A Geo-consistent Weather Augmentation Method for Improving UAV Localization Robustness under Adverse Meteorological Conditions

Volodymyr Vozniak
Khmelnytskyi National University
Oleksander Barmak
Khmelnytskyi National University
Pavlo Radiuk
radiukp@khmnu.edu.ua
Khmelnytskyi National University
Iurii Krak
Taras Shevchenko National University of Kyiv

Research Article

Keywords: Visual Place Recognition (VPR), UAV, YOLO, weather augmentation, domain shift, segmentation

Posted Date: May 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9595632/v1

Additional Declarations: No competing interests reported.

A Geo-consistent Weather Augmentation Method for Improving UAV Localization Robustness under Adverse Meteorological Conditions

Volodymyr Vozniak[1], Oleksander Barmak[1], Pavlo Radiuk[1*], Iurii Krak[2,3]

[1]Department of Computer Science, Khmelnytskyi National University, 11, Institutes str., Khmelnytskyi, 29016, Ukraine.
[2]Department of Theoretical Cybernetics, Taras Shevchenko National University of Kyiv, 64/13, Volodymyrska str., Kyiv, 01601, Ukraine.
[3]Laboratory of Communicative Information Technologies, Glushkov Cybernetics Institute, 40, Glushkov Ave., Kyiv, 03187, Ukraine.

*Corresponding author(s). E-mail(s): radiukp@khmnu.edu.ua;
Contributing authors: vozniakvz@khmnu.edu.ua; barmako@khmnu.edu.ua; iurii.krak@knu.ua;

Abstract

Global visual localization of unmanned aerial vehicles (UAVs) in GNSS-denied scenarios experiences substantial accuracy degradation due to domain shift induced by fog, snow, and nighttime illumination. This work proposes Geo-consistent Weather Augmentation (GCWA) – a guided procedure for constructing weather domains using a state-of-the-art image-to-image generative model. The scientific novelty of the approach lies in an iterative Generate–Verify–Filter loop with a fail-closed principle, which preserves the topological integrity of objects through multi-criteria verification (Mask IoU, Edge-SSIM, and the weather-shift threshold τ_{shift}). The method is applied to train a YOLO11 model that extracts structural building features and forms a robust scene descriptor. Cross-domain evaluation shows that GCWA substantially mitigates the crucial degradation of baseline models in fog, improving mean Average Precision at an Intersection-over-Union threshold of 0.50 (mAP50) from 0.397 to 0.815. In experiments on the VPAIR (Aerial Visual Place Recognition and Localization in Large-scale Outdoor Environments) dataset, integrating YOLO11+GCWA yields the highest Recall@1 (R=3 frames) in fog, night, and snow, exceeding MixVPR by 1.4–1.6× and CosPlace by 2.2–2.7× depending on the domain; in the clean Original route, MixVPR remains slightly higher (0.210 vs 0.195). The results confirm that geometric consistency of augmentations is a key factor in learning robust semantic descriptors for UAV autonomous navigation in dynamic environments.

Keywords: Visual Place Recognition (VPR), UAV, YOLO, weather augmentation, domain shift, segmentation

JIRS categories: Categories (3) , (5) , (7) , (8)

MSC Classification: 68T07 , 68T40 , 68T45 , 68U10

1 Introduction

Ensuring autonomous navigation of unmanned aerial vehicles (UAVs) under degraded or completely unavailable global navigation satellite system (GNSS) signals remains one of the most pressing challenges in modern intelligent robotics. In scenarios where satellite positioning is infeasible due to obstructions or active jamming, visual place recognition (VPR) provides a reliable alternative for global localization [1], often relying on the extraction of distinctive visual landmarks from Earth imagery to achieve robust geo-localization [2]. Previous studies [3] confirmed the effectiveness of the YOLO11 architecture for precise building segmentation as a basis for generating compact descriptors in cross-view matching between UAV imagery and satellite maps.

Conventional data augmentation methods often fail to preserve the spatial topology and geometric accuracy of objects, which is crucial for pixel-level segmentation models. Applying modern generative models to simulate weather effects without strict structural control mechanisms leads to systematic artifacts – distortions of building contours and drift of object boundaries. This induces substantial noise during training, destabilizes the network's internal representations, and reduces the discriminative power of the learned descriptors.

The central hypothesis of this study is that training a segmentation model on synthetically generated weather domains while enforcing strict spatial invariance (geo-consistency) enables the neural network to learn robust representations that are invariant to photometric perturbations. We expect that using such features as descriptors for VPR will markedly improve UAV global localization accuracy under crucially reduced visibility, ensuring substantially more stable matching compared with models trained only on data collected in favorable weather conditions.

The main contributions of this work are summarized below.

- GCWA (Geo-consistent Weather Augmentation). We propose a guided approach for generating weather domains (fog/snow/night) in an image-to-image setting using a state-of-the-art generative model and spatial constraints that preserve the topological validity of annotations and minimize label noise in pixel-level segmentation tasks.

- Multi-criteria verification of geometric consistency (fail-closed). We develop an algorithm for selecting synthetic samples based on a combination of Mask IoU, Edge-SSIM, and the weather-shift threshold τ_{shift}; when no valid candidate exists, the image is rejected, which increases the reliability of augmentation for precision segmentation.
- Systematic cross-weather evaluation of segmentation robustness. We perform cross-domain testing of YOLO11 models across baseline/fog/snow/night domains and show that mixed training on baseline+fog+snow+night yields the best worst-case stability and high average quality across domains, substantially mitigating the crucial degradation of the baseline model in fog.
- Empirical validation of GCWA for UAV VPR localization. On the urban subset of the VPAIR dataset we demonstrate that YOLO11+GCWA improves Recall@1 (R=3 frames) in adverse weather scenarios, with the largest gains in fog/night/snow domains, while MixVPR remains slightly stronger on the clean Original route; these results confirm that geo-consistent training specifically strengthens the weather invariance of the descriptor under a fixed "clean" satellite database.

The remainder of the paper is organized as follows. Section 2 reviews contemporary approaches to visual localization and generative augmentation methods. Section 3 describes the proposed GCWA method. Section 4 presents the experimental evaluation, and Section 5 discusses the results, limitations, and directions for future research.

2 Literature Review

2.1 VPR Challenges in Dynamic Environments

VPR is a key component of UAV localization systems, especially when GNSS signals are unavailable or unreliable. The main challenges arise from dynamic environmental changes such as seasonal transformations (e.g., changes in vegetation from spring to winter) and weather factors (sunlight, cloudiness, rain, fog, snow), which substantially affect image appearance [1]. These changes lead to perceptual aliasing, where different places may appear similar, or to matching failures due to the loss of visual similarity [4, 5]. In the UAV setting, where images are compared with satellite imagery, the problem is further complicated by differences in viewpoint (aerial imagery versus orthophotos) and illumination, which create a domain shift [6]. For example, datasets such as Nordland and Oxford RobotCar exhibit degraded VPR performance in multi-session scenarios, where reference images are collected in summer while queries are acquired in winter or under rain [7].

Dynamic environments also include moving objects (pedestrians, vehicles) and long-term changes (construction, vegetation growth), which increase the sensitivity of visual sensors [8]. This is crucial for UAVs because flight trajectories often span large areas under varying conditions, where traditional methods based on local descriptors (e.g., SURF or ORB) fail due to photometric variations [9]. Studies show that integrating sequence-based matching and multimodal approaches (vision + LiDAR or radar) can mitigate these effects, since radar is less sensitive to weather [10, 11].

Nevertheless, despite progress, existing methods remain highly sensitive to severe appearance changes, especially those caused by adverse weather and illumination shift, underscoring the need for robust augmentation during training [4, 12].

2.2 Generative Models in Data Augmentation for Computer Vision

Generative models such as generative adversarial networks (GANs) and diffusion models are widely used to create synthetic training samples in computer vision, particularly for tasks where real data are limited [13]. GANs generate new images through competition between a generator and a discriminator, which is effective for augmentation in medical imaging or for simulating weather conditions, such as synthesizing images with rain or fog [14]. Diffusion models, such as Stable Diffusion, often outperform GANs in realism, enabling the generation of diverse backgrounds from text prompts while emphasizing semantic consistency [15]. Image-to-image translation methods such as CycleGAN or WeatherFlux transform images from one domain to another (e.g., day to night) while preserving the underlying structure [16].

Combining diffusion models with large language models (LLMs) further strengthens the process: adaptive prompt generation increases contextual diversity, while semantic guidance helps preserve salient foreground content during synthesis [17, 18]. Object-aware pipelines that segment scene instances before applying GAN-based translation have likewise been used to improve sim-to-real transfer in robot navigation [19]. This is particularly useful for simulating weather effects, where models are trained on synthetic data to increase robustness under low visibility [20]. However, traditional and generative methods often lack built-in mechanisms for strict spatial invariance control, which leads to "geometric noise" – small shifts in object boundaries in generated images – that induce errors in neural networks [21]. Table 1 provides a comparative analysis of different data augmentation methods for computer vision.

Table 1 Comparative analysis of different data augmentation methods for computer vision

Method	Model type	Use in augmentation	Advantages	Limitations
GAN-based weather translation [14]	Adversarial	Weather simulation, noise removal	High realism	Prone to mode collapse; weak structural preservation
Diffusion (e.g., Stable Diffusion) [15]	Diffusion-based	Prompt-based background generation	Semantic consistency; diversity	Computationally expensive; risk of artifacts
LLM + Diffusion (e.g., DALDA) [17]	Hybrid	Context-aware augmentation	Foreground preservation	Depends on prompt quality
Image-to-Image (e.g., CycleGAN) [16]	Translation-based	Domain translation (day-to-night)	Style invariance	Loss of fine-grained geometry

2.3 Geometric Integrity Issues in Diffusion and Generative Models

Geometric integrity is crucial for segmentation tasks, where generative models must preserve the topological structure of an image, including object boundaries and spatial relationships [22]. Diffusion models such as GIF or SIGD optimize latent features for semantic consistency but often generate artifacts that violate geometry, especially in tasks requiring accurate segmentation [23]. Image-translation models also face limitations: when the discrepancy between domains is large, generated samples may distort structure and degrade downstream segmentation performance [24].

