Науковий вісник НЛТУ України
Scientific Bulletin of UNFU
https://doi.org/10.36930/40350615
Article received 24.10.2025
Article accepted 22.12.2025
UDC 004.[8+93+932]
Correspondence author: R. I. Ilechko roman.i.ilechko@lpnu.ua
ISSN 1994-7836 (print)
ISSN 2519-2477 (online)

R. I. Ilechko, Yu. V. Tsymbal
Lviv Polytechnic National University, Lviv, Ukraine

EFFICIENT CROSS-VIEW MATCHING OF IMAGES CAPTURED BY UAV CAMERAS

We address the problem of visual geo-localization for unmanned aerial vehicles operating under GNSS,0 degradation and study how Sample4Geo-type cross-view methods can be deployed on onboard hardware with limited resources. We analyze the behavior of lightweight feature extractors within the Sample4Geo framework and investigate how systematic backbone downscaling, together with training configuration and input image resolution, affects the accuracy of cross-view geo-localization. We introduce a unified evaluation protocol on a standard cross-view dataset and train a set of convolutional and transformer architectures (ConvNeXt-Base/Nano, ViT-Small/Tiny, MobileNetV3/V4, LeViT) under shared baseline settings at 224×224 resolution, followed by a targeted hyperparameter search for each backbone. We further examine resolution scaling by re-training the tuned models at 384×384. We construct an efficiency frontier using Recall@1 and the number of trainable parameters as axes. Our results show that compact backbones already provide competitive accuracy with a simple single-epoch training scheme, while hyperparameter tuning yields a substantial additional gain for MobileNetV4 and ViT-Small and confirms that an almost uniform training schedule is sufficient for ConvNeXt-Nano, ViT-Tiny, and MobileNetV3. We establish that increasing input resolution consistently improves all models: ConvNeXt-Nano approaches the accuracy of ConvNeXt-Base while remaining significantly smaller, and ViT-Small surpasses ViT-Tiny when richer spatial detail is available. We construct the Pareto frontier and identify MobileNetV3 and ViT-Tiny as attractive options at very low capacity, while highlighting ConvNeXt-Nano as a knee point that offers a favorable compromise between accuracy and model size for onboard deployment. In comparative experiments against recent cross-view baselines, including LPN, LCM, SAIG-D, DWDR, MBF, and MSBA, the proposed lightweight ConvNeXt-based configurations consistently outperform or match prior state-of-the-art methods under comparable or lower model complexity. Finally, we provide practical recommendations for selecting feature extractors and training settings in resource-constrained cross-view geo-localization.

Keywords: deep neural network; transformers; convolution neural network; computer vision; visual geolocalisation.

Introduction / Вступ

Vision-based geo-localization has become a core enabling technology for Unmanned Aerial Vehicles (UAVs) operating in GPS-denied or degraded settings. Modern cross-view pipeline, exemplified by Sample4Geo, combine contrastive learning with strong feature extractors to match UAV imagery to overhead maps despite viewpoint, scale, and appearance gaps. While such methods have advanced the state of the art, their deployment on embedded flight hardware remains constrained by model size, inference latency, and sensitivity to input resolution. In practice, UAV processors must balance accuracy with limited compute and memory budgets, making the choice of backbone and image resolution as consequential as the learning objective itself.

Research gap. Prior studies primarily benchmark Sample4Geo (and related pipelines) with medium-to-large backbones and fixed input sizes. Systematic evidence on how smaller feature extractors and scaled input resolutions jointly affect geo-localization quality is limited. Moreover, practical, end-to-end guidance on selecting training and inference settings that leverage lightweight networks to approach the utility of larger models, while balancing performance and size, remains underdeveloped.

Problem statement. There is a need for a principled recipe that identifies feature extractor size and resolution configurations delivering strong geo-localization under tight resource constraints and quantifies the trade-offs among accuracy, and model complexity within the Sample4Geo framework.

Object of research – comparison of images from the UAV with existing maps, given limited computing resources.

Subject of research – methods and design choices that reduce localization error and computational cost in Sample4Geo-style pipelines, with emphasis on feature-extractor capacity, training hyperparameters, and input-image resolution at inference.

The purpose of the research – to develop and validate a compact Sample4Geo configuration that delivers competitive localization accuracy at substantially lower computational cost by jointly optimizing backbone size, training hyperparameters, and input resolution.

To achieve this purpose, the following main *research objectives are identified:*

Інформація про авторів:
Ілечко Роман Ігорович, магістр, аспірант, кафедра автоматизовані систем управління. **Email:** roman.i.ilechko@lpnu.ua; https://orcid.org/0009-0006-8208-5109
Цимбал Юрій Вікторович, канд. техн. наук, доцент, кафедра автоматизовані систем управління. **Email:** yurii.v.tsymbal@lpnu.ua; https://orcid.org/0000-0001-9119-6771
Цитування за ДСТУ: Ілечко Р. І., Цимбал Ю. В. Ефективне міжракурсне зіставлення зображень, отриманих з камер БПЛА. Науковий вісник НЛТУ України. 2025, т. 35, № 6. С. 00–00.
Citation APA: Ilechko, R. I., & Tsymbal, Yu. V. (2025). Efficient cross-view matching of images captured by UAV cameras. *Scientific Bulletin of UNFU*, 35(6), 123–129. https://doi.org/10.36930/40350615

- establish a baseline Sample4Geo configuration and evaluation protocol for fair cross-model comparison, which will enable reproducible comparison across backbones, training setups, and input resolutions;
- systematically downscale feature-extractor capacity and quantify its isolated effect on geo-localization quality and memory usage, which will make it possible to disentangle the impact of network size from other factors;
- perform hyperparameter search to improve models baseline metrics, which will ensure that each backbone configuration is reasonably well tuned and that the observed performance are not artifacts of suboptimal training;
- systematically study the impact of higher input resolutions on lightweight extractors, to determine whether increased resolution improves localization performance;
- model the performance-efficiency frontier, reporting accuracy, and parameter counts to identify Pareto-optimal extractor-resolution configurations, yielding a set of optimal options for UAVs with different compute and memory budgets.

