ISSN 2707-1898 (print)
Український журнал інформаційних технологій
Ukrainian Journal of Information Technology
http://science.lpnu.ua/uk/ujit
https://doi.org/10.23939/ujit2025.01.068
Article received 22.04.2025 p.
Article accepted 01.05.2025 p.
UDC 004.8, 004.93, 004.932

Correspondence author
R. I. Ilechko
roman.i.ilechko@lpnu.ua

R. I. Ilechko, O. O. Borovyi, Y. V. Tsymbal
Lviv Polytechnic National University, Lviv, Ukraine

ENSEMBLE IMAGE SUPER-RESOLUTION FOR UAV GEO-LOCALIZATION

In this paper, we address the challenge of visual geo-localization from low-quality UAV imagery captured in real world environments. We propose a two-stage architecture, which includes Super-Resolution and visual geo-localization. We introduced novel, non-learnable Ensemble Super-Resolution (ESR) module, which first refines upscaled aerial frames, then seamlessly feeds the enhanced imagery into a visual geo-localization pipeline. Designed as a parallelizable block integrated directly into any SR computation graph, ESR combines classical Bicubic interpolation with neural SR models – boosting image fidelity and overall system accuracy without additional training and executing efficiently on most hardware accelerator. We validate our approach on a dataset of 37 000 real-world UAV images, each downscaled by a factor of four and then restored via baseline methods (Bicubic, Bilinear, Nearest Neighbour, DRCT, HMA, HAT, SwinFIR) as well as our ESR-enhanced pipeline. Quantitative evaluation shows that standalone Super-Resolution methods yield PSNR in the low 20s dB and SSIM of 0.6–0.7 – far below standard benchmarks-leading to a marked drop in geo-localization accuracy (Recall@1 and AP).

In contrast, our ESR module stabilizes SR outputs and recovers image fidelity, raising geo-localization Recall@1 to 87.0 % (vs. 84.96 % with HMA restoration) and AP to 89.1 % (against 87.41 % with HMA restoration).

Our contributions are:

Two-stage framework combining Image Super-Resolution and visual geo-localization approaches tailored for low-resolution, noisy UAV data.

Non-learnable, parallelizable ESR block that fuses Bicubic interpolation with neural restoration within the network Super-Resolution graph – requiring no retraining and fully compatible with most accelerator.

Comprehensive empirical study demonstrating that ESR substantially narrows the domain gap and boosts geo-localization performance in real-world conditions.

We conclude that embedding lightweight, hardware-agnostic ensemble strategies into SR pipelines is a promising direction for robust UAV-based visual localization. Future work will explore adaptive ensemble weighting and domain-aware SR architectures to further mitigate aerial imaging noise and variability.

Keywords: super-resolution, deep neural network, transformers, convolution neural network, computer vision.

Introduction

Unmanned Aerial Vehicles (UAVs) have rapidly evolved as indispensable tools across both military and civilian domains, yet their vision-based navigation and localization capabilities still struggle when confronted with noisy, low-resolution imagery captured in unconstrained "wild" environments. While recent advances in computer vision ranging from Convolutional Neural Networks (CNNs) and Transformer-based feature extractors to state-of-the-art geo-localization pipelines – have improved robustness to GPS denial and limited video quality, the direct integration of image Super-Resolution (SR) techniques into visual geo-localization (VG) remains underexplored.

Most existing restoration approaches, whether classical (Bicubic, Bilinear, Nearest Neighbour) or neural (DRCT, HMA, HAT, SwinFIR), were developed and benchmarked on natural or urban imagery; when applied "out of the box" to UAV data, they often introduce artifacts or wash out critical spatial details, leading to degraded downstream geo-localization performance. This mismatch highlights a pressing need for an adaptive SR strategy that can be deployed on UAV accelerators without costly retraining, stabilize outputs, and preserve the fine-grained features that geo-localization models rely upon. In this work, we bridge that gap by proposing a two-stage SR + VG framework anchored by a novel, non-learnable ESR module. ESR is designed as a fully parallelizable block – integrated directly into any SR computation graph and executable on common hardware accelerators – that fuses classical Bicubic interpolation with neural SR outputs. By combining the strengths of both analytical and learned methods, ESR suppresses artifacts, stabilizes the upscaling process, and preserves descriptive features, all without additional model retraining.

Relevance of research. As UAV deployments in complex, GPS-denied, or visually degraded environments become more widespread, resilient vision-based localization

methods are essential. Our two-stage SR + VG architecture directly addresses the impact of low-quality imagery on geo-localization accuracy, providing a lightweight, hardware-agnostic enhancement block that can be seamlessly added to existing pipelines.

The object of this research: the visual sensing subsystem of UAVs operating in real-world scenarios, where image resolution and noise often fall below the thresholds required for reliable geo-localization.

The subject of the research: computer vision methodologies for UAV visual geo-localization and image enhancement, with a focus on SR and its integration into localization pipelines.

The purpose of this work is to quantify how low-resolution, noisy UAV imagery degrades geo-localization accuracy and exposes artifacts from standalone SR methods, develop and evaluate a two-stage framework – combining SR with visual geo-localization – centered on a novel, non-learnable ESR block.