Structural control can also be introduced through explicit topology-preserving priors and deformation constraints in segmentation-aware pipelines [25]. Location-aware objectives further encourage representations to respect spatial structure during segmentation pretraining [26]. However, these methods still do not provide a fail-closed verification stage that automatically rejects geometrically inconsistent synthetic samples. This highlights the need for procedures that retain only consistent augmentations.

2.4 Robust Descriptors Based on Semantic Segmentation

Robust descriptors leverage semantic segmentation and scene-structure cues to create landmarks that are less vulnerable to photometric changes [23, 27]. Prior work uses semantic masking and discriminative feature selection to emphasize stable regions and suppress dynamic distractors during localization and place recognition [22, 28]. Complementary augmentation strategies incorporate semantic guidance to improve the robustness of downstream visual models under adverse conditions [18, 29]. However, existing aerial and cross-view VPR benchmarks primarily evaluate recognition performance under viewpoint or appearance changes rather than providing dense adverse-weather semantic annotations, which limits supervised training and analysis of segmentation-driven descriptors in this setting [12, 30].

2.5 Summary and Identified Gaps

The above analysis of the state of the art and our prior results reveals several crucial gaps in UAV visual localization.

- **Sensitivity to extreme photometric changes.** Existing VPR methods, including building-segmentation-based approaches, achieve high accuracy under ideal conditions but experience significant degradation (Recall@1 drops by 50-70%) when transitioning to low-visibility domains (fog) or specific illumination settings (night).
- **The "geometric noise" problem in augmentation.** Traditional and generative weather-augmentation methods do not provide built-in mechanisms for strict spatial-invariance control. Even slight boundary shifts (drift) in generated images break consistency with segmentation masks, inducing errors in the neural network's internal representations.
- **Lack of verified weather datasets for VPR.** Most available aerial and cross-view benchmarks focus on viewpoint or seasonal changes rather than dense adverse-weather semantic annotation, which limits the training of robust models [12, 30].

- **Absence of fail-closed verification mechanisms.** In automatic training set expansion, there are no procedures to automatically select only those samples that reliably preserve the scene's topological integrity. Overcoming these limitations requires moving from simple image generation to a guided data curation process, where each synthetic appearance change is accompanied by a mathematical confirmation that the geometric structure is preserved. The goal of this work is to minimize the impact of atmospheric and illumination disturbances on UAV global localization accuracy by ensuring weather invariance of building-based descriptors through geo-consistent training.

To achieve this goal, the following tasks must be addressed:

1. Quantitatively characterize the impact of atmospheric and illumination domain shift (fog/snow/night) on building segmentation quality and feature stability in real-time models;
2. Develop and implement the GCWA method as a Generate–Verify–Filter procedure using an image-to-image generative model and a fail-closed selection mechanism that guarantees geometric consistency of synthetic samples with the original annotations;
3. Construct and verify a multi-domain building segmentation dataset (baseline/fog/snow/night) with controlled thresholds for Mask IoU, Edge-SSIM, and the weather-shift threshold τ_{shift}, while preventing data leakage between train and test sets;
4. fine-tune YOLO11 [31] on geo-consistent data and perform cross-domain testing using mAP50 and F1, forming a "train-domain – test-domain" transfer matrix to assess robustness;
5. conduct a comparative VPR evaluation on the urban subset of VPAIR [30] with a fixed satellite database and weather perturbations applied only to UAV queries; quantitatively compare Recall@1 (R=3 frames) against CosPlace [32], MixVPR [33], and the baseline YOLO version without GCWA, interpreting the contribution of geo-consistent training.

3 Materials and Methods

3.1 Mathematical Formalization of the Approach

The proposed robust localization strategy is based on the assumption that the geometric structure of the urban environment (in particular, building contours) constitutes a stable descriptive basis that is invariant to substantial photometric shifts.

Let $\mathcal{Q} = \{q\}$ denote the set of queries acquired from UAV sensors, and $\mathcal{R} = \{r_1, \ldots, r_n\}$ the database of georeferenced satellite images. The global localization task reduces to finding a mapping $\mathcal{F} : \mathcal{I} \to \mathcal{V}$ that transforms an input image $\mathcal{I}$ into a compact feature vector (embedding) in the descriptor space $\mathcal{V} \subset \mathbb{R}^d$.

In this work, the mapping $\mathcal{F}(\bullet)$ is implemented using a YOLO11 segmenter—an architecture whose recent variants have demonstrated exceptional efficacy in diverse UAV-based visual tasks, ranging from target localization to multispectral defect detection [34]. As the image descriptor we use the output of the last convolutional layer immediately before the segmentation head, which encodes the spatial characteristics of buildings. Specifically, for each input image we compute the feature tensor from the YOLO backbone, apply global pooling, and obtain a fixed embedding vector that serves as a compact scene descriptor. Because the model is optimized for precise building segmentation, these features capture generalized information about the geometry and spatial relationships of urban development and are less sensitive to photometric changes. The resulting vector is used without additional normalization. Using the L2 distance is appropriate because it accounts for both the direction and the magnitude of the features (which reflects the confidence/expressiveness of structural patterns) and provides a stable and computationally simple comparison metric for nearest-neighbor retrieval.

The optimal match index k for a query q is obtained by minimizing the distance in the metric space $\mathcal{V}$:

$$k = \arg \min_{1 \leq j \leq n} \|\mathcal{F}(q) - \mathcal{F}(r_j)\|_2 \tag{1}$$

where q is a query image captured by the UAV, r is a satellite image, and n is the total number of satellite images.

Localization is considered correct if the spatial distance between the ground-truth coordinates $\updownarrow(q)$ and the coordinates of the retrieved satellite image $\updownarrow(r_k)$ does not exceed the predefined tolerance radius τ_{dist}:

$$\mathrm{dist}\left(\updownarrow(q), \updownarrow(r_k)\right) \leq \tau_{dist} \tag{2}$$

where q is a query image from the UAV, and r is a satellite image.

Note that in the general case the tolerance radius τ_{dist} can be interpreted as a metric distance in coordinate space. In the VPAIR protocol used in this work, we apply a discrete correctness criterion based on sequence indices: localization is considered successful if the retrieved match lies within a $\pm R$-frame window relative to the ground truth.

To overcome domain shift, we introduce a guided transformation operator G that generates an image variation $\mathcal{I}_w$ for weather condition $w \in \{fog, snow, night\}$ while preserving the topological integrity of the original segmentation mask M.

The GCWA process is formalized as the operation of a stochastic generative operator G that synthesizes an augmented image variant $\mathcal{I}_w$:

$$I_w = \mathcal{G}\left(I,\, M,\, P_w,\, \eta\right) \tag{3}$$

where M is the corresponding semantic building mask acting as a spatial constraint, P_w is the text prompt that determines the target weather domain w; η is a random noise vector reflecting the stochastic nature of the generative process.

The following conditions must be satisfied: geometric invariance, structural similarity of boundaries, and sufficient strength of the weather shift.

Geometric invariance:

$$\mathrm{IoU}\left(M,\, \mathrm{Seg}\left(I_w;\, \theta_{\mathrm{ref}}\right)\right) \geq \tau_{iou} \tag{4}$$

where $Seg(\bullet)$ is the building-segmentation function implemented using the reference segmenter.

In this work, the reference segmenter is the YOLO11 model from our previous research [3], trained on the same Building Segmentation Dataset in the original domain (without GCWA) and then kept fixed during filtering. This design reduces circular validation risk because GCWA-generated samples are evaluated by a model that is not updated with those same synthetic domains. The threshold τ_{iou} guarantees similarity between the masks of the original and augmented images.

Structural similarity of boundaries:

$$\mathrm{SSIM}\left(\nabla I,\, \nabla I_w\right) \geq \tau_{edge} \tag{5}$$

where ∇ is a gradient operator (edge map) that ensures preservation of high-frequency structural components (roof contours and building edges), and τ_{edge} is a consistency threshold that guarantees the absence of edge drift.

Effectiveness of the weather shift:

$$\mathcal{D}\left(I,\, I_w\right) \geq \tau_{shift} \tag{6}$$

where $\mathcal{D}$ is a perceptual-difference metric that ensures the generated image contains a sufficient level of photometric perturbation for learning robustness, and τ_{shift} is a consistency threshold that guarantees the effectiveness of the weather shift.

As the perceptual-difference metric $\mathcal{D}$ we use LPIPS (Learned Perceptual Image Patch Similarity) with a typical backbone architecture (AlexNet) and a resized 256×256 RGB input. LPIPS is robust to minor local noise while remaining sensitive to substantial changes in illumination, contrast, and texture, which reflects the nature of meteorological effects. The threshold $\tau_{shift} = 0.1$ is selected as a practical compromise: at lower values, frames almost identical to the original enter the set and do not promote new invariances, whereas stricter values markedly reduce Acceptance% without improving segmentation quality.

Training the model $\mathcal{F}$ on GCWA-augmented data aims to minimize the combined loss function $\mathcal{L}_{total}$, which includes segmentation and classification losses, forcing the network to focus on weather-invariant geometric features:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{seg} + \lambda_2 \mathcal{L}_{box} + \lambda_3 \mathcal{L}_{dfl} + \lambda_4 \mathcal{L}_{cls} \tag{7}$$

where λ_1, λ_2, λ_3, λ_4 are weighting hyperparameters (regularization coefficients) that define the priority of each subtask during training.