Analysis of recent research and publications. Unmanned aerial vehicles operating without reliable GPS/GNSS have moved from niche challenge to mainstream urgency: contested airspace, urban canyons, forests, and indoors routinely degrade or deny satellite signals, and the community has responded with a surge of methods that fuse cameras, IMUs, lidar, semantics, and learning. This literature is evolving fast along several complementary tracks, visual-inertial SLAM for onboard state estimation, cross-view geo-localization to find local motion to global frames, and multi-sensor fusion for robustness under dynamics and compute constraints. In research [2], a feature-based V-SLAM method formalizes map-point reliability, culls geometrically implausible points, and removes duplicates via descriptor checks. Coupled with an unreliable-map-point check and Delaunay-based selection to promote temporary points, this approach sustains real-time robustness in dynamic scenes and offers a hardware-light fallback to cross-view matching.

In study [9], the authors propose CVM-Net as an early deep-learning baseline for cross-view geo-localization, formulating the task as metric learning with a Siamese architecture that extracts local CNN features and aggregates them into global descriptors. The NetVLAD [1] layer used in CVM-Net is introduced as a learnable aggregation module that replaces hand-crafted pooling with trainable vector quantization, two variants of the architecture are presented: CVM-Net-I, with independent NetVLAD branches for each view, and CVM-Net-II, with partially shared transformations and a shared NetVLAD, both optimized using a weighted soft-margin ranking loss to accelerate convergence. The approach is shown to achieve strong retrieval accuracy on standard street to satellite benchmarks, including an analysis of loss functions and weight-sharing strategies.

In work [14] researchers introduces the SAFA method for cross-view geo-localization, explicitly modeling geometric correspondences between aerial and ground views. The authors first apply a polar transform to aerial images to coarsely align them with ground panoramas and then aggregate features using a spatial-aware attention module that embeds layout information into the global descriptor. The method is trained with a triplet objective and exhaustive in-batch mining and achieves large gains on standard cross-view benchmarks while relying on a comparatively simple backbone and no geometry-specific supervision at training time. Overall, the paper positions its contribution as decoupling geometry alignment from feature learning, thereby simplifying optimization and improving retrieval accuracy.

From another hand, in [16] scientist tackles cross-view geo-localization by explicitly leveraging contextual cues around the target, arguing that most prior work over-emphasized the central object and under-used surrounding patterns that remain informative across viewpoints. The scientist proposes a simple encoder with a square-ring feature partition that organizes features by distance to the image center, enabling end-to-end part-wise learning without extra part estimators and with robustness to rotation. It evaluates the approach on standard cross-view benchmarks and shows that injecting contextual parts yields competitive retrieval accuracy and can be fused into other pipelines. The study includes drone satellite tasks and LPN is tested under those settings. Methodologically, proposed approach is a multi-branch encoder whose aerial branches share weights, features are sliced into concentric parts, pooled, and optimized with per-part classification losses.

Study [21] proposes TransGeo as a transformer-only architecture for cross-view image geo-localization that dispenses with polar transforms and CNN-specific aggregation, arguing that global self-attention and learnable positional embeddings better bridge ground to aerial domain gaps. The method operates in two stages: an initial baseline training with a soft-margin triplet loss, followed by an attention-guided non-uniform cropping scheme that discards uninformative aerial regions and reallocates computation to higher resolution on salient areas ("attend and zoom-in"). The approach is evaluated on the aligned CVUSA dataset [18] and the unaligned VIGOR dataset [22], where it achieves strong retrieval performance with reduced computation compared to representative CNN-based pipelines. A dedicated regularization strategy is also introduced to stabilize transformer training without relying on heavy data augmentation.

In opposite to this, in the [3] authors addresses cross-view geo-localization by replacing heavy, multi-stage pipelines with a streamlined, contrastive framework built around a single, weight-shared encoder and a symmetric InfoNCE objective. The central contribution is hard-negative mining via two complementary strategies: early GPS-based neighbor sampling and Dynamic Similarity Sampling based on embedding similarity, reducing reliance on polar transforms or bespoke aggregation modules. The method is evaluated across established cross-view datasets and demonstrates strong generalization with a simplified training recipe. Conceptually, the paper positions CNN backbones ConvNeXt [15] as competitive with recent transformer/mixer designs when combined with effective sampling and symmetric losses. At the same time, the evaluation largely considers medium-to-large backbones with dataset-specific fixed input sizes, without a controlled study of how extractor capacity and inference resolution jointly influence accuracy and efficiency. Moreover, while accuracy is reported comprehensively, detailed memory trainable parameters number and a performance-efficiency analysis are not emphasized, leaving open questions about principled model selection for compute-constrained UAV deployments.

