# Weather-Robust Cross-View Geo-Localization via Prototype-Based Semantic Part Discovery

**Chi-Nguyen Tran**[1,2,4,5*†] **Duy-Minh Dao-Sy**[1,2,4,5*†] **Trung-Kiet Huynh**[1,2,4,5*†]
**Phu-Hoa Pham**[1,2,4,5†] **Phu-Quy Nguyen-Lam**[1,2,4,5†] **Long Tran-Thanh**[3]

[1]Faculty of Information Technology, University of Science, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
[3]Department of Computer Science, University of Warwick, Coventry, United Kingdom
[4]CompactAI
[5]Realtime Robotics, Ho Chi Minh City, Vietnam

*Equal contribution †Work done during an internship at Realtime Robotics.

## Abstract

Low-altitude economy has been experiencing rapid growth in recent years, with significant contributions to the global economy. While common drone tasks such as delivery, inspection, and search-and-rescue typically use Global Navigation Satellite Systems (GNSS) to navigate, there is an increasing need for developing alternative solutions as GNSS signals can be easily jammed, spoofed, or unavailable over a prolonged operational time. As such, cross-view geo-localization (CVGL), which matches an oblique drone view to a geo-referenced satellite tile, has emerged as a potent alternative that lets an autonomous drone localize itself when GNSS fails. Despite strong recent progress, three limitations persist in current CVGL methods: 1) global-descriptor designs compress the patch grid into a single vector without separating what is shared across the view gap (layout) from what is not (texture); 2) altitude-related scale variation is implicitly retained in the learned embedding rather than treated as a nuisance to be marginalized out; and 3) multi-objective training relies on hand-tuned scalars over losses that live on incompatible gradient scales. To address these limitations, we propose SKYPART, a lightweight swappable head for patch-based vision transformers (ViTs) that institutes explicit part grouping over the patch grid. SKYPART has four components grounded in established theory: (i) learnable prototypes that compete for patch tokens via a single-pass cosine assignment; (ii) altitude-conditioned linear modulation applied only during training so that the retrieval embedding is altitude-free at inference; (iii) a graph-attention readout over active prototypes, and (iv) a Kendall uncertainty-weighted multi-objective loss that learns the loss-group weights end-to-end rather than tuning them by hand. At 26.95 M parameters and 22.14 GFLOPs, SKYPART is the smallest among the top-performing methods in our comparison and sets a new state of the art on SUES-200, University-1652, and DenseUAV datasets under a single-pass, no-re-ranking, no test-time augmentation (TTA) protocol. Furthermore, its accuracy gap to the strongest baseline widens under the ten-condition WeatherPrompt corruption benchmark. Code is available at `https://github.com/chisngyen/SkyPart`.

## 1 Introduction

Autonomous drones and autonomous driving are two of the fastest maturing AI application areas in the last five years. Both are moving from research demonstration to large-scale deployment, such as commercial drone delivery and inspection fleets and self-driving vehicle pilots in dozens of cities. Both domains share the same hard requirement: the agent must know where it is on a prior map, in real

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

time, without relying on a radio signal that an adversary or a dense urban canyon can take away. This requires efficient robust visual geo-localization against a previously captured overhead map, which has drawn increasing attention from the research community. In particular, an autonomous drone that cannot localize itself cannot complete a mission. Global Navigation Satellite System (GNSS) jamming and spoofing are now routine in contested airspace and dense urban canyons [Radoš et al., 2024, Couturier and Akhloufi, 2021], and the next generation of low-altitude services, such as last-mile delivery, precision agriculture, search-and-rescue swarms, and infrastructure inspection, is being designed on the premise that radio signals may be absent or adversarial. Cross-view geo-localization (CVGL), which matches an oblique drone image to a geo-referenced satellite tile, has become one of the most important techniques to address this challenge, as it has the potential to remain robust under altitude change, weather corruption, and tight onboard compute budgets, requirements that are typical of many modern-day visual navigation tasks.

The core difficulty of CVGL is that the visual statistics that modern recognition techniques rely on (e.g., texture, colour, local contrast) are precisely the statistics that change most frequently between views and under weather. This leads to three limitations that persist in the current CVGL literature: 1) global-descriptor designs [Wu et al., 2024, Deuser et al., 2023, Du et al., 2024] compress the patch grid into a single vector with no mechanism to separate what is shared across the view gap from what is not; 2) altitude-related scale variation is implicitly retained in the learned embedding rather than marginalized as a nuisance; and 3) multi-objective training relies on hand-tuned scalars over losses on incompatible gradient scales.

We argue that to address these issues, one needs to factor out texture and keep the spatial organization of recurring components such as roads, roofs, and vegetation patches. We hypothesize that layout cues are more stable than texture under viewpoint, lighting, and sensor changes, and build SKYPART around that assumption [Marr and Nishihara, 1978, Biederman, 1987]. Following the invariance vs. sufficiency trade-off of the information bottleneck [Achille and Soatto, 2018], we adopt the working assumption that a good retrieval embedding should approximate the minimal sufficient statistic of the geolocation signal after nuisance variables have been marginalized out. Given this, our idea is that CVGL should be formulated around explicit grouping rather than ever-larger global pooling, where self-supervised ViTs can be used to expose emergent object-centric structure at the patch level [Caron et al., 2021, Oquab et al., 2024], and object-centric learning to identify a small set of competing slots to carve that structure into discrete, reusable components [Locatello et al., 2020]. The layout of those components is what is shared across the drone/satellite gap.

