. // Advances in Neural Information Processing Systems (NeurIPS). 2019. Vol. 32. P. 10090–10100.

2.°Zhu S., Shah M., Chen C. // Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR). 2022. P. 1162–1171.

3.°Zheng Z., Wei Y., Yang Y. // Proc. ACM Int. Conf. Multimedia (ACM MM). 2020. P. 1395–1403.

4.°Deuser F., Habel K., Oswald N. // Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV). 2023. P. 16847–16856.

5.°Tolstikhin I., Houlsby N., Kolesnikov A., Beyer L., Zhai X., Unterthiner T., Yung J., Steiner A., Keysers D., Uszkoreit J., Lucic M., Dosovitskiy A. // Advances in Neural Information Processing Systems (NeurIPS). 2021. Vol. 34. P. 24261–24272.

6.°Workman S., Zhai M., Crandall D., Jacobs N. Wide-area image geolocalization with aerial reference imagery // Proc. IEEE Int. Conf. Computer Vision (ICCV). 2015. P. 3961-3969.

7.°Zhu S., Shah M., Chen C. VIGOR: Cross-view image geo-localization beyond one-to-one retrieval // Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR). 2021. P. 3640-3649.

8.°Lowe D. G. Distinctive image features from scale-invariant keypoints // Int. J. Comput. Vis. 2004. Vol. 60. No. 2. P. 91-110.

9.°Dalal N., Triggs B. Histograms of oriented gradients for human detection // Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR). 2005. Vol. 1. P. 886-893.

10.°Arandjelović R., Gronat P., Torii A., Pajdla T., Sivic J. NetVLAD: CNN architecture for weakly supervised place recognition // Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR). 2016. P. 5297-5307.

11.°Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. // Proc. Int. Conf. Learning Representations (ICLR). 2021.

12.°Schroff F., Kalenichenko D., Philbin J. // Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR). 2015. P. 815–823.

13.°Chen T., Kornblith S., Norouzi M., Hinton G. // Proc. Int. Conf. Machine Learning (ICML). 2020. Vol. 119. P. 1597–1607.

14.°van den Oord A., Li Y., Vinyals O. // arXiv preprint arXiv:1807.03748. 2018.

15.°Radenović F., Tolias G., Chum O. // IEEE Trans. Pattern Anal. Mach. Intell. 2019. Vol. 41. No. 7. P. 1655–1668.

16.°He K., Zhang X., Ren S., Sun J. // Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR). 2016. P. 770–778.

17.°Deng J., Dong W., Socher R., Li L.-J., Li K., Fei-Fei L. // Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR). 2009. P. 248–255.

18.°Loshchilov I., Hutter F. // Proc. Int. Conf. Learning Representations (ICLR). 2019.

19.°van der Maaten L., Hinton G. // J. Mach. Learn. Res. 2008. Vol. 9. P. 2579–2605.

20.°Selvaraju R. R., Cogswell M., Das A., Vedantam R., Parikh D., Batra D. // Proc. IEEE Int. Conf. Computer Vision (ICCV). 2017. P. 618–626.