They balance the contribution of each loss term to the overall gradient, preventing one task (e.g., object localization) from dominating another (e.g., embedding stability).

$\mathcal{L}_{seg}$ (segmentation loss) defines the pixel-level accuracy of building boundary delineation. In modern YOLO11 architectures, this component is typically based on Binary Cross Entropy (BCE) (to classify each pixel as "object" or "background") and Dice Loss (to improve segmentation under class imbalance when buildings occupy only a small part of the frame). It ensures high-quality masks, which is crucial for learning structure-oriented features.

$\mathcal{L}_{box}$ (boundary loss) minimizes the error in estimating the spatial position of a building as an integral object. In the YOLO architecture this term usually includes IoU Loss (CIoU or DIoU), metrics that account not only for bounding-box overlap

(Intersection over Union) but also for the distance between centers and aspect-ratio consistency.

$\mathcal{L}_{dfl}$ (distribution focal loss) helps more precisely localize object boundaries when they are not sharp (e.g., due to fog or blur). It stabilizes object localization in the image, helping the network "focus" on the relevant regions for descriptor extraction.

$\mathcal{L}_{cls}$ (classification loss) is responsible for correct identification of an object class in the image. In our case only the "building" class is considered, so this component plays a secondary role but supports stable feature extraction. Although this component is not a specialized metric-learning loss for VPR, together with segmentation and localization losses it stabilizes the semantic features formed by the network during training and guides the shared feature space toward reproducing relatively stable building contours. These features are subsequently used as image descriptors in the matching procedure.

The combination of these four components guides YOLO11 to learn an internal representation of buildings that is robust to photometric disturbances (in particular, fog and snow), rather than merely reproducing masks for a specific input scene.

3.2 Geo-consistent Weather Augmentation (GCWA) Method

To mitigate segmentation degradation under adverse weather, we developed the GCWA method, which is based on an iterative "Generate–Verify–Filter" cycle. Unlike standard image stylization techniques, GCWA turns augmentation into a guided data curation process in which each synthetic sample is checked for preservation of geometric accuracy. Figure 1 shows the overall scheme of the geo-consistent weather augmentation (GCWA) method.

In the first phase of GCWA, the guided transformation operator $\mathcal{G}$ is implemented using the Gemini 2.5 Flash Image [35] (sometimes referred to as "Nano Banana") model in image-to-image mode, where the input image serves as a basis for photometric transformation into the target weather domain while preserving scene geometry. To ensure reproducibility, we provide below the configuration of generative-model calls and the prompt structure.

Generation was performed via the Google Gemini API provider (Google AI for Developers / Google AI Studio), specifying the model identifier gemini-2.5-flash-image and a fixed "model/version" string returned by the API response: "models/gemini-2.5-flash-image".

The model parameters were fixed for all runs:

- Temperature: 1.0;
- topP: 0.95;
- topK: 64;
- candidateCount: 1;
- aspect ratio: 1:1;
- MIME type: "image/png".

To ensure stable results, fixed prompt templates were used, with only the domain variable being changed, $w \in \{fog, snow, night\}$. Table 2 provides an example of the prompt structure (presented as a template and reproducible verbatim).

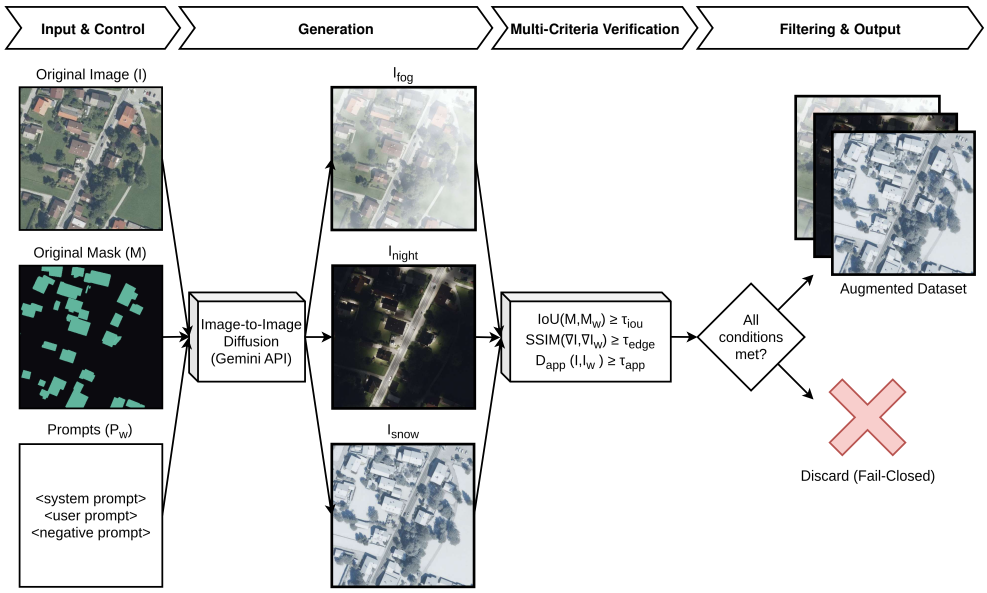

Fig. 1 General scheme of the Geo-consistent Weather Augmentation (GCWA) method based on the Generate–Verify–Filter (GVF) pipeline. The method ensures topological integrity of building footprints across multiple domains (fog, snow, night) by enforcing strict geometric and structural constraints before dataset expansion.

Table 2 Prompt configuration for Geo-consistent Weather Augmentation

Prompt type	Content
System Prompt	You are an image-to-image model. Do not alter geometry or scene layout–no warping, scaling, cropping, rotation, perspective changes. Do not add/remove/move any objects. Preserve all building footprints, roof outlines, edges, corners, and road edges exactly. Return the image at the exact same resolution.
User Prompt	Transform the image to <DOMAIN> conditions: <DOMAIN_DESCRIPTION>.
Negative Prompt	geometry change, warped buildings, shifted edges, new buildings, removed buildings, extra structures, changed roof shape, altered road layout, text, logos, artifacts, blurry edges, duplicated objects, deformation

Each generated candidate $\mathcal{I}_w$ undergoes comprehensive consistency analysis with respect to the original mask M through three independent control channels.

- Mask structural-consistency analysis. The masks of the original and generated images are compared according to Equation (4). This ensures that the main building regions remain in their original locations.
- Edge-similarity verification. Because IoU is robust to minor distortions, we additionally compute the structural similarity index (SSIM) between the gradient

maps of the original and generated images. This makes it possible to detect micro-deformations of building lines that can adversely affect descriptor precision.

- Minimum weather-shift strength control (threshold τ_{shift}). Additionally, we assess whether the generated frame differs sufficiently from the original for the augmentation to be meaningful for learning invariance. To this end, we compute the perceptual difference between the original and the candidate ($\mathcal{D}(\bullet)$, in this work – LPIPS). If $\mathcal{D} < \tau_{shift}$, the candidate is rejected as nearly devoid of weather perturbations.

GCWA follows a fail-closed principle: if none of the generated variants meets the verification thresholds, the image is excluded from augmentation. This is crucial to prevent label noise from entering the training set.

For each accepted sample, metadata are recorded (weather type, verification scores, model version), ensuring full reproducibility and enabling ablation studies of the impact of each weather condition on final VPR accuracy.

To preserve methodological rigor, augmentation is performed independently for each data split (train/test). This guarantees that the model never observes a weather-augmented version of a test image during training, ensuring an objective robustness evaluation under domain shift.

3.3 GCWA Algorithm

To systematize the expansion of the training set, the GCWA procedure is implemented as an iterative algorithm that provides automated quality control for each synthesized sample.

Line 6 in Algorithm 1 reflects the stochastic nature of the generation process: using different noise vectors η_k ensures visual diversity for the same prompt P_w. Line 13 uses the additive score $S_{iou} + S_{edge}$ as a composite geometric-quality metric, selecting a candidate that preserves both overall topology (via IoU) and boundary sharpness (via Edge-SSIM). Line 18 selects the best representative of the target domain, minimizing the contribution of "free" noise in the training set. Line 21 implements the training-safety principle: if none of the K variants satisfies the structural-stability criteria, the algorithm prefers reducing the dataset size rather than introducing potentially erroneous labels (label noise).

3.4 Datasets

To ensure a comprehensive evaluation of the proposed method, we use a combination of real and synthetically expanded datasets that cover both the training stage of the segmentation model and the final validation of the VPR system.

3.4.1 Building Segmentation Dataset

Initial training and fine-tuning of the YOLO11 model were performed on a specialized building segmentation dataset [36]. The dataset contains 9,665 high-resolution images covering diverse urban landscapes and architectural styles from the following locations: Tyrol (2,999 images), Tripoli (1,078), Kherson (1,053), Donetsk (999), Mykolaiv (739), Kharkiv (602), Mekelle (951), as well as other cities (1,244).

Algorithm 1 Geo-consistent Weather Augmentation (GCWA)

Require: Original dataset $\mathcal{D} = \{(I_i, M_i)\}_{i=1}^{N}$; weather domains $\mathcal{W}$; prompts $\{P_w\}$; thresholds $\tau_{iou}, \tau_{edge}, \tau_{shift}$; candidates per domain K.

Ensure: Augmented dataset $\mathcal{D}_{aug}$.