Research [5] tackles cross-view geo-localization by reducing the ground-overhead domain gap via novel-view synthesis. It reconstructs a scene from street-view sequences with 3D Gaussian Splatting, then renders perspective-aware, tilted aerial views that are easier to match to satellite imagery. The synthesized views are paired with a DINOv2 [12] –based feature pathway and optimized using InfoNCE for retrieval.

Manuscript [10] addresses UAV-satellite cross-view image matching by introducing an Adaptive Threshold-guided Ring Partitioning Framework that adapts local feature regions to image content. The approach augments a dual-branch CNN with brightness-alignment preprocessing to reduce cross-view style gaps, (heatmap-guided, learnable ring partitioning for context-aware local features, (a hybrid loss combining cross-entropy and ArcFace, ands post-hoc keypoint-based re-ranking of top candidates. The method reports competitive retrieval quality against contemporary baselines.

Taken together, the reviewed studies demonstrate rapid progress in cross-view geo-localization, from geometry-aware aggregation and contextual partitioning to transformer-based architectures and streamlined contrastive frameworks. However, none of these works provide a systematic analysis of lightweight Sample4Geo-style backbones under strict UAV resource constraints: feature-extractor capacity is not jointly varied with input resolution, hyperparameter regimes for compact models are not thoroughly explored, and performance-efficiency trade-offs are not formalized in terms of trainable parameters and accuracy. This gap motivates the present study, which focuses on controlled downscaling of feature extractors, resolution scaling and Pareto analysis to identify extractor-resolution configurations that are suitable for computationally constrained UAV systems and to clarify the associated performance-efficiency trade-offs.

Materials and methods of research. This study investigates Sample4Geo under resource-constrained conditions. We base our study on the publicly available University-1652 UAV imagery collection [19]. Primary metrics are Recall@1/5/10, AP. Efficiency is characterized by parameter count. All experiments were executed on a single workstation with an NVIDIA RTX 3090 Ti GPU.

Research results and their discussion / Результати дослідження та їх обговорення

Initial experiments. Phase of initial experiments establishes a unified, fast-probe baseline across all backbones under identical constraints. For every model, inputs are fixed to 224×224, training runs for one epoch with batch size = 16, and optimization uses Adam W. [11] ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $wd = 0.05$) with an initial learning rate 3×10^{-4}, 10 % linear warm-up, and cosine decay to 1×10^{-6}. Mixed-precision (AMP/fp16) is enabled. Models are evaluated on the validation split and the checkpoint is selected by best Recall@1. We report Recall@{1,5,10}, and AP, to provide like-for-like reference points before any hyperparameter tuning or resolution scaling. This protocol prioritizes comparability and peed, making the results a clean baseline against which subsequent improvements are measured.

Table 1. Initial models hyperparameter settings / Початкові налаштування гіперпараметрів моделі

Setting	Value
Epochs	1
Batch size	16
Optimizer	AdamW
Learning rate	η=3 *e*-4
Precision	AMP (fp16)

Table 2 summarizes the initial performance metrics for a representative set of lightweight backbone models evaluated under a uniform training configuration. All models were trained for one epoch with fixed input resolution (224×224) and default hyperparameters. The metrics reported include Recall@1, Recall@5, Recall@10, and Average Precision (AP), computed on a held-out validation set. Among the evaluated models, ConvNeXt-Nano achieved the highest Recall@1 (83.84 %) and AP (86.36 %), followed by ViT-Tiny (Recall@1 = 76.77 %, AP = 79.91 %). MobileNetV4 and ViT-Small exhibited comparable performance in the mid-60 s to low-70 s across Recall and AP metrics. LeViT yielded substantially lower results across all measures, with Recall@1 of 14.27 % and AP of 18.74 %, indicating limited performance under the default setup.

Table 2. Models performance after training with initial settings / Показники моделей після тренування з початковими налаштуваннями

Model	Recall@1	Recall@5	Recall@10	AP
MobileNetV4 [13]	65.55	85.14	90.05	69.93
ViT small [6]	67.40	86.25	90.96	71.62
ConvNext nano [5]	83.84	95.08	96.89	86.36
LeViT [7]	14.27	31.01	41.05	18.74
ViT tiny [6]	76.77	91.53	92.64	79.91
MobileNetV3 [8]	68.69	85.92	91.23	71.08

Hyperparameters search. The goal of this stage was to identify training configurations that improve performance without increasing model size or input resolution. A limited hyperparameter search was conducted per model, adjusting epochs, batch size, and initial learning rate to optimize accuracy under constrained compute settings.