This idea reformulates the aerial CVGL problem as a layout-centric factorization problem, from which four architectural choices follow directly: prototype-based part discovery, nuisance-aware training modulation, multi-objective training, and active prototypes identification. To implement this pipeline, we develop SKYPART, which consists of:

1. Single-pass cosine assignment module for prototype-based part discovery [Locatello et al., 2020, Luo et al., 2023, Sun et al., 2018];
2. Altitude-conditioned Feature-wise Linear Modulation (FiLM) [Perez et al., 2018, Dumoulin et al., 2018], which is applied only during training, so altitude is a nuisance to marginalize rather than an inference feature;
3. Kendall-weighted multi-objective training [Kendall et al., 2018] over four loss groups (alignment, part quality, altitude, distillation), with the group weights learned end-to-end rather than hand-tuned;
4. A graph-attention readout [Veličković et al., 2018, Kipf and Welling, 2017] over active prototypes, which carries the pairwise compatibility of discovered components alongside the global image representation derived from the classification token (CLS) of the ViT.

We also apply a lightweight swappable head for patch-based ViTs that operationalizes the reformulation and trains end-to-end with a grouped multi-objective loss called GEOPARTLOSS. As such, the inference pipeline of SKYPART is a single forward pass with cosine similarity, no re-ranking, no test-time augmentation, and no query expansion. Note that SKYPART ranks first on SUES-200, University-1652, and DenseUAV under our evaluation protocol (single-pass, no-re-ranking, no-TTA; see Sec. 4.2), and is the most parameter-efficient among the top-accuracy-tier methods in the comparison. Under the reported WeatherPrompt protocol [Wen et al., 2025], SKYPART shows the strongest mean robustness, with especially large gains under fog+snow and darkness; SUES-200 and University-1652 baselines are taken from their published configurations and SKYPART is evaluated

under a shared corruption pipeline, so resolution and training-budget differences (detailed in Sec. 4.4) are reported explicitly rather than fully resolution-controlled. Overall, SKYPART is among the most parameter- and compute-efficient methods in the top-accuracy tier (26.95 M, 22.14 GFLOPs), retains a strong relative robustness margin under the WeatherPrompt [Wen et al., 2025] corruption benchmark, and we further include a lessons-learned catalogue of the methods we tried that underperformed (Appendix A5.3).

We also apply SKYPART to the problem of zero-shot cross-dataset transfer (University-1652 → SUES-200). Here, SKYPART leads all Drone→Satellite cells and remains competitive on Satellite→Drone, where dedicated baselines retain advantages in some R@1/AP cells (see Appendix A3.5 for the full table).

## 2 Related Work

**Cross-View Geo-Localization.** CVGL began as a ground→aerial retrieval problem [Workman et al., 2015, Lin et al., 2015, Hu et al., 2018], solved by NetVLAD-style aggregators [Arandjelović et al., 2016] and triplet objectives [Schroff et al., 2015], and sharpened by geometry-aware priors such as polar warping [Shi et al., 2020]. The UAV benchmarks [Zheng et al., 2020, Zhu et al., 2023, Dai et al., 2024] removed most of that prior because a drone's oblique angle and continuous altitude variation replace a clean radial warp with a soft projective distortion. Recent UAV CVGL has responded on two fronts: a scale-up line stacking larger ConvNeXt-B backbones with hard-negative sampling and attention-based partitioning [Dai et al., 2022, Shen et al., 2024, Deuser et al., 2023, Chen et al., 2024, Du et al., 2024, Wu et al., 2024, Chen et al., 2025, Hou et al., 2025], and an efficiency line that shrinks the model at a cost in accuracy [Wang et al., 2022, Liu et al., 2024, Xu et al., 2025a, Ping et al., 2025, Gao et al., 2025, Sun et al., 2024]. Both lines retain a global-descriptor commitment, sometimes augmented by a fixed spatial partition; partial-overlap settings [Ye et al., 2026] expose the low-altitude limit of both. SKYPART departs from this commitment and replaces the global pool with an explicit prototype grouping that the retrieval objective can reward directly. Closest in spirit among graph-based efficiency methods, SeGCN [Liu et al., 2024] relies on semantic categories from an auxiliary module to define graph nodes, while SKYPART discovers its prototype nodes end-to-end from patch-token competition without auxiliary semantic supervision. Classical vision-aided navigation can instead align video frames to views rendered from a geo-registered 3D model using geometric correspondences, inertial cues, and temporal optimization [Chiu et al., 2014]. SKYPART ranks flat satellite tiles from a single image and may serve as a coarse front end for such a system.

**Part-based Representations and Object-Centric Grouping.** Part-based recognition has a long cognitive-vision history [Marr and Nishihara, 1978, Biederman, 1987, Sun et al., 2018], and its modern incarnation is slot attention [Locatello et al., 2020]: a small set of latents competes for input tokens through iterative attention, producing object-centric groupings without supervision. SegCLIP [Luo et al., 2023] projects this onto ViT patch tokens [Dosovitskiy et al., 2021, Vaswani et al., 2017] for open-vocabulary segmentation, and masked autoencoders and feature-masking variants [He et al., 2022, Zhou et al., 2022, Baevski et al., 2022] pursue analogous grouping through reconstruction. None was designed for cross-view retrieval, and the altitude nuisance specific to aerial imagery has no analogue in their benchmarks. SKYPART uses a single-pass cosine assignment (a fixed-point view of slot iteration that is sufficient when the downstream objective is retrieval rather than reconstruction), applies masking and distillation at the *part* level rather than the patch level so that reconstruction pressure targets layout rather than texture, and conditions the clustering pathway on altitude via FiLM [Perez et al., 2018, Dumoulin et al., 2018].