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

1. Quantify how low resolution, noisy UAV data degrades geo-localization accuracy, identify the artifacts introduced by standalone SR methods.
2. Develop and evaluate a two-stage SR + VG framework tailored for UAV imagery under adverse conditions.
3. Design and perform a comprehensive analysis of the introduced non-learnable, parallelizable ESR block.
4. Demonstrate that ESR module raises downstream geo-localization metrics (Recall@1, AP) significantly, with minimal computational overhead.
5. Proposing recommendations for future research and practical applications aimed at enhancing UAV operations in challenging real-world environments.

Materials and methods of research. This study introduces a two-stage framework that combines image enhancement with visual geo-localization, utilizing large-scale UAV imagery preprocessing to simulate wild conditions. Research features a novel, non-learnable ESR block – integrated directly into the SR computation graph – that fuses classical and neural Super-Resolution methods, and leverages hardware-agnostic, parallelizable components for efficient execution on common accelerators.

Dataset and preprocessing.

The experiments were conducted using a publicly available UAV imagery dataset obtained from the University-1652 [1] UAV image collection. To simulate image degradation, all original images were downscaled by a factor of four using the INTER_AREA interpolation method, which is known for minimizing aliasing effects during reduction. A total of 37,000 images were processed for the subsequent experiments.

Upscaling methods.

Three classical interpolation methods were selected for evaluation: Bilinear, Bicubic [2], and Nearest Neighbor. These methods were implemented using the OpenCV library. In addition, four neural network-based upscaling models were included in the comparison: DRCT [3], HAT [4], HMA [5], and SwinFIR [6]. The official implementations of each neural network model were utilized to ensure reproducibility. For each method, a separate Python environment was created where all officially recommended dependencies were installed, thus avoiding version conflicts and ensuring consistency with the original implementations. To improve the visual quality and robustness of image enhancement results obtained from multiple algorithms, we propose a Ensemble Super-Resolution technique. This method constructs a single composite image by probabilistically selecting and assembling patches from a set of enhanced images generated by different algorithms. The process captures the strengths of individual algorithms while maintaining spatial coherence.

Let: $I = \{I_1, I_2, \ldots, I_N\}$ – a set of N enhanced images, each $I_j \in R^{H \times W \times C}$, where denote height, width, and number of channels, respectively. The side length of square patches used to divide each image denote as s, then the total nu mber of non-overlapping patches per image define as:

$$M = \frac{H}{S} \times \frac{W}{S},$$

$P = \{p_1, p_2, \ldots, p_N\}$ – a discrete probability distribution over the N algorithms, where:

$$p \in [0, 1], \sum_{i=1}^{N} p_i = 1.$$

For each patch position $j = \{1, 2, \ldots, M\}$ an algorithm index k_j is sampled according to the categorical distribution defined by P:

$$k_j \approx Categorical(P).$$

The resulting patch at location j in the final ensemble image I^* is then selected as:

$$I_j^* = I_{k,j},$$

where $I_{k,j}$ denotes the j-th patch of the k_j-th image. The ensemble image I^* is reconstructed by placing each selected patch into its corresponding position, forming a composite that leverages the complementary strengths of the contributing algorithms.

Experimental Procedure.

Each upscaling method, both classical and neural network-based, processed the same set of pre-downscaled images independently. After restoration, the upscaled images were saved to disk for further analysis. The performance of each method was evaluated using two standard image quality metrics: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM). Mean PSNR and SSIM values were calculated for each method across the entire dataset. These metrics were compared not only within the dataset but also against typical benchmark values reported in the literature for vision datasets.

Application to visual geo-localization.

To assess the practical impact of the upscaling methods, an additional evaluation was conducted by applying the

restored images to the task of visual geo-localization, one of the most prominent tasks in UAV-based computer vision. The Sample4Geo [7] method was employed as the visual geo-localization framework. The performance of the model using the restored images was compared to that obtained with the original high-resolution images, thus providing insights into how different upscaling methods affect downstream vision tasks.

Hardware and software environment.

All experiments were conducted on a workstation equipped with an NVIDIA RTX 3090 Ti GPU. Software environments were configured using Python 3.10, with specific library versions matching the official recommendations for each neural network model.

Analysis of recent research and publications.

1. Visual Geo-localization.

Cross-view feature representation.

Cross-view geo-localization faces significant challenges due to viewpoint variations and environmental dynamics. Early methods focused on joint representation learning and contextual feature extraction to bridge the gap between aerial and ground perspectives. At [8] researchers pioneered this direction with RK-Net, which integrates representation learning and keypoint detection using a Unit Subtraction Attention Module (USAM). USAM employs a Unit Subtraction Convolution to highlight local feature differrences, enabling robust keypoint matching across viewpoints without additional annotations. Evaluated on datasets like University-1652 [1], RK-Net achieved competitive accuracy by leveraging subtle appearance variances caused by viewpoint shifts. Building on this, the Local Pattern Network [9] (LPN) introduced a square-ring partition strategy to exploit spatial-contextual relationships in images. Inspired by human visual processing, LPN divides feature maps into concentric regions, assuming geographic targets cluster at the center. Using ResNet-50 [10] backbones with weight-sharing between aerial branches, LPN maps multi-view features (satellite, drone, ground) into a shared semantic space. This approach improved alignment by addressing feature distribution mismatches, demonstrating the value of hierarchical spatial reasoning for cross-view tasks.