$\mathcal{D}_{aug} \leftarrow \mathcal{D}$
for all $(I, M) \in \mathcal{D}$ **do**
    for all $w \in \mathcal{W}$ **do**
        $\mathcal{V} \leftarrow \emptyset$ ▷ valid candidates for domain w
        for $k = 1$ **to** K **do**
            $\tilde{I}_w^{(k)} \leftarrow \mathcal{G}(I, M, P_w, \eta_k)$ ▷ Generate
            $\tilde{M}_w^{(k)} \leftarrow \mathrm{Seg}(\tilde{I}_w^{(k)}; \theta_{\mathrm{ref}})$
            $s_{iou} \leftarrow \mathrm{IoU}(M, \tilde{M}_w^{(k)})$
            $s_{edge} \leftarrow \mathrm{SSIM}(\nabla I, \nabla \tilde{I}_w^{(k)})$
            $d_{perc} \leftarrow \mathcal{D}(I, \tilde{I}_w^{(k)})$ ▷ LPIPS
            if $s_{iou} \geq \tau_{iou}$ **and** $s_{edge} \geq \tau_{edge}$ **and** $d_{perc} \geq \tau_{shift}$ **then**
                add $(\tilde{I}_w^{(k)},\ s_{iou} + s_{edge})$ to $\mathcal{V}$ ▷ Verify
            end if
        end for
        if $\mathcal{V} \neq \emptyset$ **then**
            select $\tilde{I}_w^{*} \in \mathcal{V}$ maximizing $(s_{iou} + s_{edge})$ ▷ Filter
            add $(\tilde{I}_w^{*}, M)$ to $\mathcal{D}_{aug}$
        else
            continue ▷ fail-closed: discard domain w for this image
        end if
    end for
end for
return $\mathcal{D}_{aug}$

Such geographic representativeness enables the model to learn generalized geometric building features, minimizing the risk of overfitting to a specific region. All images were resized to 640×640 pixels, which is the standard resolution for the YOLO11 architecture.

3.4.2 VPAIR Dataset for Localization Evaluation

To validate VPR robustness, we use the VPAIR (Aerial Visual Place Recognition and Localization in Large-scale Outdoor Environments) dataset [30], an aerial visual place recognition and localization benchmark for large-scale outdoor environments, recorded from a light aircraft at an altitude of over 300 m with a downward-facing camera. Each query frame is paired with a high-detail reference render accompanied by a depth map and a 6-DoF pose; the positional component is derived from GPS coordinates, while orientation is provided in a local coordinate system (NED convention). The dataset covers a trajectory of more than 100 km and contains 2,706 “query – reference” pairs, as well as a set of distractor images to make the retrieval scenario more challenging.

In this study we focus on the urban subset of VPAIR, where buildings and road infrastructure provide the most discriminative visual landmarks for descriptor formation. Therefore, the robustness claims reported in this paper are validated only for dense built-up urban areas and are not yet experimentally confirmed for rural or unstructured environments.

3.4.3 Synthetic Weather Domains

To evaluate cross-weather robustness, the original images from the building-segmentation dataset as well as UAV query images from VPAIR were transformed using GCWA. Three main domains were constructed:

- fog – characterized by reduced contrast, depth-dependent light scattering, and degradation of high-frequency background details;
- snow – includes simulation of snow cover on horizontal surfaces (roofs, roads, fields) and changes in the scene's color temperature;
- night – assumes a global reduction in exposure with the introduction of local artificial light sources while preserving geometric building contours.

Owing to the verification procedure in the GCWA algorithm, all synthetic domains preserve geometric consistency with the original georeferenced annotations within the specified thresholds, which enables an objective comparison of localization accuracy across meteorological conditions under an unchanged satellite database.

To prevent information leakage (data leakage), the split into training and test subsets (80/20) was performed prior to augmentation. Final training of the YOLO model was carried out using seven folds to ensure the statistical significance of the results (mean and standard deviation).

3.5 Experimental Setup

The experimental part of the study aims to test the hypothesis of weather invariance of the learned descriptors and consists of two stages: fine-tuning the segmentation network and evaluating VPR performance.

3.5.1 Segmentation Model Configuration

To extract structural features, we use the YOLO11 architecture [31] (segmentation version). Fine-tuning was performed with the following hyperparameters.

- Number of epochs: 100.
- Input resolution: 640×640 pixels.
- Optimizer: AdamW with an initial learning rate 10^{-3} and weight decay of 0.05.
- Loss function: the weighting coefficients in Equation (7) were set empirically as $\lambda_1 = 7.5$ (segmentation), $\lambda_2 = 7.5$ (box), $\lambda_3 = 1.5$ (DFL), and $\lambda_4 = 0.5$ (classification) to ensure a dominant influence of geometric features on gradient formation.

3.5.2 GCWA Verification Parameters

During the execution of Algorithm 1 (GCWA), the following crucial thresholds were set for filtering synthetic samples.

- Geometric-consistency threshold (τ_{iou}): ≥ 0.85. Choosing this high value guarantees that building masks remain precisely accurate even after radical photometric changes.
- Boundary-similarity threshold (τ_{edge}): ≥ 0.7. This ensures preservation of high-frequency information about built-up contours.
- Weather-shift threshold (τ_{shift}): ≥ 0.1 (LPIPS). This filters out samples that are visually almost indistinguishable from the original, ensuring that the augmentation meaningfully contributes to robustness training.
- Number of candidates (K): 1. One candidate is generated for each input image and each target weather domain (given the computational cost of generation).

3.5.3 Computational Environment

All computations were performed on an Ubuntu workstation using the following components.

- Hardware: Nvidia ASUS RTX 5070 Ti GPU (16 GB VRAM), AMD Ryzen 9.
- Software: Python 3.10, the PyTorch library, and the Ultralytics framework for YOLO11.
- Generative module: the Gemini 2.5 Flash Image model integrated via an API for batch image processing.

3.5.4 Evaluation Criteria

Segmentation quality was evaluated using mean Average Precision at an Intersection-over-Union threshold of 0.50 (mAP50) and F1-score metrics [37]. Localization effectiveness was assessed using Recall@1 with a localization radius of R=3 frames [12] (± 3 neighboring indices along the flight trajectory), following the VPAIR protocol. This setup is strict because it requires correct ranking of the nearest match within a local window of sequential frames. Because VPAIR verifies correctness using sequence indices, we use the notation R=3 frames in the remainder of the text to avoid ambiguity with metric units.

3.6 Evaluation Setup

The proposed approach is evaluated in two stages: analysis of building-segmentation robustness and measurement of UAV global localization accuracy under cross-domain shift.

3.6.1 Segmentation Evaluation Protocol

To quantitatively assess the robustness of YOLO11 to atmospheric disturbances, we implement a cross-weather building-segmentation robustness strategy that includes three scenarios.

- Baseline mode. The model is trained exclusively on original images under favorable conditions and tested on the Fog, Snow, and Night domains. This measures degradation under pure domain shift.
- Augmented mode (proposed). The model is trained on a combined dataset expanded using GCWA. We assess the network's ability to generalize features under the simultaneous influence of different photometric disturbances.
- Weather-specific analysis. Separate models are trained for each weather domain and then tested in an "each-on-each" scheme to build a cross-domain transfer matrix.

Segmentation quality is evaluated using mAP50 and F1-score, where the latter serves as a balanced indicator of robustness against missed detections under challenging conditions.

3.6.2 VPR Evaluation Protocol

The main evaluation of localization effectiveness is carried out on the urban segment of the VPAIR trajectory. The evaluation protocol is based on the following principles.

- Fixed reference. The satellite database consists exclusively of "clean" original images. This mimics a realistic scenario in which satellite imagery is static and typically has high visual quality.
- Variable query. UAV queries are subjected to weather perturbations via GCWA, which makes it possible to isolate the impact of adverse weather specifically on the airborne part of the localization system.
- Success metric. Recall@1 (R=3 frames) is used. A query is considered successful if the top retrieval result, measured by L2 distance in descriptor space, corresponds to a position within ± 3 neighboring frames (sequence indices) around the ground-truth location along the VPAIR trajectory.

To verify the competitiveness of the proposed method, we compare against:

- CosPlace [32] – a modern state-of-the-art baseline for global descriptors that uses a classical CNN backbone to extract features without semantic object anchoring;
- MixVPR [33] – a strong holistic global-descriptor baseline based on feature mixing, which improves retrieval quality without explicit semantic anchoring;
- YOLO without GCWA [3] – the baseline version of our approach without geo-consistent training, to assess the isolated contribution of GCWA.

This evaluation setup not only confirms the effectiveness of the proposed algorithm but also enables a detailed analysis of how geometric segmentation stability translates into reliable UAV navigation under extreme conditions.

4 Results

4.1 Sensitivity Analysis of GCWA Thresholds

Geo-consistent Weather Augmentation (GCWA) includes the Verify stage, during which each synthetic sample is filtered using three thresholds: τ_{iou} (mask geometric invariance), τ_{edge} (preservation of contour structure), and τ_{shift} (minimum strength of the weather shift). To justify the operating configuration, a sensitivity analysis of the system was conducted. The evaluation was performed on a subset of 1,000 source images with parameter $K = 1$ (one candidate per domain).

Acceptance% denotes the fraction of source images for which at least one synthetic variant passed the Verify filter. Avg valid candidates (per image) reflects the average number of accepted candidates per source image across all domains. The resulting segmentation quality after training on the constructed augmented data is reported using the mAP50 metric in Table 3.