Table 3. Models performance after hyperparameter optimization / Показники моделей після тренування з оптимізованим налаштуваннями гіперпарематра

Model	Recall@1	Recall@5	Recall@10	AP
MobileNetV4	74.46	90.43	93.90	78.06
ViT small	75.28	90.93	94.04	78.79
ConvNext nano	86.34	95.70	97.21	88.46
LeViT	14.27	31.01	41.05	18.74
ViT tiny	78.27	91.88	94.70	81.33
MobileNetV3	69.70	87.81	92.22	73.71

Table 3 presents the performance of the same set of lightweight backbone models after applying targeted hyperparameter optimization. The evaluation metrics remain consistent with the previous setup and reported on the same validation split. Following parameter search, ConvNeXt-Nano maintained its leading position, improving Recall@1 to 86.34 % and AP to 88.46 %, further extending its margin over other models. ViT-Tiny exhibited notable gains across all metrics, achieving Recall@1 of 78.27 % and AP of 81.33 %, overtaking ViT-Small, which reached 75.28 % Recall@1 and 78.79 % AP. MobileNetV4 also improved substantially (Recall@1 = 74.46 %, AP = 78.06 %), ranking just behind the ViT variants. MobileNetV3 showed moderate improvement, while LeViT-128 s remained unchanged, indicating limited sensitivity to the explored hyperparameter configurations. Before optimization, ConvNeXt-Nano led the performance rankings across all metrics, followed by ViT-Tiny, ViT-Small, and MobileNetV3, with MobileNetV4 trailing slightly behind. After parameter tuning, the performance of most models improved. ConvNeXt-Nano retained its top position, showing further gains in both Recall@1 and AP. MobileNetV4 showed notable improvement, moving ahead of MobileNetV3, which had more modest gains. The ranking shift highlights the impact of hyperparameter search on model effectiveness, particularly for transformer-based architectures and newer convolutional designs, while also exposing stability or limitations in specific models under the same resource conditions.

Table 4. Models hyperparameters after optimization / Гіперпараметри моделі після оптимізації

Model	Epochs	Batch size	Initial LR
MobileNetV4	1	64	3e-4
ViT small	2	32	1e-4
ConvNext nano	1	64	3e-4
LeViT	1	64	1e-3
ViT tiny	1	64	1e-4
MobileNetV3	1	64	3e-4

Table 4 lists the optimized training configurations for each model, including the number of training epochs, batch size, and initial learning rate. These settings were selected through a targeted parameter search aimed at improving geo-localization performance under fixed resolution and resource constraints.

Input Resolution Scaling. To further explore the performance limits of compact models, we conducted a series of experiments with increased input resolution while keeping all other settings fixed. This phase investigates whether resolution scaling can compensate for reduced model capacity and enhance localization accuracy under constrained conditions.

Table 5. Models performance with 384×384 image input / Показники моделі з вхідним зображенням розміром 384×384

Model	Recall@1	Recall@5	Recall@10	AP
ConvNext base	92.37	97.46	98.33	93.55
ConvNext nano	90.51	97.21	98.03	98.12
ViT small	82.14	94.06	96.11	96.33
ViT tiny	81.47	94.25	96.36	96.0
MobileNetV3	78.67	92.03	94.67	94.93

Table 5 reports the performance of selected models under increased input resolution, with all other optimal training settings for each model, allowing comparison of resolution-driven improvements across architectures. ConvNeXt-Nano approaches the performance of the larger ConvNeXt-Base model, achieving competitive scores at a fraction of the size. Both ViT-Small and ViT-Tiny benefit from resolution scaling, with ViT-Small slightly outperforming ViT-Tiny across all metrics. MobileNetV3 also shows a notable performance boost, though it remains behind the transformer-based models in this configuration.

Table 6. Performance-capacity summary of evaluated backbones / Підсумок продуктивності розглянутих архітектур

Model	Trainable parameters	Recall@1
ConvNext base	88.5 M	92.37
ConvNext nano	15.59 M	90.51
ViT small	22.88 M	82.14
ViT tiny	5.72 M	81.47
MobileNetV3	5.48 M	78.67

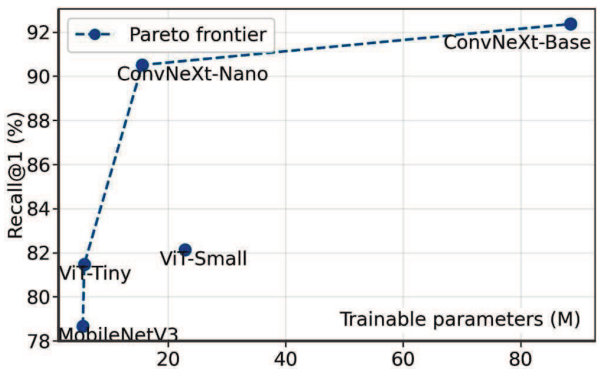

Fig. Pareto frontier: Recall@1 vs. model size / Межа Парето: Recall@1 проти розміру моделі

Figure, displays the Recall@1 vs. number of trainable parameters for the evaluated backbones. The line indicates the Pareto frontier, models lying on this curve are not dominated in terms of accuracy-capacity trade-off. ConvNeXt-Nano represents a knee point on the frontier, delivering high Recall@1 at substantially lower capacity than ConvNeXt-Base.

Table 7. Comparison to SOTA approaches / Порівняння з сучасними передовими методами

Model	Recall@1	AP
ConvNext base	92.37	93.55
ConvNext nano	90.51	98.12
ViT tiny	81.47	96.0
MobileNetV3	78.67	94.93
LPN [21]	75.93	79.14
LCM [4]	66.65	70.82
SAIG-D [23]	78.85	81.62
DWDR [17]	86.41	88.41
MBF [20]	89.05	90.61
MSBA [24]	86.61	88.55

Table 7 summarizes the comparison between the proposed configurations and existing state-of-the-art approaches (LPN, LCM, SAIG-D, DWDR, MBF, MSBA) using Recall@1 and AP as evaluation metrics. ConvNeXt-Base attains the highest Recall@1, while ConvNeXt-Nano achieves a slightly lower Recall@1 and AP. ViT-Tiny and MobileNetV3 also reach higher Recall@1 and AP than LPN and SAIG-D, and DWDR and MBF occupy intermediate positions between MobileNetV3/ViT-Tiny and the ConvNeXt variants in terms of both metrics.