Transformer-based feature segmentation.

Recent advances in transformer architectures have enabled more robust feature alignment for UAV-view geo-localization. The Feature Segmentation and Region Alignment [11] (FSRA) method addresses position shifts and scale uncertainty by combining transformer-based feature extraction with part-based segmentation. FSRA divides input images into patches, encodes positional embeddings, and uses a Heatmap Segmentation Module (HSM) to categorize regions (e. g., buildings, roads). A Heatmap Alignment Branch (HAB) then pools and aligns region-specific features, while a multiple sampling strategy mitigates training instability through augmented satellite images. By focusing on semantic segmentation rather than fine-grained details, FSRA improves robustness against viewpoint-induced distortions.

Contrastive learning with hard negative sampling.

Sampling strategies play a critical role in contrastive learning for geo-localization. Sample4Geo [7] introduces a novel framework using hard negative sampling to enhance model discriminability. It employs a symmetric InfoNCE [12] loss with two strategies: (1) geographic proximity sampling, selecting neighboring locations as hard negatives, and (2) visual similarity sampling, targeting geographically distinct but visually similar image pairs. Sample4Geo outperformed state-of-the-art methods by simplifying the training pipeline and eliminating complex pre-processing. Its success highlights the importance of challenging samples in improving generalization across diverse environments.

Table 1. Comparison of visual geo-localization methods on the University-1652 dataset. The best results are highlighted in bold

Method	AP, %	Recall@1, %
Sample4Geo	**93.81**	**92.65**
FSRA	84.82	82.25
LPN+USAM	80.55	77.6
LPN	79.14	75.93
RK-Net	70.23	66.13

2. Single Image Super-Resolution (SISR).

Classic and early neural-network-based methods.

Early SISR methods relied on interpolation techniques such as bicubic and bilinear resampling, which estimate pixel values using weighted averages of neighboring pixels. While computationally efficient, these methods often produce overly smooth outputs with limited high-frequency detail recovery. Example-based approaches like nearest neighbors (NN) and sparse coding [13] later emerged, leveraging neighbor embedding or learned dictionaries to map low-resolution (LR) to high-resolution (HR) patches. NN-based methods, for instance, searched for structurally similar LR patches in a training dataset and reconstructed HR patches by aggregating contributions from their k-closest matches. These methods improved edge preservation compared to interpolation but struggled with generalization due to their reliance on handcrafted similarity metrics (e. g., Euclidean distance for patch matching) and small patch-wise processing, which limited their ability to model global image structures.

The advent of deep learning revolutionized SISR by enabling end-to-end mapping of LR to HR images. SRCNN [14] pioneered this shift with a three-layer CNN, outperforming traditional methods by learning hierarchical features directly from data. Subsequent works introduced architectural refinements: FSRCNN [15] accelerated inference by adopting a compact design with transposed convolutions for upsampling, while VDSR [16] deepened networks and incorporated residual learning to stabilize

training. DRCN [17] further enhanced performance with recursive layers, and EDSR [18] removed batch normalization to enable larger model capacities.

Attention mechanisms and advanced residual structures later dominated the field. RDN [19] aggregated multi-level features via dense residual blocks, and RCAN [20] introduced channel attention to prioritize informative features. These CNN-based methods established foundational principles-residual learning, multi-scale fusion, and attention-that remain central to modern SISR architectures.

SwinIR: Transformer-Based Image Restoration.

SwinIR [21] represents a pivotal advancement in SISR by integrating shifted window attention from the Swin Transformer architecture into low-level vision tasks. Unlike CNN-based methods, SwinIR leverages hierarchical window partitioning to balance global context modeling and computational efficiency. Its architecture comprises three modules:

1. Shallow Feature Extraction: A 3×3 convolutional layer captures low-level spatial details.
2. Deep feature extraction: Residual Swin Transformer Blocks (RSTB) combine Swin Transformer layers (with shifted windows for cross-region interactions) and residual connections, enabling multi-scale feature aggregation.
3. Reconstruction module: Task-specific layers (e. g., sub-pixel convolution) fuse shallow and deep features to generate high-resolution outputs.

SwinIR outperformed CNN-based models (e. g., EDSR, RCAN) by 0.14–0.45 dB PSNR on benchmarks like Urban100 and Manga109 while reducing parameters by up to 67 %. Its success inspired extensions such as SwinFSR [22] for stereo SR and omnidirectional SR models addressing equirectangular projection distortions. By bridging transformer-based global modeling with CNN-inspired locality, SwinIR established a robust baseline for subsequent SISR innovations.

Hybrid attention mechanisms.