Table 3 Sensitivity analysis of GCWA verification thresholds $\boldsymbol{\tau}$

τ_{iou}	τ_{edge}	τ_{shift}	Acceptance%	Avg valid candidates	Augmented size	mAP50
0.8	0.7	0.1	99	2.25	2253	0.778
0.85	**0.7**	**0.1**	**96**	**1.9**	**1902**	**0.781**
0.9	0.7	0.1	88	1.55	1548	0.778
0.85	0.6	0.1	99	2.35	2354	0.778
0.85	0.8	0.1	86	1.45	1447	0.777
0.85	0.7	0.05	98	2.05	2051	0.779
0.85	0.7	0.15	88	1.65	1646	0.776

Experimental results show that lowering the thresholds (τ_{iou} or τ_{edge}) increases Acceptance% to 99% and expands the resulting dataset to 2253–2354 images, but mAP50 remains at 0.778. Conversely, increasing the thresholds (τ_{iou} or τ_{edge}) reduces Acceptance% to 88% and 86%, respectively; Augmented size decreases to 1548 and 1447 samples, and mAP50 decreases to 0.778–0.777 for the respective configurations ($\tau_{iou} = 0.8$ and $\tau_{edge} = 0.6$; $\tau_{iou} = 0.9$ and $\tau_{edge} = 0.8$).

The maximum mAP50 value of 0.781 is achieved for the configuration $\tau_{iou} = 0.85$, $\tau_{edge} = 0.7$, $\tau_{shift} = 0.1$, where Acceptance% is 96% and the sample size is 1902. Varying the threshold shows that at 0.05 mAP50 is 0.779 (Acceptance% = 98%), whereas at 0.15 it decreases to 0.776 with more candidates filtered out (Acceptance% = 88%). These quantitative relationships indicate the presence of a characteristic break point (knee-point) in the threshold parameter space.

4.2 Cross-Domain Segmentation Evaluation

To evaluate the robustness of YOLO11 segmentation models to changes in weather and illumination conditions, we used a single fixed split of the original image set into train/test (80/20). Based on this split, four data variants were created by applying GCWA visual transformations simulating fog, snow, and night, along with the baseline variant without modifications. In addition, an augmented scenario was formed in which the training and test subsets were expanded fourfold by merging all four domains (baseline + fog + snow + night).

Visual examples of the generated weather augmentations are shown in Figure 2. During generation, geometric/contour consistency of the scene and building outlines is preserved, while photometric characteristics (contrast, illumination, atmospheric effects) change according to the target domain.

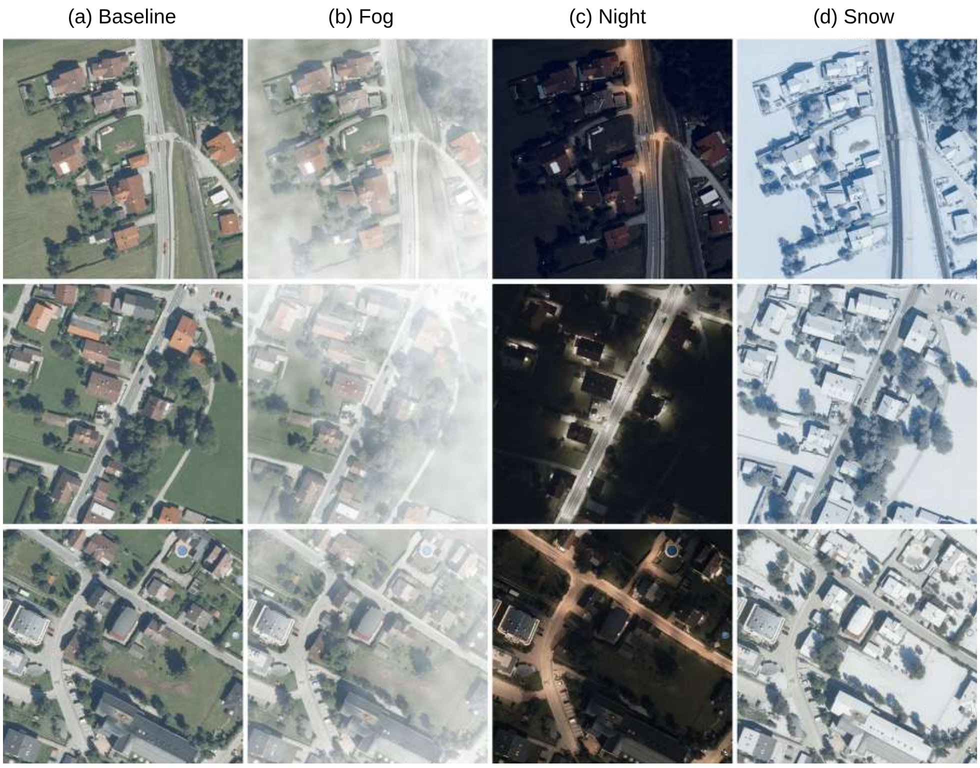

Fig. 2 Visual examples of weather augmentations generated using the GCWA method for 3 random samples of building segmentation dataset: (a) original image (baseline); (b) fog; (c) night with artificial lighting; (d) snow.

For each domain, the model was trained on the corresponding train set and then evaluated on the test set across five domains. The results are presented as a cross-testing matrix, where rows correspond to the training domain and columns to the testing domain. The intersection of a row and a column reflects segmentation quality: the matrix diagonal corresponds to in-domain performance, whereas off-diagonal elements represent quality under domain shift. Augmented uses the union of data from all domains, i.e., Baseline + Fog + Night + Snow.

Table 4 reports the cross-validation results in terms of mAP50, which was selected as the primary metric for these experiments.

A similar robustness evaluation was performed using the F1-score metric (Table 5), which makes it possible to assess the balance between segmentation precision and recall.

According to the results, the maximum metric values are achieved on the diagonal (in-domain): baseline-baseline (0.826 mAP50), fog-fog (0.816 mAP50), snow-snow (0.751 mAP50), and night-night (0.700 mAP50). When moving from “clean” conditions (baseline test) to fog (baseline-fog), mAP50 drops from 0.826 to 0.397 (-0.429

Table 4 Cross-condition performance matrix (mAP50) for YOLO11 segmentation models

Training Domain	Baseline	Fog	Night	Snow	Augmented
Baseline	**0.826**	0.397	0.577	0.573	0.590
Fog	0.820	**0.816**	0.606	0.606	0.714
Night	0.742	0.390	0.700	0.675	0.623
Snow	0.722	0.349	0.529	0.751	0.583
Augmented	0.824	0.815	**0.724**	**0.753**	**0.781**

Table 5 Cross-condition performance matrix (F1) for YOLO11 segmentation models

Training Domain	Baseline	Fog	Night	Snow	Augmented
Baseline	**0.816**	0.448	0.631	0.616	0.629
Fog	0.814	**0.811**	0.656	0.655	0.736
Night	0.751	0.441	0.718	0.694	0.660
Snow	0.721	0.383	0.569	0.769	0.618
Augmented	0.815	**0.811**	**0.739**	**0.772**	**0.784**

absolute points). In contrast, the model trained in the augmented scenario remains stable across all test domains: mAP50 ranges from 0.724 (night) to 0.824 (baseline).

4.3 YOLO Fine-tuning Results

To ensure robustness of the visual localization system, we further trained YOLO11 on the combined set of original and augmented data (augmented scenario). Figure 3 shows the loss curves (box_loss, seg_loss, cls_loss, dfl_loss) on the training (train) and validation (val) sets over 100 epochs.

According to the curves, train loss monotonically decreases throughout the entire training cycle. Validation loss decreases rapidly at the early stages (up to epochs 20-30) and then reaches a plateau.

Figure 4 shows the confusion matrix for the final model. Evaluation was performed for two classes: "Building" (target object) and "Background". The total number of correctly identified building fragments is 32,957, whereas the numbers of false positives (background-building) and missed detections (building-background) are 12,312 and 11,985, respectively.

Typical false-positive cases are visualized in Figure 5. In the left example, the model highlights objects in a coastal area near a road, where geometric patterns imitate roof contours. In the right example, the model responds to elongated objects of coastal infrastructure (piers/boats) near the water surface.

Quantitative performance metrics of the model across seven different data splits (folds) are reported in Table 6. For each metric, the mean value and the standard deviation were computed.

Figure 6 shows the Precision-Recall curve for the best train/test split. The area under the curve (AUC) for the model is 0.76.

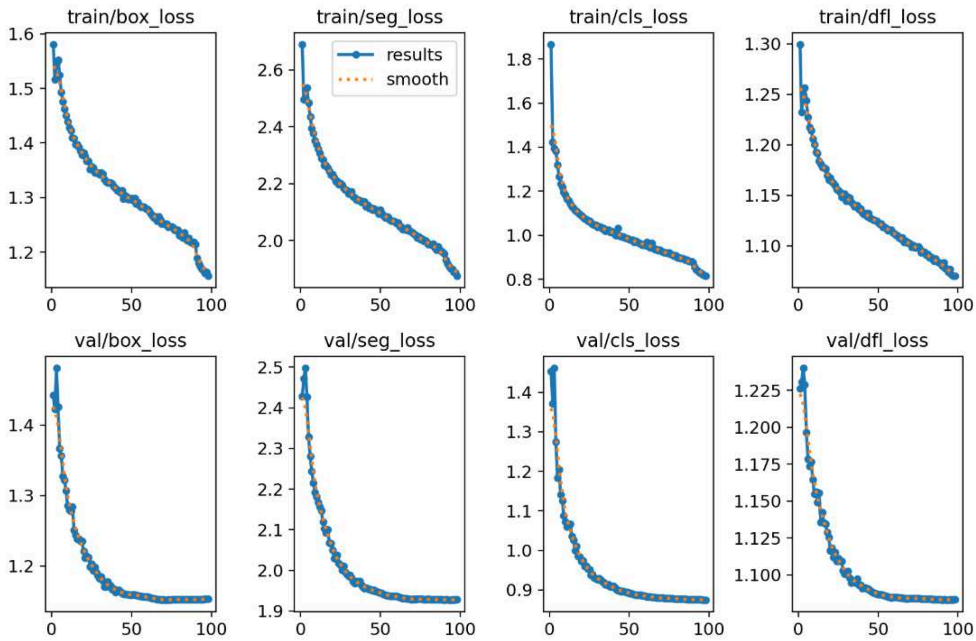

Fig. 3 Loss curves for the training (train) and validation (val) datasets.