**Metric Learning and Multi-Task Optimization.** Our retrieval objective groups Circle [Sun et al., 2020], proxy-anchor [Kim et al., 2020], InfoNCE [van den Oord et al., 2018], and patch-level NCE into a single *alignment* task, and combines that alignment task with three further tasks (part quality, altitude regression, and distillation) via Kendall homoscedastic uncertainty weighting [Kendall et al., 2018], read through the alignment/uniformity decomposition of contrastive learning [Wang and Isola, 2020, Chen et al., 2020]: the decomposition predicts that alignment carries the signal and uniformity regularizes it, which the ablation in Sec. 4.5 confirms. The general theorem of Sener & Koltun [Sener and Koltun, 2018] guarantees that any positive-weighted scalarization of the four task losses is Pareto-stationary in the vector objective, and Kendall weighting produces such positive weights $\exp(-s_g)$ automatically; what it buys us over a hand-tuned scalar combination (which is also Pareto-stationary) is that the four weights are learned end-to-end rather than swept manually. The

relational readout over prototype nodes [Veličković et al., 2018, Kipf and Welling, 2017] complements the global CLS pathway; denoising [Feng et al., 2024] and domain-adversarial [Ganin and Lempitsky, 2015] alternatives are orthogonal. Distillation [Hinton et al., 2015, Gou et al., 2021, Tian et al., 2020, Ma et al., 2024, Park et al., 2019] is used only to stabilize the teacher-space projector during training, so the deployed pipeline has no external detector, diffusion front-end, or inference-time adaptation.

# 3 Method: SKYPART

## 3.1 Architecture Overview

We start with the shape of the head which follows from the framing we described in Sec. 1. If layout is the invariant we want the embedding to preserve across the drone/satellite gap, then the representation should expose an identity axis (what the recurring components are) and an arrangement axis (how they are spatially related), and it should do so without letting altitude, which is a nuisance observed at training but unreliable at test, leak into the retrieval embedding. Four choices fall out of this framing. First, the part branch is built as a competition between a small bank of prototypes for a larger set of patch tokens, with the softmax running over prototype indices rather than patch positions, so that each patch commits to a prototype rather than each prototype averaging over all patches. This is the single-pass analogue of slot attention [Locatello et al., 2020], which we treat as sufficient in our retrieval setting because the downstream objective does not require a reconstruction from the grouping. Second, a graph readout over the active prototype nodes can produce a separate arrangement embedding, complementing the identity-heavy part descriptor and the global CLS pathway, which refers to the branch that uses the CLS token as a global image descriptor [1]. Third, altitude conditions the clustering pathway through a feature-wise affine modulation [Perez et al., 2018, Dumoulin et al., 2018] but is marginalized at inference through mean FiLM parameters. This yields an exact equivalence only at the modulation-operator level (linearity in $\gamma, \beta$), not a claim that training-conditioned features are globally unchanged (formal statement in Appendix A1.2). Fourth, the four training heads (alignment, part quality, altitude, and distillation) live on incompatible gradient scales. To combine them we use Kendall homoscedastic uncertainty weighting [Kendall et al., 2018]. Sener & Koltun's general theorem [Sener and Koltun, 2018] guarantees Pareto-stationarity for any positive-weighted scalarization of the four task losses; Kendall weighting produces such positive weights $\exp(-s_g)$ automatically, so its advantage over a hand-tuned scalar combination (which is also Pareto-stationary) is that the four weights are learned end-to-end rather than swept manually. The theoretical justification of this idea can be found in Appendix A1.2.

Figure 1 depicts the resulting full pipeline. A shared patch-based encoder, DINOv2 ViT-S/14 in our experiments, chosen because its self-distillation pretext produces patch tokens with emergent object-centric structure [Caron et al., 2021, Oquab et al., 2024] that the prototype assignment can cluster on, processes both views to produce a CLS token and a set of patch tokens. The three branches read from these tokens: a global branch that keeps the CLS, a part branch that runs prototype competition and returns a part-pooled descriptor, and a graph branch that runs a two-layer Graph Attention Network (GAT) [Veličković et al., 2018] over per-view active prototype nodes (fully connected within each view; no explicit drone-satellite edges in the graph). A lightweight fusion gate combines the three into a single 768-D retrieval vector. Altitude-FiLM acts only as a metadata-conditioned branch during training; at inference it uses fixed mean FiLM parameters and requires no altitude input. The auxiliary signals (masked part reconstruction, teacher distillation, EMA self-ensembling) are discarded at inference. The deployed model is the student backbone plus the head, with cosine similarity and no re-ranking or test-time augmentation.