Recent advancements in SISR have leveraged hybrid attention mechanisms to enhance spatial feature utilization in Transformer architectures. The Hybrid Attention Transformer (HAT) [4] addresses limited pixel engagement in window-based Transformers by integrating channel attention (capturing global statistics) with window self-attention (modeling local details). An overlapping cross-attention module (OCAB) further broadens receptive fields by facilitating cross-window interactions. Combined with same-task pre-training on ImageNet, HAT achieves state-of-the-art performance, improving PSNR by up to 1.2 dB over SwinIR. Attribution analysis via Layer Attribution Maps (LAM) confirms its expanded spatial utilization. Building on this, the Hybrid Multi-Axis Aggregation Network (HMANet) [5] introduces Residual Hybrid Transformer Blocks (RHTB) to fuse Swin Transformer layers with CNNs via Fused Attention Blocks (FAB), balancing local-global feature extraction. To exploit structural self-similarity, Grid Attention Blocks (GAB) partition features into intervals, enabling sparse cross-region attention. HMANet's task-specific pre-training strategy, initializing models with parameters from different scales, yields consistent gains (0.05–0.09 dB PSNR), achieving top results on Urban100 and Manga109. However, its computational overhead (69.9M parameters) highlights trade-offs between performance and efficiency.

Table 2. Quantitative comparison of the several SISR methods on benchmark datasets for x4 upscaling. The best results are highlighted in bold [2–6, 16, 20, 21]

Method	Set5		Set14		BSD100		Urban100		Manga109	
	PSNR dB	SSIM	PSNR dB	SSIM	PSNR dB	SSIM	PSNR dB	SSIM	PSNR dB	SSIM
Bicubic	28.42	0.8104	26	0.7027	25.69	0.6675	23.14	0.6577	24.89	0.7866
SRCNN	30.48	0.8628	27.5	0.7513	26.9	0.7101	24.53	0.7221	27.58	0.8555
FSRCNN	30.72	0.8666	27.61	0.7555	26.98	0.715	24.62	0.728	27.9	0.861
DRCN	31.53	0.8841	28.04	0.7704	27.24	0.7243	25.14	0.7518	28.99	0.8891
EDSR	32.46	0.8968	28.8	0.7876	27.71	0.742	26.64	0.8033	31.02	0.9148
RDN	32.47	0.899	28.81	0.7871	27.72	0.7419	26.61	0.8028	31	0.9151
RCAN	32.63	0.9002	28.87	0.7889	27.77	0.7436	26.82	0.8087	31.22	0.9173
VDSR	31.35	0.882	28.02	0.7681	27.29	0.0711	25.18	0.751	28.83	0.887
SwinIR	32.92	0.9044	29.09	0.795	27.92	0.7489	27.45	0.8254	32.03	0.926
SwinFIR	33.08	0.9048	29.21	0.7971	27.98	0.7508	27.87	**0.9348**	32.52	0.9292
HAT	33.04	0.9056	29.23	0.7973	28	0.7517	27.97	0.8368	32.48	0.9292
HMA	**33.15**	0.906	29.32	**0.7996**	28.05	0.753	**28.42**	0.845	**32.97**	**0.932**
DRCT	33.11	**0.9064**	**29.35**	0.7984	**28.18**	**0.7532**	28.06	0.8378	32.59	0.9304

Frequency-spatial feature fusion.

Integrating frequency-domain processing has emerged as a powerful strategy for global dependency modeling. SwinFIR [6] augments SwinIR with Spatial Frequency Blocks (SFB), combining Fast Fourier Convolution (FFC) for global frequency features and residual CNNs for local spatial details. This hybrid design mitigates SwinIR's reliance on local attention, while Charbonnier loss stabilizes training and pixel-domain augmentations (Mixup, channel shuffle) enhance generalization. Zero-cost feature ensemble strategy merges parameters from multiple checkpoints, improving performance without inference overhead. SwinFIR outperforms SwinIR by up to 0.8 dB PSNR on Manga109, demonstrating the efficacy of frequency-spatial fusion.

Efficient architectures and information flow.

Addressing information bottlenecks in deep SR models, DRCT [3] introduces dense residual connections within Swin Transformer blocks (SDRCBs) to stabilize feature propagation and preserve spatial details. By mitigating abrupt feature suppression in deeper layers, DRCT [3] achieves superior reconstruction with fewer parameters than HAT or SwinIR. A Same-task Progressive Training Strategy (SPTS) combines $L1/L2$ losses to enhance high-frequency recovery, validated by competitive NTIRE 2024 Challenge results (31.44 dB validation PSNR). Despite its efficiency, DRCT [3] lacks theoretical analysis of its dense connections' role in mitigating information loss, warranting future exploration.

Research results and their discussion

In this study, we evaluated a range of image upscaling methods applied to UAV aerial imagery under real-world, noisy conditions. Two sets of experiments were conducted. The first assessed the performance of geo-localization models on restored images compared to the original high-quality data (Baseline). Results are displayed in Table 3. The second focused on quantifying the restoration quality using standard image quality metrics, results are provided in Table 4.

The proposed two-stage framework for visual geo-localization on low quality images displayed in the Fig. 1.