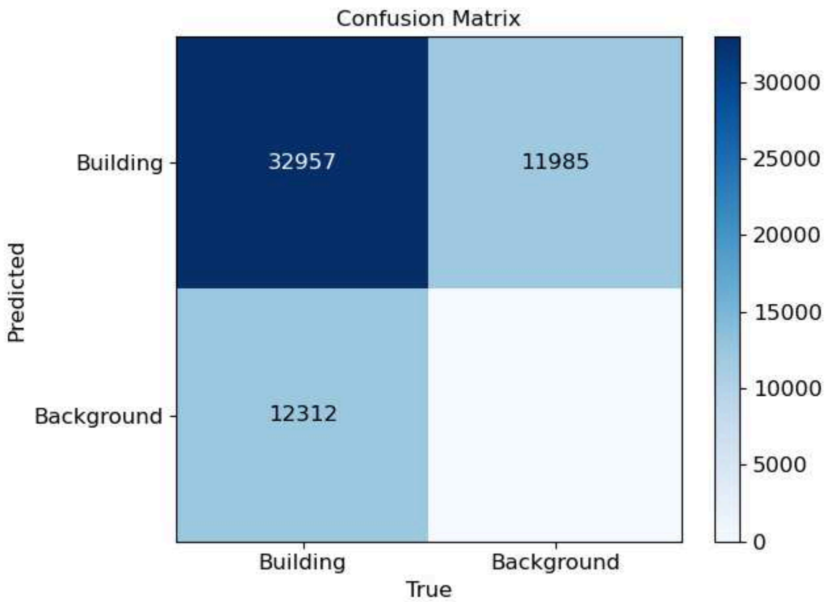

Fig. 4 Confusion matrix for the fine-tuned YOLO11 model on the building-segmentation dataset.

Fig. 5 Examples of false positives of the building-segmentation model.

Table 6 Metric values of the fine-tuned YOLO11 model on the building-segmentation dataset

Metric	Split	Fold							Mean±SD
		1	2	3	4	5	6	7	
mAP50	Train	0.792	0.795	0.796	0.791	0.794	0.793	0.795	0.794±0.0018
	Test	0.723	0.725	0.727	0.722	0.726	0.724	0.725	**0.725±0.0017**
Recall	Train	0.706	0.709	0.712	0.705	0.710	0.707	0.708	0.708±0.0024
	Test	0.663	0.666	0.668	0.662	0.667	0.660	0.665	**0.664±0.0029**
Precision	Train	0.812	0.815	0.817	0.811	0.816	0.813	0.814	0.814±0.0022
	Test	0.772	0.775	0.776	0.771	0.774	0.770	0.773	**0.773±0.0022**
F1	Train	0.755	0.758	0.761	0.754	0.759	0.756	0.757	0.757±0.0023
	Test	0.713	0.716	0.718	0.712	0.717	0.711	0.715	**0.715±0.0026**

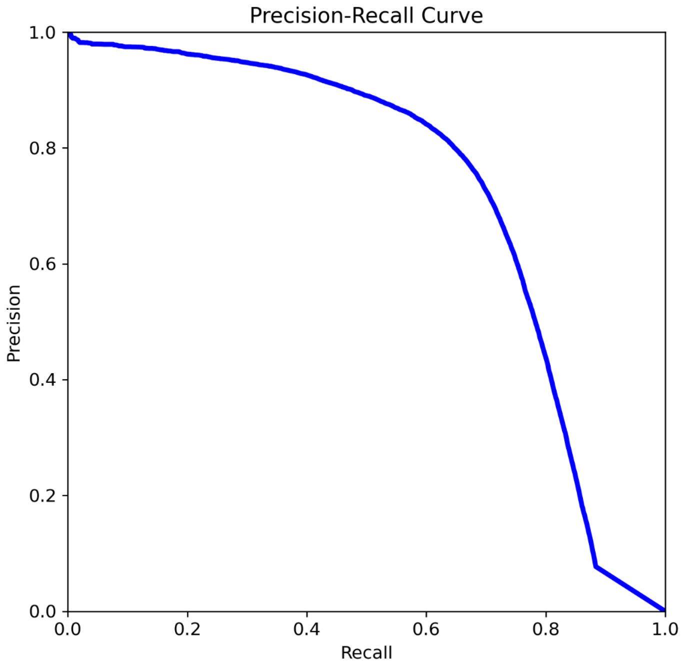

Fig. 6 Precision–Recall curve for the fine-tuned YOLO11 model on the building-segmentation dataset.

4.4 VPR Pipeline Using VPAIR Dataset

The effectiveness of the proposed method for global localization was evaluated on the urban segment of the VPAIR route. The experimental protocol uses a fixed satellite database ("clean" images), while UAV-side queries are subjected to weather perturbations via GCWA.

Table 7 reports a comparison of four approaches using Recall@1 (R=3 frames), where R=3 denotes a ±3-frame tolerance in the VPAIR sequence: the global-descriptor baselines CosPlace [32] and MixVPR [33], the baseline YOLO version without geo-consistent training [3], and the proposed YOLO method with GCWA.

Table 7 Recall@1 (R=3 frames) on the urban area of the VPAIR dataset under different weather conditions

Method	Original	Fog	Night	Snow
CosPlace [32]	0.140	0.045	0.051	0.056
MixVPR [33]	**0.210**	0.074	0.085	0.089
YOLO without GCWA [3]	0.181	0.079	0.083	0.091
YOLO with GCWA (Ours)	0.195	**0.121**	**0.139**	**0.125**

According to the results, MixVPR achieves the highest Recall@1 in the Original domain (0.210), slightly outperforming YOLO11+GCWA (0.195) and YOLO11 without GCWA (0.181). This indicates that under clean conditions a strong holistic descriptor can exploit fine photometric cues very effectively.

Under adverse weather, however, YOLO11+GCWA becomes the best-performing method in all perturbed domains. In the Fog domain, Recall@1 reaches 0.121, compared with 0.074 for MixVPR, 0.045 for CosPlace, and 0.079 for YOLO11 without GCWA. In the Night domain, YOLO11+GCWA reaches 0.139, compared with 0.085 for MixVPR, 0.051 for CosPlace, and 0.083 for the baseline YOLO11. In the Snow domain, the proposed method achieves 0.125, whereas YOLO11 without GCWA, MixVPR, and CosPlace obtain 0.091, 0.089, and 0.056, respectively.

Relative to MixVPR, YOLO11+GCWA gains +0.047 in fog, +0.054 in night, and +0.036 in snow, while MixVPR retains a +0.015 advantage in the clean Original domain. Relative to the baseline YOLO11, the gains in challenging domains are +0.042 for fog, +0.056 for night, and +0.034 for snow.

5 Discussion

5.1 Interpretation of Results and Weather Invariance

The sensitivity-analysis results (Table 3) confirm the crucial role of multi-criteria sample selection. The observed "knee-point" phenomenon at $\tau_{iou} = 0.85$ indicates that for precision building segmentation, simple visual similarity (stylization) is insufficient. Even minor geometric boundary drift in generated images (drift), which may be imperceptible to the human eye, induces substantial label noise. This destabilizes gradient-based training, because the network receives conflicting signals about pixel-level boundary locations. The fail-closed principle filters out invalid samples, leading to higher mAP50 even with a smaller training set.

Cross-domain evaluation (Table 4 and Table 5) reveals pronounced asymmetry: models trained on "difficult" domains (fog, night) generalize to "clean" conditions (baseline) better than vice versa. The most substantial performance drop of the baseline model in fog (from 0.826 to 0.397 mAP50) is consistent with the physical nature of this phenomenon: fog simultaneously reduces edge sharpness and scene contrast. Standard photometric augmentations do not reproduce such a combination of effects, whereas modern image-to-image generative models can synthesize weather artifacts while accounting for scene context. Training on the mixed (augmented) set encourages the network to extract structural building features that are relatively invariant to photometric disturbances, effectively serving as semantic regularization.

The comparative Recall@1 evaluation (Table 7) reveals a clear trade-off between clean-condition peak accuracy and adverse-weather robustness. Among holistic global descriptors, MixVPR is clearly stronger than CosPlace: in the Original domain it reaches 0.210 versus 0.140, and it also improves the perturbed domains to 0.074/0.085/0.089 for fog/night/snow. This confirms that feature mixing provides a more powerful global aggregation strategy than the CosPlace baseline when the descriptor must represent the entire scene.

However, YOLO11+GCWA yields the best results in all adverse-weather domains, with Recall@1 values of 0.121, 0.139, and 0.125 for fog, night, and snow, respectively. The absolute gains over MixVPR are +0.047, +0.054, and +0.036, which correspond to relative improvements of approximately 1.64×, 1.64×, and 1.40×. The key observation is that MixVPR remains dependent on holistic appearance cues distributed over the full image; these cues are strongly perturbed by fog scattering, snow cover, and nighttime illumination. In contrast, YOLO11+GCWA focuses on building contours as relatively stable elements of the urban landscape. Geo-consistent training increases descriptor invariance and preserves correct ranking of the nearest matches in a fixed satellite database even under low photometric similarity between the UAV query and the satellite reference. This confirms that, for reliable UAV localization in GNSS-denied scenarios, the topological integrity of semantic landmarks should be prioritized over the photorealism of synthetic transformations when adverse-weather robustness is the main objective.