## 3.2 Semantic Part Discovery and Three-Way Fusion

**Prototype bank and assignment.** Our representation is organized around a small bank of $K_{\max}{=}12$ learnable prototypes $\{\mathbf{p}_k\}_{k=1}^{K_{\max}}$, each acting as a soft detector for a recurring aerial region (buildings, vegetation, roads, water); the per-image salience gate (next paragraph) selects an active subset of typical size $K{=}10$, so $K_{\max}$ is the bank capacity and $K$ the typical active count throughout the paper. These prototypes are latent functional parts rather than supervised semantic categories; the region names are illustrative. Each projected, altitude-modulated patch token $\mathbf{z}_i$ commits to one prototype

[1] The CLS (classification) token is a learnable, special vector added to the input sequence of a Vision Transformer (ViT) to represent the entire image for classification tasks [Lappe and Giese, 2025].

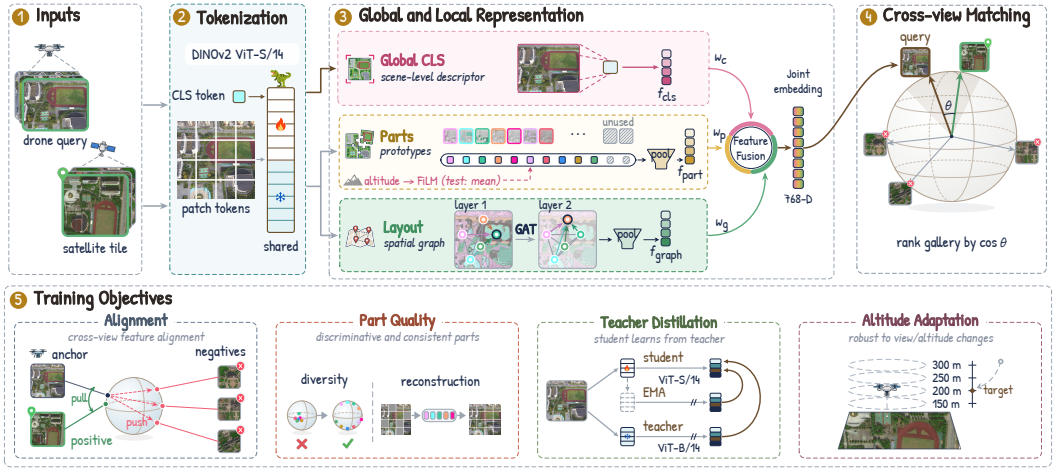


Figure 1: **SkyPart overview. (1–2)** A shared DINOv2 ViT-S/14 tokenizes the drone query and the satellite tile into a CLS token and patch tokens. **(3)** The head reads the tokens in three ways: the global CLS token; a part branch in which patch tokens are assigned by cosine similarity to $K_{\max}=12$ learnable prototypes and pooled into $\mathbf{f}_{\text{part}}$; and a two-layer GAT over the active prototypes that encodes their spatial layout. Altitude conditions the part branch through FiLM during training only; at test time FiLM uses its mean parameters and needs no altitude input. A fusion gate with weights $w_c, w_p, w_g$ combines the three readouts into one $\ell_2$-normalized 768-D embedding. **(4)** Retrieval is a single cosine ranking of the satellite gallery. **(5)** The four Kendall-weighted loss groups (alignment, part quality, teacher distillation, altitude) are used only during training.

through a single-pass temperature-scaled cosine assignment

$$\mathbf{A}_{i,k} = \frac{\exp(\langle \bar{\mathbf{p}}_k, \bar{\mathbf{z}}_i \rangle / \tau)}{\sum_j \exp(\langle \bar{\mathbf{p}}_j, \bar{\mathbf{z}}_i \rangle / \tau)}, \tag{1}$$

where $\bar{\mathbf{p}}_k$ and $\bar{\mathbf{z}}_i$ denote the $\ell_2$-normalized prototype and patch token (so $\langle \bar{\mathbf{p}}_k, \bar{\mathbf{z}}_i \rangle$ is a cosine similarity); the softmax runs *over prototype indices* rather than over patch positions, which is what turns the bank into a competing set of slots [Locatello et al., 2020] rather than a linear reweighting. Prototype count, dimensionality, and temperature are given in Appendix A2.1.

**Per-image salience gating.** Because different scenes activate different subsets of the bank (an urban tile rarely needs a "water" prototype), we learn a per-image Gumbel-sigmoid salience gate over the prototype indices that zeroes out inactive prototypes at the image level while keeping a small active set. The gate, the residual MLP that refines each active part, and the spatial-centroid bookkeeping are detailed in Appendix A1.3.

**Three-way fusion.** The retrieval embedding combines three readouts. A global branch keeps the CLS token ($\mathbf{f}_{\text{cls}}$), a part branch pools the salience-weighted part descriptors into $\mathbf{f}_{\text{part}}$, and a graph branch passes the active prototype nodes through a two-layer GAT [Veličković et al., 2018] over a fully-connected adjacency (feasible for this small prototype bank) to yield a layout readout $\mathbf{f}_{\text{graph}}$. The part and CLS pathways carry identity-heavy information (what the components are), and the graph readout carries arrangement-heavy information (how they are related), consistent with the layout-invariance framing of Sec. 1 and formalized as an information-bottleneck factorization in Appendix A1.2. A lightweight gating MLP learns non-negative weights summing to one, with

$$[w_p, w_c, w_g] = \text{softmax}\big(\mathbf{W}_2\, \text{ReLU}(\mathbf{W}_1[\mathbf{f}_{\text{part}}; \mathbf{f}_{\text{cls}}; \mathbf{f}_{\text{graph}}])\big), \tag{2}$$

initialized to start biased toward the part branch and re-balance during training. The output is the L2-normalized sum $\mathbf{f} = \text{L2Norm}(w_p \mathbf{f}_{\text{part}} + w_c \mathbf{f}_{\text{cls}} + w_g \mathbf{f}_{\text{graph}})$.