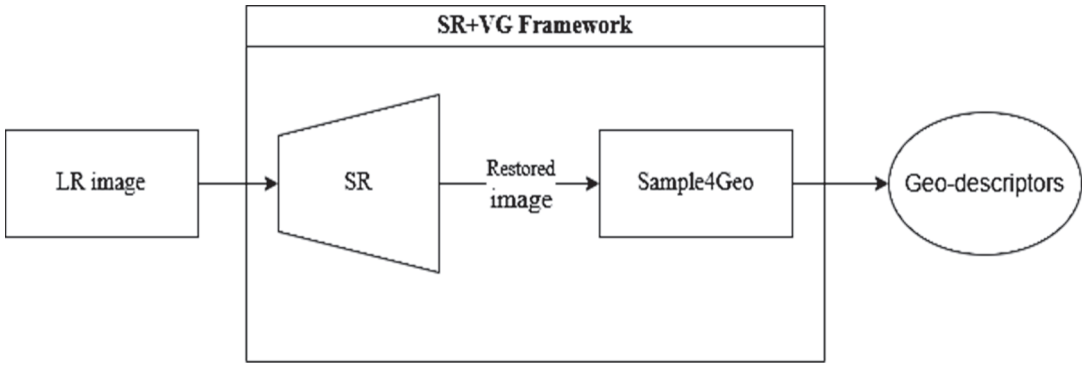

Fig. 1. Diagram of the proposed two stage framework

Table 3. Quantitative comparison of Sample4Geo performance on original aerial imagery (Baseline) and after image restoration on the University-1652 dataset

Method	Recall@1, %	AP, %
Baseline	92.65	93.81
Bicubic	88.21	90.11
Bilinear	88.22	90.06
NN	69.66	73.72
DRCT	84.43	86.96
HMA	84.96	87.41
HAT	83.36	87.01
SwinFIR	84.2	86.63
DRCT+ESR	87.01	89.09

A low-resolution (LR) UAV image is first passed through a Super-Resolution (SR) module, which reconstructs a higher-quality version of the input. The restored image is then fed into the visual geo-localization component (Sample4Geo), where spatial features are extracted and encoded as geo-descriptors. These descriptors serve as the final output for downstream localization.

Table 4. Performance of Restoration Methods on the University-1652 dataset. The best result is highlighted in bold

Method	PSNR dB	SSIM
Bicubic	22.61	0.64
Bilinear	22.08	0.6
NN	21.68	0.6
DRCT	**23.1**	**0.69**
HMA	22.88	0.69
HAT	22.6	0.62
SwinFIR	22.58	0.61
DRCT+ESR	22.77	0.63

For the geo-localization task, the Baseline model (applied to the original images) achieved a Recall@1 of 92.65 % and an AP of 93.81 %. Classical interpolation methods exhibited

varying degrees of performance degradation when processing the downscaled and subsequently upscaled images. Bicubic and Bilinear interpolation methods produced Recall@1 values of 88.21 % and 88.22, and AP values of 90.11 and 90.06, respectively. Notably, the Nearest Neighbour approach suffered a substantial drop, yielding a Recall@1 of 69.66 and an AP of 73.72. Among the neural network-based approaches, DRCT, HMA, HAT, and SwinFIR demonstrated intermediate performance, with Recall@1 values ranging from 83.36 to 84.96 and AP values from 86.63 to 87.41. Although these methods provided slightly higher restoration metrics than the Nearest Neighbour method, the overall geo-localization performance still lagged behind the Baseline.

The quality of the restored images was further quantified using the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM). The classical methods returned PSNR values of 22.61 dB (Bicubic), 22.08 dB (Bilinear), and 21.68 dB (Nearest Neighbour), with corresponding SSIM values around 0.64, 0.60, and 0.60, respectively. The neural network-based approaches yielded modest improvements, with DRCT [3] achieving a PSNR of 23.1 dB and SSIM of 0.69, HMA with 22.88 dB PSNR and 0.69 SSIM, HAT with 22.6 dB PSNR and 0.62 SSIM, and SwinFIR with 22.58 dB PSNR and 0.61 SSIM. To further reduce artifacts from SR methods and improve both analytical fidelity and perceptual quality, we propose a patch-based ensemble. By probabilistically combining patches from neural restoration outputs with those from classical interpolation and analytical techniques, our approach harnesses their complementary strengths – dampening hallucination artifacts while retaining fine structural detail. In particular, blending DRCT [3] with bicubic interpolation trough ESR module improves DRCT's accuracy and boosts overall performance, yielding a Recall@1 of 87.01 % while maintaining balanced perceptual quality.

The process of improved image restoration illustrated in Fig. 2. A low-resolution (LR) UAV image is simultaneously fed into a standard Super-Resolution network and a classical Bicubic interpolation block. The SR network produces an initial upscaled output, while the Bicubic block generates a parallel interpolation result. Both outputs are then combined within the non-learnable, parallelizable ESR module, which fuses the neural and analytical upscaling streams to suppress artifacts and preserve fine spatial details. The module's output is the final restored image, optimized for subsequent visual geo-localization tasks. This ensemble aims to improve the stability and reliability of neural network-based restoration by leveraging the consistency of classical methods like Bicubic interpolation. The fused output is then passed to the Sample4Geo module, which extracts geo-descriptors necessary for the downstream geo-localization task. This design effectively balances perceptual quality with analytical performance, helping mitigate the artifacts introduced by individual models.