At the same time, MixVPR attains the highest score in the clean Original domain (0.210 versus 0.195 for YOLO11+GCWA). This indicates that holistic descriptors can better exploit fine-grained photometric and texture cues when query and reference domains are closely matched. Therefore, the proposed semantic approach should be interpreted not as a universal replacement for modern global descriptors, but as a robustness-oriented alternative that sacrifices a small amount of clean-condition peak accuracy in exchange for significantly better worst-case behavior under weather shift.

The absolute Recall@1 values nevertheless remain modest (0.195 in the Original domain and 0.121–0.139 in adverse weather), meaning that top-1 retrieval still misses a large fraction of queries. We therefore interpret the current results as an incremental but meaningful step toward robust UAV localization rather than a fully mature standalone navigation solution. In this study, we intentionally evaluate a lightweight single-frame CNN pipeline optimized for computational efficiency, without sequence-level re-ranking, geometric verification, or multimodal fusion. Under this strict and fast setting, the observed relative gains over CosPlace, MixVPR, and YOLO without GCWA are practically important, and they define a strong foundation for further improvements via temporal filtering and sensor fusion.

5.2 Limitations and Open Questions

Despite the demonstrated effectiveness of GCWA in improving localization robustness, the study reveals a number of limitations that should be considered for practical deployment.

The GCWA procedure implements an iterative Generate–Verify–Filter loop with a single candidate generation per input image for each target domain (K=1). With multiple weather domains, the number of calls to the generative model scales linearly with the dataset size and the number of domains. Additional computational overhead is determined by multi-criteria verification (computing Mask IoU, Edge-SSIM, and the weather-shift metric $\mathcal{D}(\bullet)$ for τ_{shift}), which increases data-preparation time compared with standard augmentations. In practice, this constrains rapid expansion of training sets to millions of images without parallel processing and efficient management of API calls.

In this study, the generative transformation operator is implemented via a cloud API (Google Gemini), which can impose practical limitations related to service availability, usage cost, and model-version changes over time. In addition, the inability to run GCWA fully autonomously offline reduces its potential applicability in closed environments. The use of a cloud provider also introduces reproducibility risks: since generative models operate as "black boxes", the provider can update weights or adjust internal safety policies and filtering parameters, which may affect the quality of synthesized weather effects even under identical prompts. Moreover, reliance on stable Internet connectivity and API rate limits restricts the possibility of autonomous fine-tuning directly onboard the UAV or in field conditions. Importantly, the cloud generative module (Gemini) is used exclusively at the data-preparation and offline-training stage. During deployment (onboard inference), the UAV performs only a forward pass of YOLO11 and nearest-neighbor retrieval in the descriptor database, which does not require Internet access and does not depend on the external API.

The effectiveness of the method crucially depends on the choice of thresholds τ_{iou} and τ_{edge}. Setting overly strict criteria leads to a low Acceptance%, which can cause underfilling of the training set and loss of potentially useful hard cases. Conversely, overly lenient thresholds increase the risk of admitting "geometric noise", which degrades segmentation accuracy. Another consequence of the fail-closed principle is a possible imbalance in data difficulty: scenes that are hardest to generate/verify (e.g., dense urban layouts with complex shadows) are rejected more often, which may shift the training distribution toward simpler geometric structures.

This study focuses on three of the most crucial factors (fog, snow, night). However, real UAV operating conditions include a wider spectrum of disturbances: heavy rain (with wet reflections and lens distortions), sandstorms, haze, or radical seasonal changes in vegetation. The question of how invariant the learned descriptors remain to combined effects (e.g., night rain or fog with snow) remains open and requires further experimental validation. In addition, the current protocol assumes a static and "clean" satellite database, which does not account for potential domain shift on the reference side (outdated imagery or seasonal discrepancies).

The current implementation is based on building segmentation as the primary invariant landmark class in urban environments. In non-urban or natural scenes (fields, forests), the effectiveness of such a descriptor may be lower, which motivates extending the set of segmented classes and adapting the GCWA procedure to other semantic categories. Accordingly, robustness outside dense built-up areas has not been experimentally validated in the present study.

5.3 Segmentation Reliability and Interpretation of Model Errors

Cross-validation results across seven independent random splits (Table 6) confirm the high reliability of the proposed data-preparation pipeline. The low standard deviation across all key metrics (SD does not exceed 0.3%) indicates that the reported segmentation quality is not an artifact of a particular train/test split or an accidentally favorable subset. This supports reproducibility of the results and statistical stability of YOLO11 under variations in urban development. The loss dynamics (Figure 3) show proper convergence: a monotonic decrease of the loss on the training set and its stabilization on validation without signs of divergence indicate the absence of overfitting even under substantial expansion of the training set with synthetic data.

The confusion matrix (Figure 4) details the nature of the model's misses and indicates the absence of an undesirable bias toward a single error type. Comparable values of false positives (background-building) and misses (building-background) mean that the model is neither overly "aggressive" (with FP dominating) nor overly "conservative" (with FN dominating). This corresponds to stable segmenter behavior, where most errors stem from typical hard cases: low-contrast roofs, thin building fragments, or background with similar texture. The shape of the Precision-Recall curve (Figure 6) with AUC=0.76 confirms the possibility of flexibly tuning the operating threshold depending on mission requirements – either to minimize spurious landmarks (high precision) or to maximize detection of all present objects (high recall).

The method was compared with CosPlace and MixVPR as representative global-descriptor baselines. However, there remains a broad range of modern VPR methods that may demonstrate better performance in specific domains. Future work will extend the comparison to a larger set of baselines to fully assess the competitiveness of GCWA.

Analysis of erroneous segmentations (Figure 5) indicates that the majority of false positives arise from texture patterns resembling building roofs: coastal infrastructure, boats, and piers. This points to the need for additional semantic context or introducing "water" and "road" classes to reduce spurious detections. Some errors are also associated with complex roof geometries in dense urban development, which may require higher input resolution and stronger boundary-aware losses during training. A qualitative analysis indicates that these errors are not random but semantically driven: semantic aliasing arises from visual similarity between objects of different categories in remote-sensing imagery. Importantly, such false positives are usually local and occupy a small area, which makes them amenable to post-processing using simple geometric filters (e.g., compactness or a minimum-area criterion) without degrading the overall Recall of the localization system.

5.4 Directions for Future Research

The results of this work lay the groundwork for several promising directions for further development in UAV visual navigation.

Future studies should integrate faster diffusion models or local generative networks to enable scalable augmentation without reliance on external APIs. Batch generation and parallel processing can also be used to reduce dataset-preparation time. To address the high computational cost of the generative loop, we will investigate knowledge distillation methods. This would make it possible to train a compact local model that imitates weather effects without calling heavy cloud APIs. Mechanisms for parallelization and intelligent caching of generated artifacts are also promising and can substantially accelerate preparation of multi-domain datasets for large territories.

Future research will focus on replacing fixed verification thresholds τ_{iou} and τ_{edge} with adaptive algorithms based on neural networks. This will allow dynamic adjustment of geometric-consistency requirements depending on landscape type and scene complexity. A separate direction is the development/calibration of the metric for τ_{shift} ($\mathcal{D}(\bullet)$) while accounting for the specifics of weather effects (e.g., a learned "weather strength" predictor or a domain-specific perceptual metric) to reliably filter out nearly identical frames even in cases of weak or ambiguous visual changes.

Although the present work focuses on urban built-up areas, adapting GCWA to forest and agricultural landscapes is an important step toward a universal localization system. This will require extracting other semantic landmarks (road networks, hydrography, and forest boundaries) and accounting for substantial seasonal changes in vegetation appearance. Future work will evaluate robustness under a "double" domain shift, where both the UAV query and the satellite database are affected by different temporal or weather distortions.

Future work will focus on optimizing the inference pipeline and implementing the full localization stack on embedded platforms. This includes reducing the computational cost of segmentation and descriptor extraction and accelerating retrieval (e.g., via approximate nearest-neighbor search) to enable near real-time localization onboard the UAV. The ultimate goal is a lightweight architecture capable of performing segmentation and feature extraction in real time on embedded computers (e.g., the NVIDIA Jetson series). This would enable a fully autonomous navigation system that not only uses pre-trained robust features but is also capable of self-adaptation (online domain adaptation) during flight in dynamically changing meteorological conditions.

6 Conclusion

This study addresses the degradation of UAV visual localization in GNSS-denied scenarios under adverse weather using the Geo-consistent Weather Augmentation (GCWA) framework. Employing a Generate–Verify–Filter loop with strict spatial constraints (Mask IoU, Edge-SSIM, τ_{shift}), GCWA prevents geometric noise during pixel-level segmentation training. Empirical results confirm GCWA mitigates baseline performance collapse, improving YOLO11 segmentation mAP50 in fog from 0.397 to 0.815. On the VPAIR dataset, YOLO11+GCWA achieved the highest Recall@1 under fog (0.121), night (0.139), and snow (0.125), outperforming MixVPR by up to 1.6× and CosPlace by 2.7×, despite MixVPR's slight edge in clean conditions (0.210 vs 0.195). However, limitations include high computational overhead during data preparation, reliance on external cloud generative APIs, static verification thresholds, and a narrow focus on urban building footprints.

Future research will address these constraints by leveraging knowledge distillation for fast, local image generation, introducing adaptive neural-driven thresholds, and expanding semantic categories to encompass rural landscapes (e.g., roads, hydrography). Ultimately, optimizing the inference pipeline for resource-constrained embedded hardware will enable a fully autonomous, real-time UAV navigation system capable of robust, domain-adaptive localization across diverse meteorological conditions.