### 3.3 Altitude as a Training-Time Nuisance (FiLM)

Altitude changes scale and perspective but is not part of the location identity we want the retrieval embedding to encode, so we treat it as a nuisance to marginalize rather than a feature to keep. During




Figure 2: **Part-level evidence under weather shifts.** Rows show clean drone inputs, their part-level activations, paired satellite views, satellite part activations, weather-corrupted drone queries, and the corresponding part activations. Columns cover different corruptions and mixed weather conditions. Across substantial appearance changes, the part-discovery head continues to produce spatially structured activations over buildings, roads, fields, and open regions, rather than collapsing to a single dominant texture cue. These visualizations support prototype-based layout as a robust cross-view abstraction learned from the retrieval objective alone, without segmentation supervision.

training, FiLM conditions projected patch tokens on the available altitude bin before prototype assignment. At inference, no altitude input is used: altitude-specific FiLM parameters are replaced by their marginalized mean and this fixed modulation is applied to both views. Thus the deployed embedding has no dependence on test-time altitude metadata, while still benefiting from altitude-aware regularization during training. A paired three-seed control with the recipe, checkpoint rule, and evaluator fixed shows that active altitude-FiLM adds only $+0.14$ pp clean D$\rightarrow$S R@1 but $+0.62$ pp on the ten-condition mean (Appendix A4.2), so we treat FiLM as a robustness regularizer rather than a main accuracy driver. The formal mean-FiLM argument is given in Appendix A1.2. Masked Part Reconstruction and distillation further shape the part pathway during training and are detailed in Appendix A1.3.

### 3.4 Training Objective: GeoPartLoss

**GeoPartLoss**. We use a multi-objective training loss that jointly optimizes four complementary objectives via Kendall homoscedastic uncertainty weighting [Kendall et al., 2018]: $\mathcal{L}_{\text{GeoPartLoss}} = \sum_{g\in\mathcal{G}}\left[\exp(-s_g)\,\mathcal{L}_g + s_g\right]$, where $\mathcal{G} = \{\text{align, part, distill, alt}\}$ are the four loss groups, and $s_g = \log\sigma_g^2$ is a per-group learnable log-variance. The weight $\exp(-s_g)$ is always positive, so a larger $s_g$ down-weights that group; the additive $s_g$ term penalizes collapse to zero. By construction, any stationary point of this loss is a Pareto-stationary point of the four-objective vector $(\mathcal{L}_{\text{align}}, \mathcal{L}_{\text{part}}, \mathcal{L}_{\text{distill}}, \mathcal{L}_{\text{alt}})$ [Sener and Koltun, 2018], removing the need for hand-tuned loss scalars.

**Four loss groups.** *(i) Alignment* ($\mathcal{L}_{\text{align}}$): Circle loss [Sun et al., 2020], proxy-anchor [Kim et al., 2020], and InfoNCE [van den Oord et al., 2018] applied to the fused 768-D embedding, pulling same-location drone–satellite pairs together and pushing different-location pairs apart, plus a patch-level NCE term on part-assigned patches (Appendix A1.5). *(ii) Part quality* ($\mathcal{L}_{\text{part}}$): masked part reconstruction (MAR) and prototype diversity, encouraging prototypes to cover distinct and spatially coherent scene regions. *(iii) Distillation* ($\mathcal{L}_{\text{distill}}$): feature-level alignment toward a frozen DINOv2 ViT-B/14 teacher and an exponential moving average (EMA) self-ensemble [Tarvainen and Valpola, 2017], providing a stable training signal for the smaller student backbone. *(iv) Altitude* ($\mathcal{L}_{\text{alt}}$): SmoothL1 regression on normalized altitude, active only when metadata is available and omitted otherwise. More details of this loss can be found in Appendix A1.5.

**Uncertainty-Aware Proxy Alignment (UAPA).** Within the alignment group, we modulate the distillation temperature using the relative entropy gap between the drone and satellite proxy distributions. More uncertain drone queries receive softer teacher targets while confident queries retain the default temperature, with gradient magnitude preserved. Full formulation and ablation are in Appendix A1.4.

**Inference.** SKYPART reduces at test time to a single forward pass through the backbone and head. The teacher, EMA self-ensemble, altitude-supervision branch, and all GEOPARTLOSS terms are discarded; altitude-specific FiLM is replaced by fixed mean-FiLM parameters requiring no altitude metadata. The deployed model is the 26.95 M student and its 768-D embedding. Retrieval uses cosine similarity with no re-ranking, TTA, query expansion, multi-crop inference, or post-processing.

# 4 Experiments

## 4.1 Datasets and Protocol

We evaluate SKYPART on SUES-200 [Zhu et al., 2023], University-1652 [Zheng et al., 2020], and DenseUAV [Dai et al., 2024] under a strict single-pass protocol: a single-scale $448\times448$ crop, the full confusion gallery when defined by the benchmark, and no test-time augmentation, re-ranking, query expansion, multi-crop inference, or other post-processing. Published baselines in the SOTA tables are reported from their original papers and may follow method-specific protocols, as detailed in Sec. 4.3. Altitude is treated as an optional training-time nuisance signal, not a required input feature. We use FiLM bins when metadata is available (150/200/250/300 m on SUES-200; dataset-provided altitude bins on DenseUAV), and mean-FiLM otherwise, omitting the altitude-supervision term on datasets without altitude labels. Thus the same head supports datasets with and without altitude annotations without pseudo-labeling or folder-derived heuristics. For weather robustness we adopt the WeatherPrompt protocol [Wen et al., 2025] verbatim. Split sizes, per-method weather fine-tuning recipes, and dataset quirks appear in Appendix A2.2 and A3.4. All results use official splits and full benchmark galleries. WeatherPrompt corrupts drone queries only.