The comparison between Tables 2 and 3 shows that all models except LeViT benefit from even a hyperparameter search, but the magnitude of improvement varies notably by architecture. Under the initial settings, ConvNeXt-Nano already dominates other lightweight models, with the highest Recall@1 and AP, followed by ViT-Tiny and then ViT-Small, MobileNetV3, and MobileNetV4. After tuning, ConvNeXt-Nano still holds the leading position but gains a further ~2-point increase in Recall@1 and AP, indicating that the default recipe was already close to its natural operating point and only needed minor refinement. In contrast, MobileNetV4 and ViT-Small show much larger absolute gains in Recall@1 and AP, suggesting that their baseline configuration under-exploited their capacity. ViT-Tiny improves more modestly, remaining ahead of MobileNetV3 and ending up as a strong "second tier" option beneath ConvNeXt-Nano. LeViT remains at essentially the same performance level before and after tuning, which points to an architectural or pretraining mismatch with the Sample4Geo objective rather than purely suboptimal hyperparameters.

Table 4 helps explain these patterns by revealing that most backbones converge best under very similar training regimes: a single epoch, batch size 64, and initial learning rate $3 \cdot 10^{-4}$ work well for ConvNeXt-Nano, MobileNetV3, and MobileNetV4, and serve as a reasonable choice for ViT-Tiny as well. This consistency suggests that, for this dataset and contrastive loss, a simple, unified training recipe is sufficient for a wide range of compact CNN and small-transformer architectures. The main exception is ViT-Small, which prefers a slightly gentler schedule – two epochs, a smaller batch size, and a lower initial learning rate – consistent with larger ViT variants typically requiring more optimization steps and more conservative updates to

reach good generalization. Overall, these results indicate that hyperparameter tuning can meaningfully reshuffle the ranking among mid-range models (especially MobileNetV4 and ViT-Small), while the strongest and weakest architectures, ConvNeXt-Nano and LeViT, remain at the respective ends of the performance spectrum.

Table 5 shows that increasing the input resolution to 384×384 boosts backbones, but the notion of an "optimal" model becomes ambiguous. ConvNeXt-Base attains the highest Recall@1, remaining the strongest option if parameters and compute are unconstrained. However, ConvNeXt-Nano now approaches it very closely in Recall@1 while achieving the highest AP of all models, which indicates more consistent ranking quality across candidates despite much lower capacity. ViT-Small and ViT-Tiny both benefit from the larger input, with ViT-Small slightly overtaking ViT-Tiny at this resolution, suggesting that additional spatial detail is better exploited by the higher-capacity transformer. MobileNetV3 also improves but remains behind the ConvNeXt and ViT variants. Taken together, the results at 384×384 do not point to a single universally optimal backbone. ConvNeXt-Base is preferable when peak Recall@1 is the only objective, ConvNeXt-Nano offers a near-SOTA Recall@1 with superior AP and substantially fewer parameters, and ViT-Small/Tiny provide intermediate options that balance transformer-style representations with moderate resource use.

Table 6 and Figure summarize the trade-off between model capacity and geo-localization accuracy using Recall@1 as the primary performance indicator and the number of trainable parameters as a proxy for complexity. The Pareto frontier connects the non-dominated configurations: MobileNetV3 at the extreme low-capacity end, ViT-Tiny as a slightly larger but clearly more accurate option, ConvNeXt-Nano as a mid-size high-accuracy model, and ConvNeXt-Base as the highest-accuracy but heaviest backbone. ViT-Small is strictly dominated by ConvNeXt-Nano, which achieves higher Recall@1 with fewer parameters, and therefore does not lie on the frontier.

Within this set of Pareto-optimal points, ConvNeXt-Nano emerges as the most balanced configuration. It achieves Recall@1 very close to ConvNeXt-Base while using roughly six times fewer parameters, and at the same time it offers a substantial accuracy margin over ViT-Tiny and MobileNetV3 for a moderate increase in capacity. In other words, ConvNeXt-Nano sits near the "knee" of the Pareto curve, where additional parameters above its size bring only marginal accuracy gains, whereas reducing capacity below this point leads to a disproportionate drop in Recall@1. For resource-constrained UAV scenarios, this makes ConvNeXt-Nano a natural default choice when both performance and model size must be taken into account.

Discussion of research results. Study [4] proposes a Location Classification based Matching approach for UAV to satellite cross-view geo-localization, targeting the challenging viewpoint gap on the University-1652 dataset. The authors reformulate cross-view retrieval as a location-classification task, explicitly addressing the imbalance between the large number of satellite images and the comparatively few UAV-view samples, and treating similarity between UAV and satellite images through a shared feature space. Within this framework, the retrieval stage is simplified to a classification problem during training. In this comparison, LCM [4] yields the lowest Recall@1 and AP among the evaluated methods, indicating that its location-classification formulation with relatively simple feature modeling is less effective than modern metric-learning pipelines. While model remains conceptually straightforward and practically appealing, the results suggest that it does not fully exploit the representational capacity of recent backbones.