Declarations

- **Funding:** The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

- **Competing interests:** The authors have no relevant financial or non-financial interests to disclose.
- **Ethics approval:** This study did not involve human participants, human data, animals, or biological material.
- **Consent to participate:** Not applicable.
- **Consent for publication:** Not applicable.
- **Data availability:** The publicly available datasets used in this study are the Buildings instance segmentation dataset and the VPAIR dataset. Details of these datasets, including repository/location information and identifiers where available, are provided in the cited dataset references. The GCWA-generated weather-domain images, trained model checkpoints, and experiment logs generated during the current study are available from the corresponding author upon reasonable request.
- **Materials availability:** Not applicable. This study utilized commercially available software and open-source datasets. No new physical materials or proprietary hardware were developed.
- **Code availability:** The scripts for GCWA generation, verification, YOLO11 training, and VPR evaluation are available from the corresponding author upon reasonable request.
- **Author contributions:** All authors contributed to the study conception and design. Volodymyr Vozniak and Oleksander Barmak conceived and conducted the experiments, Volodymyr Vozniak, Oleksander Barmak, and Pavlo Radiuk validated

the results, Volodymyr Vozniak, Oleksander Barmak, and Iurii Krak made formal analysis, Volodymyr Vozniak did investigation, Pavlo Radiuk and Iurii Krak conducted data curation, Volodymyr Vozniak and Oleksander Barmak prepared the draft of the manuscript, Pavlo Radiuk and Iurii Krak reviewed and edited the manuscript, Volodymyr Vozniak and Pavlo Radiuk made visualizations, Iurii Krak and Oleksander Barmak provided the supervision. All authors read and approved the final version of the manuscript.

References

[1] Schubert, S., Neubert, P., Garg, S., Milford, M., Fischer, T.: Visual place recognition: A tutorial. IEEE Robotics and Automation Magazine **31**(3), 139–153 (2024)

[2] Ostrovskyi, Z., Barmak, O., Radiuk, P., Krak, I.: Unsupervised knowledge extraction of distinctive landmarks from earth imagery using deep feature outliers for robust UAV geo-localization. Machine Learning and Knowledge Extraction **7**(3), 81 (2025)

[3] Vozniak, V., Barmak, O., Krak, I.: Method for matching satellite and UAV images for visual place recognition with cross-view color normalization. Radioelectronic and Computer Systems **2025**(3), 150–165 (2025)

[4] Zhang, X., Wang, L., Su, Y.: Visual place recognition: A survey from deep learning perspective. Pattern Recognition **113**, 107760 (2021)

[5] Borovyk, D., Barmak, O., Radiuk, P., Krak, I.: Hierarchical deep learning model for identifying similar targets in UAV imagery. Drones **9**(11), 743 (2025)

[6] Fan, J., Zheng, E., He, Y., Yang, J.: A cross-view geo-localization algorithm using UAV image and satellite image. Sensors **24**(12), 3719 (2024)

[7] Schmidt, F., *et al.*: ROVER: A multiseason dataset for visual SLAM. IEEE Transactions on Robotics **41**, 4005–4022 (2025)

[8] Yin, P., Cisneros, I., Zhao, S., Zhang, J., Choset, H., Scherer, S.: iSimLoc: Visual global localization for previously unseen environments with simulated images. IEEE Transactions on Robotics **39**(3), 1893–1909 (2023)

[9] Yin, P., *et al.*: General place recognition survey: Toward real-world autonomy. IEEE Transactions on Robotics **41**, 3019–3038 (2025)

[10] Pivoňka, T., Přeučil, L.: On model-free re-ranking for visual place recognition with deep learned local features. IEEE Transactions on Intelligent Vehicles **9**(12), 7900–7911 (2024)

[11] Svystun, S., Melnychenko, O., Radiuk, P., Savenko, O., Sachenko, A., Lysyi, A.:

Thermal and RGB images work better together in wind turbine damage detection. International Journal of Computing **23**(4), 526–535 (2024)

[12] Zaffar, M., *et al.*: VPR-Bench: An open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change. International Journal of Computer Vision **129**(7), 2136–2174 (2021)

[13] Rahat, F., Hossain, M.S., Ahmed, M.R., Jha, S.K., Ewetz, R.: Data augmentation for image classification using generative AI. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2025), pp. 4173–4182. IEEE, Tucson, AZ, USA (2025). https://doi.org/10.1109/WACV61041.2025.00410

[14] Yang, H., Carballo, A., Zhang, Y., Takeda, K.: Framework for generation and removal of multiple types of adverse weather from driving scene images. Sensors **23**(3), 1548 (2023)

[15] Shen, F., *et al.*: W-ControlUDA: Weather-controllable diffusion-assisted unsupervised domain adaptation for semantic segmentation. IEEE Robotics and Automation Letters **10**(5), 4204–4211 (2025)

[16] Seunghyun, B., Kim, S., Kim, E.: WeatherFLUX: Universal weather translation with diffusion models (2025). https://openreview.net/pdf?id=KYLWylUQsu

[17] Jung, K., Seo, Y., Cho, S., Kim, J., Min, H., Choi, S.: DALDA: Data augmentation leveraging diffusion model and LLM with adaptive guidance scaling. In: Bue, A.D., Canton, C., Pont-Tuset, J., Tommasi, T. (eds.) Computer Vision – ECCV 2024 Workshops, pp. 182–200. Springer, ??? (2025). https://doi.org/10.1007/978-3-031-91907-7_11

[18] Li, B., *et al.*: Semantic-guided generative image augmentation method with diffusion models for image classification. Proceedings of the AAAI Conference on Artificial Intelligence **38**(4), 3018–3027 (2024)

[19] López-Francos, I.G., Lipkis, R., Vlastos, P.G., Agogino, A.: Improving sim-to-real transfer in vision-based robot navigation via instance-level GAN-based data augmentation. American Institute of Aeronautics and Astronautics, ??? (2025). https://doi.org/10.2514/6.2025-1343

[20] Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. Journal of Big Data **6**, 60 (2019)

[21] Qiao, X., Fan, P., Li, J., Zhang, Y., Liu, X.: AWDAN: Adverse-weather data augmentation network for enhanced traffic object detection. In: Sixth International Conference on Signal Processing and Computer Science (SPCS 2025), pp. 363–368. SPIE, Xi'an, China (2025). https://doi.org/10.1117/12.3089556

[22] Mao, L., Soussan, R., Coltin, B., Smith, T., Biswas, J.: Semantic masking and

visual feature matching for robust localization. In: 2024 International Conference on Space Robotics (iSpaRo 2024), pp. 1–7. IEEE, Luxembourg, Luxembourg (2024). https://doi.org/10.1109/iSpaRo60631.2024.10688001

[23] Chen, K., *et al.*: Semantic visual simultaneous localization and mapping: A survey. IEEE Transactions on Intelligent Transportation Systems **26**(6), 7426–7449 (2025)

[24] Kang, M., Chikontwe, P., Won, D., Luna, M., Park, S.H.: Structure-preserving image translation for multi-source medical image domain adaptation. Pattern Recognition **144**, 109840 (2023)

[25] Zhang, H., Lui, L.M.: A learning-based framework for topology-preserving segmentation using quasiconformal mappings. Neurocomputing **600**, 128124 (2024)

[26] Caron, M., Houlsby, N., Schmid, C.: Location-aware self-supervised transformers for semantic segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2024), pp. 117–127. IEEE, Waikoloa, HI, USA (2024). https://doi.org/10.1109/WACV57701.2024.00019

[27] Miao, J., *et al.*: A survey on monocular re-localization: From the perspective of scene map representation. IEEE Transactions on Intelligent Vehicles **10**(4), 2519–2550 (2025)

[28] Tian, Y., *et al.*: Discriminative and semantic feature selection for place recognition towards dynamic environments. Pattern Recognition Letters **153**, 75–82 (2022)

[29] Chen, C., *et al.*: MPSD-net: A novel multiple progressive self-distillation network for semantic segmentation in adverse weather. Neurocomputing **654**, 131286 (2025)

[30] Schleiss, M., Rouatbi, F., Cremers, D.: VPAIR - Aerial visual place recognition and localization in large-scale outdoor environments. arXiv preprint arXiv:2205.11567 (2022)

[31] Ultralytics: YOLO11 models documentation (2024). https://docs.ultralytics.com/models/yolo11/ Accessed 2026-04-28

[32] Berton, G., Masone, C., Caputo, B.: Rethinking visual geo-localization for large-scale applications. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022), pp. 4878–4888. IEEE Computer Society, New Orleans, LA, USA (2022). https://doi.org/10.1109/CVPR52688.2022.00483

[33] Ali-Bey, A., Chaib-Draa, B., Giguère, P.: MixVPR: Feature mixing for visual place recognition. In: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2023), pp. 2997–3006. IEEE, Waikoloa, HI, USA (2023). https:

//doi.org/10.1109/WACV56688.2023.00301

[34] Svystun, S., Radiuk, P., Melnychenko, O., Savenko, O., Sachenko, A.: YOLO ensemble for UAV-based multispectral defect detection in wind turbine components. In: 2025 IEEE 13th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), pp. 11–16. IEEE, ??? (2025). https://doi.org/10.1109/IDAACS68557.2025.11322062

[35] Google AI Studio: Gemini 2.5 flash image (Nano banana) (2026). https://aistudio.google.com/models/gemini-2-5-flash-image/ Accessed 2026-04-28

[36] Roboflow: Buildings instance segmentation dataset ¿ Overview (2025). https://universe.roboflow.com/roboflow-universe-projects/buildings-instance-segmentation/dataset/4 Accessed 2026-04-28

[37] Rainio, O., Teuho, J., Klén, R.: Evaluation metrics and statistical tests for machine learning. Scientific Reports **14**(1), 6086 (2024)