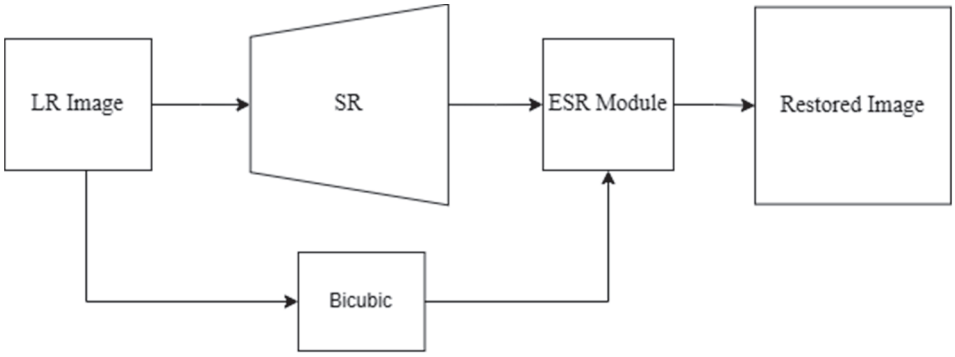

Fig. 2. Diagram of the image restoration with proposed ESR module

Fig. 3. Example of image restoration using the Ensemble (Our), Bicubic, DRCT, and HMA algorithms (from left to right)

Fig. 3. Example of image restoration using the Ensemble (Our), Bicubic, DRCT, and HMA algorithms (from left to right)

Discussion of research results. A deeper analysis of the experimental results reveals several critical insights. In our findings selected restoration methods yield similar PSNR and SSIM values, however, on standard vision datasets, neural methods typically achieve PSNR values in the range of 30–40 dB and SSIM values exceeding 0.85. The lower values observed in this study PSNR in the low 20s and SSIM around 0.6–0.7 are indicative of a pronounced domain shift. This shift arises because the neural network models, which are usually optimized on controlled or ideal datasets, encounter significant challenges when applied to low-quality, noisy UAV imagery acquired under wild conditions.

The substantial drop in Recall@1 and AP between the Baseline and the restored images highlights the sensitivity of geo-localization algorithms to image quality. In particular, the drastic performance decrease observed with the Nearest Neighbour method (Recall@1 dropping to 69.66 and AP to 73.72) underscores the limitations of basic interpolation techniques when dealing with severely degraded input data.

Moreover, although the neural network-based restoration methods (DRCT [3], HMA [5], HAT [4], and SwinFIR [6]) performed at levels comparable to classical approaches, they still fell short of the high restoration quality typically expected from such advanced models. This gap directly affects the downstream geo-localization task, where even small deviations in image fidelity can lead to significant errors in target identification and navigation accuracy. Despite our ensemble's ability to boost the accuracy of neural-based restorations and sustain overall performance – raising Recall@1 to 87.01 % – a notable gap remains relative to the baseline (original high-resolution data), underscoring the challenge of fully closing the performance difference with restored imagery.

Furthermore, PSNR and SSIM emphasize pixel-level precision and structural similarity but overlook the semantic details essential for geolocation. An image that appears statistically "sharper" may still lose the subtle cues that Sample4Geo depends on. For example, in Image 3, network-based restoration methods introduce high-frequency artifacts that compromise features critical for geolocation, whereas Bicubic interpolation, despite yielding a lower PSNR, better preserves the edges necessary for Sample4Geo compared to DRCT [3] or HMA [5].

The Scientific novelty of the obtained research results. First of all, we propose a two-stage architecture for geo-localization of low-quality UAV imagery. Second, we introduce a novel, non-learnable Ensemble Super-Resolution block that seamlessly fuses classical interpolation with neural SR models within a single computation graph – stabilizing outputs and suppressing artifacts without any additional training. Additionally, our ESR module is hardware-agnostic and parallelizable, requiring no specialized retraining or custom accelerator support, and can be integrated plug-and-play into existing SR pipelines. Moreover, we validate our approach on 37 000 real "wild" UAV frames, showing that ESR raises Recall@1 from 84.96 % to 87.00 % and Average Precision from 87.41 % to 89.10 %.

The Practical significance of the research results. The proposed ESR block seamlessly integrates with Super-Resolution approach without requiring additional training, improving image quality and reducing artifacts. It's hardware-agnostic, parallelizable design ensures that it can run on resource-constrained UAV platforms and edge accelerators. By producing higher-fidelity imagery, ESR directly enhances geo-localization accuracy, leading to more precise visual geo-localization. Also, in this study we highlighted the importance of addressing domain shifts when developing restoration algorithms tailored for UAV imagery.

Conclusions

This study presented a two-stage framework, which combine Super-Resolution and visual geo-localization approaches, engineered for low-resolution, noisy UAV imagery. Central to our design is the Ensemble Super-Resolution (ESR) module – a non-learnable, parallelizable block that fuses classical Bicubic interpolation with neural SR outputs within any Super-Resolution graph, requiring no retraining and running efficiently on standard accelerators.