In [23] tackles the need for a backbone that is tailored to cross-view geo-localization but remains architecturally simple and computationally efficient. The authors introduce Simple Attention-based Image Geo-localization network, which replaces heavy, multi-stage pipelines with a compact network that models long-range dependencies and cross-view relations using multi-head self-attention, while a shallow convolutional stem preserves local structure without aggressive patchification. The backbone follows a "slim but deep" design to increase representational power without excessive cost and is intended to work without explicit feature-alignment modules or handcrafted aggregation blocks. In our evaluation on University-1652, SAIG-D reaches Recall@1 and AP that are clearly higher than those of earlier methods such as LPN and LCM, but it is still outperformed by all proposed ConvNeXt-based and ViT-Tiny configurations. In particular, ConvNeXt-Base and ConvNeXt-Nano achieve higher Recall@1 and substantially stronger AP than SAIG-D, indicating that, despite SAIG's architectural simplicity and good reported results, modern CNN and ViT backbones with appropriate training and resolution scaling offer a more favorable accuracy-capacity trade-off in this setting.

In [17] addresses cross-view geo-localization by focusing on a largely overlooked aspect: redundancy within the embedding itself rather than only distances between embeddings. The authors propose Dynamic Weighted Decorrelation Regularization (DWDR), a simple regularizer that encourages different channels of the learned descriptor to carry complementary information by driving the correlation matrix of the embedding toward an identity-like, sparse structure. The "dynamic" component assigns higher weights to channels that remain strongly correlated during training, allowing the network to progressively decorrelate only the most redundant components instead of penalizing all channels uniformly. In our evaluation, DWDR reaches 86.41 % Recall@1 and 88.41 % AP, clearly outperforming earlier methods such as LPN (75.93 % / 79.14 %) and LCM (66.65 % / 70.82 %), as well as SAIG-D (78.85 % / 81.62 %). However, DWDR still lags behind the proposed ConvNeXt-Base and ConvNeXt-Nano configurations (92.37 % / 93.55 % and 90.51 % / 98.12 % respectively), indicating that while embedding decorrelation is beneficial, it still far behind current SOTA.

In [20] addresses UAV visual geo-localization by explicitly modeling how additional status information and multi-scale scene structure can help align drone and satellite views. The authors propose MBF, a feature-fusion network that combines image embeddings with UAV status encoded as word-like vectors, which are injected into a transformer block to obtain representations that are aware of the platform's state. In parallel, the method leverages both global and local image regions from the same viewpoint and couples them via hierarchical bilinear pooling, so that fine-grained local cues and broader scene context reinforce each other in the final descriptor. MBF attains around 89.05 % Recall@1 with an AP of 90.61. At the same time, its Recall@1 and AP still remain below ConvNeXt-Base and close to ConvNeXt-Nano approaches.

In [24] targets the challenge of matching UAV and satellite views under strong changes in scale and target position, and argues that many existing methods underuse multiscale structure and cross-view relationships. The authors build on a local pattern network and introduce Multiscale Block Attention, which splits each image into patches at several scales and applies self-attention within and across these blocks to improve feature extraction at different spatial extents. A multi-branch architecture is used to process different views, with the branches designed to interact so that each stream can benefit from information coming from the other viewpoint, encouraging a more consistent cross-view representation. MSBA emerges as the one of the strongest of the previously published baselines, with Recall@1 and AP values that lie just below those of ConvNeXt-Base (92.37 % / 93.55 %) yet clearly above ViT-Tiny (81.47 % / 96.0 %). It is therefore the close competitor to the proposed ConvNeXt-Base configuration while still outperforming transformer-based ViT backbones.

Taken together, the discussed results confirm the relevance and necessity of the proposed study. While prior work has substantially advanced cross-view geo-localization through new architectures, feature fusion strategies, and regularization schemes, none of these methods systematically explore lightweight Sample4Geo-style backbones under strict UAV resource constraints, jointly varying backbone capacity, training regime, and input resolution. The presented experiments, including hyperparameter tuning, resolution scaling, and Pareto-style analysis of accuracy versus model size, therefore fill a clear gap by revealing where compact models can approach or match heavier baselines and where they fundamentally diverge.

So, based on the results of the work performed, it is possible to formulate the following scientific novelty and practical significance of the research results.

Scientific novelty of the obtained research results – the Sample4Geo pipeline configuration compactness verification method has been further developed, which, unlike the existing one, provides competitive localization accuracy at significantly lower computational costs by jointly optimizing the size of the backbone, training hyperparameters, and input data resolution, and also formalizes the best compromise between localization quality and model size for resource-constrained UAV deployment.

Practical significance of the research results – the obtained configurations and co-optimization strategy can be used to design and deploy UAV visual geo-localization systems that apply compact Sample4Geo-based models with tuned training settings and increased input resolution, thereby achieving competitive localization accuracy with reduced model size and enabling implementation of cross-view matching in resource-constrained onboard environments.

Conclusions / Висновок

In this work, a set of lightweight Sample4Geo-based configurations for UAV visual geo-localization under resource-constrained conditions has been developed and comprehensively evaluated, which made it possible to determine the advantages, limitations, and practical applicability of the proposed approach. Based on the results of the research the following main conclusions can be drawn.