## 4.2 Implementation Details

The deployed model is a DINOv2 ViT-S/14 student with the SKYPART head, totaling 26.95 M parameters and 22.14 GFLOPs for a single $448\times448$ forward pass. Training uses AdamW, identity-balanced mini-batches, standard crop/colour augmentation, and the GEOPARTLOSS groups described in Sec. 3.4. DenseUAV additionally uses rotation augmentation during training because its satellite crops have arbitrary basemap orientation; SUES-200 and University-1652 use canonical orientations. At inference, the teacher, EMA, FiLM-altitude auxiliaries, and all training objectives are discarded; retrieval is cosine similarity between 768-D embeddings, with no re-ranking, TTA, query expansion, multi-crop inference, or post-processing. Results for SKYPART follow this deployed protocol. Across three independently trained seeds (42, 123, 2024; final-epoch checkpoints, no test-set selection), clean D$\rightarrow$S R@1 is $98.74 \pm 0.11$ on SUES-200, $96.47 \pm 0.25$ on University-1652, and $91.85 \pm 0.26$ on DenseUAV (mean $\pm$ sample SD). Full hyperparameters and dataset-specific training details are in Appendix A2.1 and Appendix A2.2.

## 4.3 Comparison with State of the Art

**SUES-200.** Table 1 reports the SUES-200 comparison. Under the single-pass protocol (a single $448\times448$ crop, one forward pass, the full 200-tile confusion gallery, and *no* re-ranking, TTA, query expansion, or multi-crop inference, while baselines retain whatever inference-time wrappers their original papers use), SKYPART sets a new state of the art across every altitude band on D$\rightarrow$S with the smallest model in the table. The margin grows at the lowest altitude, where the field of view is narrow and only partial facades are visible: global descriptors lose traction precisely in the regime where part-level layout becomes the deciding cue, exactly as the framing in Sec. 1 predicts.

**University-1652 and DenseUAV.** Table 2 extends the same picture to the other two benchmarks. On University-1652, SKYPART leads R@1 and AP at the lowest parameter budget in the comparison while running the inference path with mean-FiLM (no altitude metadata), evidence that the architecture itself, rather than altitude conditioning, carries the advantage. DenseUAV is the harder benchmark: its dense urban gallery makes neighbouring locations visually ambiguous and three flight heights stress the altitude module-yet SKYPART still has the highest reported R@1 and the highest reported SDM@1; some high-R@1 baselines do not report SDM@1.

Table 1: **SUES-200 per-altitude comparison.** R@1/AP (%) at four UAV altitudes for D→S and S→D, with mean R@1 per direction in the rightmost column of each block. SKYPART is evaluated with a single-scale 448×448 crop, the full 200-satellite confusion gallery, and no post-processing; the SKYPART row is one checkpoint (three-seed mean D→S R@1: 98.74 ± 0.11, Sec. 4.2); non-SKYPART numbers are reported from the original publications. Cell shading scales with R@1/AP (green = high, red = low). **Red**: best; **blue**: second; SKYPART has the lowest Params/GFLOPs.