Our evaluation on 37000 real-world UAV frames showed that conventional restoration methods achieve only modest fidelity gains (PSNR in the low 20s, SSIM of 0.6–0.7) and limited performance in geo-localization. To address these shortcomings, we developed a ESR module that probabilistically combines patches from neural restoration

outputs with those from classical interpolation and analytical techniques. By tuning the sampling distribution to favor higher-accuracy predictors in each region. Our geo-localization framework with ESR module achieved a Recall@1 of 87.01 % – surpassing the standalone DRCT [3] method (84.43 %) – and delivered balanced perceptual quality. These findings demonstrate that lightweight ensemble strategies are a powerful tool for closing the domain gap in UAV imagery.

Future work should focus on restoration models and ensemble strategies tailored specifically to the noisy, real-world characteristics of UAV imagery. Mitigating domain shift through advanced training protocols, adaptive patch selection, and fine-tuning of neural architectures will be critical to further improving both image restoration and geo-localization accuracy.

References

[1] Zheng, Z., Wei, Y., & Yang, Y. (2020). University-1652: A Multi-view Multi-source Benchmark for Drone-based Geo-localization. In Proceedings of the 28th ACM International Conference on Multimedia (pp. 1395–1403). Association for Computing Machinery. https://doi.org/10.1145/3394171. 3413896

[2] Li, K., Yang, S., Dong, R., Wang, X., & Huang, J. (2020). Survey of single image super-resolution reconstruction. *IET Image Processing*, 14(11), 2273–2290. https://doi.org/ 10.1049/iet-ipr.2019.1438

[3] Hsu, C.C., Lee, C.M., & Chou, Y.S. (2024). DRCT: Saving Image Super-Resolution Away from Information Bottleneck. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (pp. 6133–6142). https://doi.org/10.1109/CVPRW63382.2024.00618

[4] Chen, X., Wang, X., Zhou, J., Qiao, Y., & Dong, C. (2023). Activating More Pixels in Image Super-Resolution Transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 22367–22377). https://doi.org/10.1109/CVPR52729. 2023.02142

[5] Chu, S. C., Dou, Z. C., Pan, J. S., Weng, S., & Li, J. (2024). HMANet: Hybrid Multi-Axis Aggregation Network for Image Super–Resolution. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (pp. 6257–6266). https://doi.org/10.1109/ CVPRW63382. 2024.00629

[6] Dafeng Zhang, Feiyu Huang, Shizhuo Liu, Xiaobing Wang, & Zhezhu Jin (2023). SwinFIR: Revisiting the SwinIR with Fast Fourier Convolution and Improved Training for Image Super-Resolution. https://doi.org/10.48550/ arXiv.2208.11247

[7] Deuser, F., Habel, K., & Oswald, N. (2023). Sample4Geo: Hard Negative Sampling For Cross-View Geo-Localisation. In 2023 IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 16801–16810). https://doi.org/10.1109/ ICCV51070.2023.01545

[8] Lin, J., Zheng, Z., Zhong, Z., Luo, Z., Li, S., Yang, Y., & Sebe, N. (2022). Joint Representation Learning and Keypoint Detection for Cross-View Geo-Localization. *IEEE Transactions on Image Processing*, 31, 3780–3792. https://doi.org/10.1109/TIP.2022.3175601

[9] 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

[10] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90

[11] Dai, M., Hu, J., Zhuang, J., & Zheng, E. (2022). A Transformer–Based Feature Segmentation and Region Alignment Method for UAV-View Geo-Localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7), 4376–4389. https://doi.org/10.1109/TCSVT.2021.3135013

[12] Aäron van den Oord, Yazhe Li, & Oriol Vinyals (2018). Representation Learning with Contrastive Predictive Coding. *ArXiv*, abs/1807.03748. http://dx.doi.org/10.48550/arXiv. 1807.03748

[13] Yang, J., Wright, J., Huang, T., & Ma, Y. (2010). Image Super–Resolution Via Sparse Representation. IEEE Transactions on Image Processing, 19(11), 2861–2873. https://doi.org/ 10.1109/TIP.2010.2050625

[14] Dong, C., Loy, C., He, K., & Tang, X. (2016). Image Super-Resolution Using Deep Convolutional Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2), 295–307. https://doi.org/10.1109/TPAMI. 2015.2439281

[15] Dong, X. (2016). Accelerating the Super-Resolution Convolutional Neural Network. In Computer Vision – ECCV 2016 (pp. 391–407). Springer International Publishing. https://doi.org/10.1007/978-3-319-46475-6_25

[16] Kim, J., Lee, J., & Lee, K. (2016). Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 1646–1654). https://doi.org/ 10.1109/CVPR. 2016.182

[17] Kim, J., Lee, J., & Lee, K. (2016). Deeply-Recursive Convolutional Network for Image Super-Resolution. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 1637–1645). https://doi.org/ 10.1109/CVPR. 2016.181

[18] Lim, B., Son, S., Kim, H., Nah, S., & Lee, K. (2017). Enhanced Deep Residual Networks for Single Image Super-Resolution. In 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (pp. 1132–1140). https://doi.org/10.1109/CVPRW.2017.151

[19] Zhang, Y., Tian, Y., Kong, Y., Zhong, B., & Fu, Y. (2018). Residual Dense Network for Image Super-Resolution. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2472–2481). https://doi.org/10.1109/ CVPR. 2018.00262