1. Systematized the evaluation of lightweight Sample4Geo configurations under a unified baseline protocol, showing that compact backbones such as ConvNeXt-Nano, ViT-Tiny and MobileNetV3 already provide competitive geo-localization accuracy relative to a ConvNeXt-Base reference at 224×224 resolution.
2. Analyzed the effect of systematic backbone downscaling on geo-localization accuracy and parameter count, demonstrating that architectures like ConvNeXt-Nano and ViT-Tiny retain strong Recall@1 at much lower capacity, whereas LeViT remains poorly suited to the Sample4Geo pipeline even after hyperparameter tuning.
3. Investigated the influence of hyperparameter selection (number of epochs, batch size, initial learning rate) for each backbone and confirmed that a simple, near-uniform training schedule is sufficient for most models, while MobileNetV4 and ViT-Small exhibit particularly large performance gains when appropriately tuned.
4. Demonstrated that increasing input resolution from 224×224 to 384×384 consistently improves performance for all evaluated models and disproportionately benefits higher-capacity backbones, enabling ConvNeXt-Nano to approach the accuracy of ConvNeXt-Base while remaining significantly smaller in size.
5. Established the performance-efficiency frontier for the evaluated feature extractors by jointly considering Recall@1 and the number of trainable parameters, identifying MobileNetV3 and ViT-Tiny as attractive options at very low capacity and highlighting ConvNeXt-Nano as a Pareto-optimal knee point with a favorable trade-off between accuracy and model size for onboard UAV deployment.

References

1. Arandjelović, R., Gronat, P., Torii, A., Pajdla, T., & Sivic, J. (2018). NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6), 1437–1451. https://doi.org/10.1109/TPAMI.2017.2711011
2. Campos, C., Elvira, R., Rodríguez, J. J. G., M. Montiel, J. M., & D. Tardós, J. (2021). ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual – Inertial, and Multimap SLAM. *IEEE Transactions on Robotics*, 37(6), 1874–1890. https://doi.org/10.1109/TRO.2021.3075644
3. Deuser, F., Habel, K., & Oswald, N. (2023). Sample4Geo: Hard Negative Sampling For Cross-View Geo-Localisation. *arXiv* [*Cs.CV*]. https://doi.org/10.48550/arXiv.2303.11851
4. Ding, L., Zhou, J., Meng, L., & Long, Z. (2021). A Practical Cross-View Image Matching Method between UAV and Satellite for UAV-Based Geo-Localization. *Remote Sensing*, 13(1), article ID 47. https://doi.org/10.3390/rs13010047
5. Ding, X., Zhang, X., Song, S., Li, B., Hui, L., & Dai, Y. (2025). Cross-View Geo-Localization via 3D Gaussian Splatting-Based Novel View Synthesis. *Remote Sensing*, 17(22), article ID 3673. https://doi.org/10.3390/rs17223673
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., … Houlsby, N. (2021). An Image is Worth 16x16 Words: *Transformers for Image Recognition at Scale. arXiv* [*Cs.CV*]. https://doi.org/10.48550/arXiv.2010.11929
7. Graham, B., El-Nouby, A., Touvron, H., Stock, P., Joulin, A., Jégou, H., & Douze, M. (2021, October). LeViT: A Vision Transformer in ConvNets Clothing for Faster Inference. *Proceedings of the IEEE/CVF International Conference on Computer Vision* (*ICCV*), 12259–12269. https://doi.org/10.1109/ICCV48922.2021.01204
8. Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., … Adam, H. (2019, October). Searching for MobileNetV3. *Proceedings of the IEEE/CVF International Conference on Computer Vision* (*ICCV*). https://doi.org/10.1109/ICCV.2019.00140
9. Hu, S., Feng, M., Nguyen, R. M. H., & Lee, G. H. (2018). CVM-Net: Cross-View Matching Network for Image-Based Ground-to-Aerial Geo-Localization. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7258–7267. https://doi.org/10.1109/CVPR.2018.00758
10. Liao, Y., Su, J., Ma, D., & Niu, C. (2025). UAV-Satellite Cross-View Image Matching Based on Adaptive Threshold-Guided