| Method | Venue | Params↓ | GFLOPs↓ | Drone→Satellite | | | | | | | | | Satellite→Drone | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 150m | | 200m | | 250m | | 300m | | Mean | 150m | | 200m | | 250m | | 300m | | Mean |
| | | | | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 |
| LPN [Wang et al., 2022] | TCSVT'22 | 62.4 M | 36.78 | 61.58 | 67.23 | 70.85 | 75.96 | 80.38 | 83.80 | 81.47 | 84.53 | 73.57 | 83.75 | 66.78 | 88.75 | 75.01 | 92.50 | 81.34 | 92.50 | 85.72 | 89.38 |
| Baseline [Zhu et al., 2023] | TCSVT'23 | – | – | 55.65 | 61.92 | 66.78 | 71.55 | 72.00 | 76.43 | 74.05 | 78.26 | 67.12 | 75.00 | 55.46 | 85.00 | 66.05 | 86.25 | 69.94 | 88.75 | 74.46 | 83.75 |
| Sample4Geo [Deuser et al., 2023] | ICCV'23 | 87.6 M | 90.24 | 92.60 | 94.00 | 97.38 | 97.81 | 98.28 | 98.64 | 99.18 | 99.36 | 96.86 | 92.50 | 94.00 | 97.38 | **97.81** | 98.28 | **98.64** | **99.18** | **99.36** | 96.84 |
| MCCG [Shen et al., 2024] | TCSVT'24 | 56.6 M | 51.04 | 82.22 | 85.47 | 89.38 | 91.41 | 93.82 | 95.04 | 95.07 | 96.20 | 90.12 | 93.75 | 89.72 | 93.75 | 92.21 | 96.25 | 96.14 | 98.75 | 96.64 | 95.62 |
| SeGCN [Liu et al., 2024] | JSTARS'24 | **28.0 M** | **24.58** | 90.80 | 92.32 | 91.93 | 93.41 | 92.53 | 93.90 | 93.33 | 94.61 | 92.15 | 93.75 | 92.45 | 95.00 | 93.65 | 96.25 | 94.39 | 97.50 | 94.55 | 95.62 |
| CCR [Du et al., 2024] | TCSVT'24 | 156.6 M | 160.6 | 87.08 | 89.55 | 93.57 | 94.90 | 95.42 | 96.28 | 96.82 | 97.39 | 93.22 | 92.50 | 88.54 | 97.50 | 95.22 | 97.50 | 97.10 | 97.50 | 97.49 | 96.25 |
| CAMP [Wu et al., 2024] | TGRS'24 | 91.4 M | 90.24 | 95.40 | 96.38 | 97.63 | 98.16 | 98.05 | 98.45 | 99.33 | 99.46 | 97.60 | 96.25 | 93.69 | 97.50 | 96.76 | **98.75** | 98.10 | **100.00** | 98.85 | 98.12 |
| DAC [Xia et al., 2024] | TCSVT'24 | 96.5 M | 90.24 | **96.80** | **97.54** | 97.48 | 97.97 | 98.20 | 98.62 | 97.58 | 98.14 | 97.52 | **97.50** | 94.06 | **98.75** | 96.66 | **98.75** | 98.09 | 98.75 | 97.87 | 98.44 |
| MEAN [Chen et al., 2025] | TGRS'25 | 36.5 M | 26.18 | 95.50 | 96.46 | **98.38** | **98.72** | **98.95** | **99.17** | **99.52** | **99.63** | **98.09** | **97.50** | **94.75** | **100.00** | 97.09 | **100.00** | 98.28 | **100.00** | 99.21 | **99.38** |
| **SKYPART (Ours)** | – | **26.95 M** | **22.14** | **97.25** | **97.90** | **98.75** | **99.33** | **99.30** | **99.64** | **99.60** | **99.77** | **98.725** | **100.00** | **97.36** | **100.00** | **97.95** | **100.00** | **99.98** | **100.00** | **99.95** | **100.00** |

Table 2: **University-1652 (left) and DenseUAV (right).** Side-by-side state-of-the-art comparison. SKYPART leads both benchmarks while remaining among the smallest models in the comparison. Same cell shading as Table 1 (green = high R@1/AP). **Red**: best; **blue**: second. †DenseUAV: 256×256 input, as reported in the comparison of [Xu et al., 2025c]; *param count estimated from backbone (ViT-S≈22 M, ConvNeXt-T≈28 M, ConvNeXt-B≈88 M, DINOv2-B≈86 M).

**(a) University-1652**

| Method | Venue | Params | D→S | | S→D | |
|---|---|---|---|---|---|---|
| | | | R@1 | AP | R@1 | AP |
| LPN [Wang et al., 2022] | TCSVT'22 | 62.4 M | 77.71 | 80.80 | 90.30 | 78.78 |
| Sample4Geo [Deuser et al., 2023] | ICCV'23 | 87.6 M | 92.65 | 93.81 | 95.14 | 91.39 |
| MCCG [Shen et al., 2024] | TCSVT'24 | 56.6 M | 89.28 | 91.01 | 94.29 | 89.29 |
| CCR [Du et al., 2024] | TCSVT'24 | 156.6 M | 92.54 | 93.78 | 95.15 | 91.80 |
| CAMP [Wu et al., 2024] | TGRS'24 | 91.4 M | **94.46** | **95.38** | **96.15** | **92.72** |
| MEAN [Chen et al., 2025] | TGRS'25 | **36.5 M** | 93.55 | 94.53 | 96.01 | 92.08 |
| **SKYPART (Ours)** | – | **26.95 M** | **96.47** | **97.77** | **98.43** | **98.24** |

**(b) DenseUAV**

| Method | Venue | Params | R@1 | R@5 | SDM@1 |
|---|---|---|---|---|---|
| DenseUAV baseline† [Dai et al., 2024] | TIP'24 | ∼22 M* | 83.01 | 95.58 | 86.50 |
| Sample4Geo† [Deuser et al., 2023] | ICCV'23 | 87.6 M | 49.38 | 78.29 | 61.72 |
| MCCG† [Shen et al., 2024] | TCSVT'24 | ∼28 M* | 83.14 | 93.39 | 85.94 |
| CAMP [Wu et al., 2024] | TGRS'24 | 91.4 M | 88.72 | – | – |
| Yang et al.† [Yang et al., 2025] | RA-L'25 | ∼86 M* | 86.27 | **96.83** | 88.87 |
| CEUSP† [Xu et al., 2025c] | PRCV'25 | ∼28 M* | 89.45 | 96.05 | **91.01** |
| MEAN [Chen et al., 2025] | TGRS'25 | 36.5 M | **90.18** | – | – |
| **SKYPART (Ours)** | – | 26.95 M | **91.85** | **97.81** | **93.59** |

**Protocol note for Tables 1 and 2.** SKYPART is evaluated under the strict single-pass protocol of Sec. 4.1; the baseline numbers are taken directly from the original publications, where some methods use inference-time wrappers such as $k$-reciprocal re-ranking, multi-scale TTA, or query expansion. We do not apply these wrappers to the headline SKYPART results; when combined on SKYPART, they increase D→S R@1 by +0.66, +1.25, and +1.83 pp on SUES-200, University-1652, and DenseUAV, respectively (three seeds; Appendix A4.3). Conversely, SKYPART uses a 448×448 input, larger than the 384×384 or 256×256 inputs of several baselines. Tables 1 and 2 should therefore be read as comparisons of published operating points rather than fully protocol-normalized results.