[20] Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., & Fu, Y. (2018). Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In ECCV. https://doi.org/ 10.1007/978-3-030-01234-2_18

[21] Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., & Timofte, R. (2021). SwinIR: Image Restoration Using Swin Transformer. In 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW) (pp. 1833–1844). https://doi.org/10.1109/ICCVW54120.2021.00210

[22] Chen, K., Li, L., Liu, H., Li, Y., Tang, C., & Chen, J. (2023). SwinFSR: Stereo Image Super-Resolution using SwinIR and Frequency Domain Knowledge. In 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (pp. 1764–1774). https://doi.org/ 10.1109/ CVPRW59228.2023.00177

Р. І. Ілечко, О. О. Боровий, Ю. В. Цимбал
Національний університет "Львівська політехніка", Львів, Україна

АНСАМБЛЕВЕ ПІДВИЩЕННЯ РОЗДІЛЬНОЇ ЗДАТНОСТІ ЗОБРАЖЕНЬ ДЛЯ ГЕОЛОКАЦІЇ БПЛА

Розглянуто проблему візуальної геолокалізації з використанням низькоякісних зображень з БПЛА, отриманих у реальних умовах. Ми пропонуємо двоетапну архітектуру, яка передбачає відновлення роздільної здатності зображення та візуальну геолокалізацію. Наведено новий модуль Ensemble Super-Resolution (ESR), який спочатку покращує збільшені аерофотокадри, а потім подає відновлені зображення в конвеєр візуальної геолокалізації. Розроблений як розпаралелюваний блок, інтегрований безпосередньо в будь-який обчислювальний граф відновлення зображення, ESR поєднує класичну бікубічну інтерполяцію із нейронними моделями відновлення зображення, поліпшуючи якість зображення і підвищуючи загальну точність системи без додаткового навчання, та ефективно виконується на більшості апаратних прискорювачів. Ми перевірили наш підхід на наборі даних з 37 000 реальних зображень з БПЛА, кожне з яких було зменшено в чотири рази, а потім відновлено за допомогою базових методів (бікубічної, білінійної інтерполяції, найближчого сусіда, DRCT, HMA, HAT, SwinFIR), а також нашого пайплайну з ESR. Кількісна оцінка показує, що окремі методи надвисокої роздільної здатності забезпечують PSNR в межах 20 дБ і SSIM 0.6–0.7 – набагато нижче від стандартних показників, що призводить до помітного зменшення точності геолокалізації (Recall@1 і AP).

На противагу цьому, наш модуль ESR стабілізує результати відновлення та реконструює якість зображення, підвищуючи точність геолокації Recall@1 до 87,0 % (порівняно з 84,96 % у разі відновлення HMA) і AP до 89,1 % (порівняно з 87,41 % у випадку відновлення HMA).

Наш внесок охоплює:

1. Двоетапний фреймворк, що поєднує підходи підвищення роздільної здатності зображення та візуальної геолокації, адаптований для роботи з низькороздільними зашумленими даними з БПЛА.

2. Розпаралелюваний блок ESR, що не потребує навчання, який поєднує бікубічну інтерполяцію з нейронним відновленням у межах мережевого графа і повністю сумісний із більшістю прискорювачів.

3. Комплексне емпіричне дослідження, яке демонструє, що ESR істотно зменшує розрив між доменами та підвищує продуктивність геолокалізації у реальних умовах.

Ми дійшли висновку, що додавання легких, апаратонезалежних ансамблевих стратегій в конвеєри з відновлення зображень є перспективним напрямом для надійної візуальної локалізації на основі БПЛА. Подальша робота полягатиме у дослідженні адаптивного зважування ансамблю та архітектур відновлення зображень з урахуванням домену для подальшого зменшення шуму та мінливості аерофотозображень.

Ключові слова: надвисока роздільна здатність, глибока нейронна мережа, трансформатори, згорткова нейронна мережа, комп'ютерний зір.

Інформація про авторів:

Ілечко Роман Ігорьович, магістр, аспірант, кафедра автоматизованих систем управління. **Email:** roman.i.ilechko@lpnu.ua; https://orcid.org/0009-0006-8208-5109

Боровий Орест Олегович, магістр, аспірант, кафедра автоматизованих систем управління. **Email:** orest.o.borovyi@lpnu.ua; https://orcid.org/0009-0008-2069-5336

Цимбал Юрій Вікторович, канд. техн. наук, доцент, кафедра автоматизованих систем управління. **Email:** yurii.v.tsymbal@lpnu.ua; https://orcid.org/0000-0001-9119-6771

Цитування за ДСТУ: Ілечко Р. І., Боровий О. О., Цимбал Ю. В. Ансамблеве підвищення роздільної здатності зображень для геолокації. *Український журнал інформаційних технологій*. 2025, т. 7, № 1. С. 68–76.

Citation APA: Ilechko, R. I., Borovyi, O. O., & Tsymbal, Y. V. (2025). Ensemble Image Super-Resolution for UAV Geo-localization. *Ukrainian Journal of Information Technology*, 7(1), 68–76. https://doi.org/10.23939/ujit2025.01.068