Ring Partitioning Framework. *Remote Sensing*, 17(14), article ID 2448. https://doi.org/10.3390/rs17142448
11. Loshchilov, I., & Hutter, F. (2019). Decoupled Weight Decay Regularization. *arXiv* [*Cs.LG*]. https://doi.org/10.48550/arXiv.1711.05101
12. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., … Bojanowski, P. (2024). DINOv2: Learning Robust Visual Features without Supervision. *arXiv* [Cs.CV]. https://doi.org/10.48550/arXiv.2304.07193
13. Qin, D., Leichner, C., Delakis, M., Fornoni, M., Luo, S., Yang, F., … Howard, A. (2024). MobileNetV4: Universal Models for the Mobile Ecosystem. *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XL*, 78–96. Presented at the Milan, Italy. https://doi.org/10.1007/978-3-031-73661-2_5
14. Shi, Y., Liu, L., Yu., X., & Li, H. (2019). Spatial-Aware Feature Aggregation for Image based Cross-View Geo-Localization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, & R. Garnett (Eds). *Advances in Neural Information Processing Systems* (Vol. 32). URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/ba2f0015122a5955f8b3a50240fb91b2-Paper.pdf
15. Todi, A., Narula, N., Sharma, M., & Gupta, U. (2023). ConvNext: A Contemporary Architecture for Convolutional Neural Networks for Image Classification. 2023 3rd *International Conference on Innovative Sustainable Computational Technologies* (CISCT), pp. 1–6. https://doi.org/10.1109/CISCT57197.2023.10351320
16. Wang, T., Zheng, Z., Yan, C., Zhang, J., Sun, Y., Zheng, B., & Yang, Y. (2022). Each Part Matters: Local Patterns Facilitate Cross-View Geo-Localization. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(2), 867–879. https://doi.org/10.1109/TCSVT.2021.3061265
17. Wang, T., Zheng, Z., Zhu, Z., Sun, Y., Yan, C., & Yang, Y. (2024). Learning Cross-View Geo-Localization Embeddings via Dynamic Weighted Decorrelation Regularization. *IEEE Transactions on Geoscience and Remote Sensing*, *62*, 1–12. https://doi.org/10.1109/TGRS.2024.3491757
18. Workman, S., Souvenir, R., & Jacobs, N. (2015). Wide-Area Image Geolocalization with Aerial Reference Imagery. *IEEE International Conference on Computer Vision* (ICCV), 1–9. https://doi.org/10.1109/ICCV.2015.451
19. Zheng, Z., Wei, Y., & Yang, Y. (2020). University-1652: A Multi-view Multi-source Benchmark for Drone-based Geo-localization. *Proceedings of the 28th ACM International Conference on Multimedia*, 1395–1403. Presented at the Seattle, WA, USA. https://doi.org/10.1145/3394171.3413896
20. Zhu, R., Yang, M., Yin, L., Wu, F., & Yang, Y. (2023). UAVs Status Is Worth Considering: A Fusion Representations Matching Method for Geo-Localization. *Sensors*, *23*(2), article ID 720. https://doi.org/10.3390/s23020720
21. Zhu, S., Shah, M., & Chen, C. (2022). TransGeo: Transformer Is All You Need for Cross-view Image Geo-localization. *arXiv* [*Cs.CV*]. https://doi.org/10.48550/arXiv.2204.00097
22. Zhu, S., Yang, T., & Chen, C. (2021). VIGOR: Cross-View Image Geo-localization beyond One-to-one Retrieval. *arXiv* [*Cs.CV*]. https://doi.org/10.48550/arXiv.2011.12172
23. Zhu, Y., Yang, H., Lu, Y., & Huang, Q. (2023). Simple, Effective and General: A New Backbone for Cross-view Image Geo-localization. *arXiv* [*Cs.CV*]. https://doi.org/10.48550/arXiv.2302.01572
24. Zhuang, J., Dai, M., Chen, X., & Zheng, E. (2021). A Faster and More Effective Cross-View Matching Method of UAV and Satellite Images for UAV Geolocalization. *Remote Sensing*, 13(19), article ID 3979. https://doi.org/10.3390/rs13193979

Р. І. Ілечко, Ю. В. Цимбал

Національний університет "Львівська політехніка", м. Львів, Україна

ЕФЕКТИВНЕ МІЖРАКУРСНЕ ЗІСТАВЛЕННЯ ЗОБРАЖЕНЬ, ОТРИМАНИХ З КАМЕР БПЛА

Розглянуто завдання візуальної геолокації безпілотних літальних апаратів за умови деградації GNSS та проаналізовано особливості застосування крос-в'ю підходів типу Sample4Geo на бортовому обладнанні за обмежених ресурсів. Досліджено поведінку компактних екстракторів ознак у фреймворку Sample4Geo та проаналізовано вплив систематичного зменшення їх ємності, конфігурації навчання та роздільної здатності вхідних зображень на точність крос-в'ю геолокації. Запроваджено уніфікований протокол оцінювання фреймворку на стандартному крос-в'ю наборі даних та виконано навчання низки згорткових і трансформерних архітектур (ConvNeXt-Base/Nano, ViT-Small/Tiny, MobileNetV3/V4, LeViT) за спільними базовими налаштуваннями за роздільної здатності 224×224 із подальшим цілеспрямованим пошуком гіперпараметрів. Досліджено особливості масштабування роздільної здатності зображення шляхом повторного навчання налаштованих моделей для 384×384 та змодельовано перелік ефективності за показниками Recall@1 і кількості навчовуваних параметрів. Показано, що компактні бекбони забезпечують конкурентну точність уже за простої одноепохової схеми навчання, а налаштування гіперпараметрів дає істотний додатковий приріст для MobileNetV4 і ViT-Small та підтверджує достатність майже однакового режиму навчання для ConvNeXt-Nano, ViT-Tiny і MobileNetV3. Встановлено, що підвищення роздільної здатності вхідних зображень стабільно покращує результати роботи всіх моделей. При цьому ConvNeXt-Nano наближається за точністю до ConvNeXt-Base, залишаючись значно компактнішою, а ViT-Small перевершує ViT-Tiny за наявності детальнішої просторової інформації. Побудовано множину Парето, на якому MobileNetV3 та ViT-Tiny визначено як привабливі варіанти за дуже низької ємності, а ConvNeXt-Nano окреслено як "точку згину" з вигідним компромісом між точністю та розміром моделі для бортового застосування. Сформульовано практичні рекомендації щодо вибору екстракторів ознак і параметрів навчання для ресурсно-обмеженої крос-в'ю геолокації та окреслено напрями етапів подальшого дослідження, пов'язаних з аналізом затримок та енергоспоживання на реальних платформах БПЛА.

Ключові слова: глибока нейронна мережа; трансформери; згорткова нейронна мережа; комп'ютерний зір; візуальна геолокалізація.