### 4.4 Weather Robustness (WeatherPrompt)

Clean-data accuracy is a weak proxy for a drone flying through fog at dusk. WeatherPrompt [Wen et al., 2025] stress-tests this setting by applying nine weather and visibility corruptions to the drone query (ten conditions with the clean Normal case) while keeping the satellite gallery clean. Since the deployment-critical direction is drone-query localization, Table 3 reports D→S R@1 on SUES-200, University-1652, and DenseUAV; full per-condition R@1/AP results for both D→S and S→D are deferred to Appendix A3.4.

On SUES-200 and University-1652, baselines are reported from WeatherPrompt at 384×384, while SKYPART is retrained at 392×392 (yielding a 28×28 patch grid for DINOv2 ViT-S/14) under the same corruption pipeline, so these two blocks should be read as reported-baseline rather than fully resolution-controlled comparisons. Two controls bound the resolution difference (Appendix A4.4): re-running WeatherPrompt at 392×392 with its own recipe changes the ten-condition mean by only +0.18, +0.10, and +0.12 pp on SUES-200, University-1652, and DenseUAV; retraining SKYPART at 378×378, below WeatherPrompt's 384-px input, lowers its own mean by 0.51, 0.61, and 1.63 pp, respectively. On DenseUAV, we retrain every method under the same 60-epoch budget, batch size, online corruption seeds, and official gallery with the WeatherPrompt pipeline, while retaining method-specific input resolutions and learning rates; SKYPART also ranks first on this denser urban robustness benchmark. Across all datasets, SKYPART preserves the highest mean robustness and degrades most

Table 3: **Weather robustness summary.** Drone→Satellite R@1 (%) under the 10-condition WeatherPrompt corruption pipeline with clean satellite galleries. SUES-200 and University-1652 baselines are reported from WeatherPrompt [Wen et al., 2025] at 384×384; SKYPART is retrained at 392×392. DenseUAV uses the same corruption pipeline with reproduced baselines, yielding a third weather-robustness benchmark. Full results for D→S and S→D are in Appendix A3.4. *: official pretrained weights; †: official weights pretrained on University-1652 (markers as in WeatherPrompt). **Red**: best; **blue**: second.

| Method | Venue | Params↓ | Normal | Fog | Rain | Snow | F+R | F+S | R+S | Dark | Over-exp | Wind | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **SUES-200** | | | | | | | | | | | | | |
| Sample4Geo† [Deuser et al., 2023] | ICCV'23 | 87.6 M | 74.93 | 72.58 | 34.60 | 28.95 | 35.10 | 12.95 | 20.05 | 34.18 | 38.40 | **67.80** | 41.95 |
| Safe-Net* [Lin et al., 2025] | TIP'25 | 25.6 M | 76.31 | **73.53** | 54.15 | 48.94 | 45.12 | 40.05 | 25.95 | 29.74 | 54.86 | 58.10 | 50.68 |
| CCR† [Du et al., 2024] | TCSVT'24 | 156.6 M | 73.22 | 70.95 | 60.14 | 50.31 | 45.87 | 45.80 | 31.25 | 31.03 | 59.97 | 52.02 | 52.06 |
| WeatherPrompt [Wen et al., 2025] | NeurIPS'25 | 87.6 M | **76.72** | 68.49 | **71.77** | **59.95** | **58.24** | **64.36** | **58.49** | **40.42** | **61.57** | 65.19 | **62.52** |
| **SKYPART (Ours)** | – | 26.95 M | **97.21** | **96.27** | **96.06** | **95.33** | **94.71** | **90.44** | **95.17** | **92.06** | **93.34** | **96.54** | **94.71** |
| **University-1652** | | | | | | | | | | | | | |
| Sample4Geo* [Deuser et al., 2023] | ICCV'23 | 87.6 M | **92.70** | **88.70** | 62.44 | 52.76 | 52.70 | 19.79 | 38.19 | 46.34 | 75.77 | **81.54** | 61.10 |
| Safe-Net* [Lin et al., 2025] | TIP'25 | 25.6 M | 86.98 | 82.12 | 67.13 | 60.50 | 54.80 | 32.12 | 25.83 | 41.10 | 69.87 | 74.32 | 59.48 |
| CCR* [Du et al., 2024] | TCSVT'24 | 156.6 M | 92.54 | 85.57 | 67.46 | 55.16 | 63.11 | 27.74 | 23.06 | 51.10 | **75.90** | 81.31 | 62.30 |
| WeatherPrompt [Wen et al., 2025] | NeurIPS'25 | 87.6 M | 82.78 | 81.46 | **80.34** | **77.60** | **78.75** | **73.38** | **78.41** | **67.22** | 74.20 | 77.26 | **77.14** |
| **SKYPART (Ours)** | – | 26.95 M | **95.15** | **93.78** | **93.44** | **92.05** | **90.05** | **88.51** | **89.47** | **83.26** | **88.23** | **89.68** | **90.36** |
| **DenseUAV** | | | | | | | | | | | | | |
| Sample4Geo [Deuser et al., 2023] | ICCV'23 | 87.6 M | **72.37** | **72.42** | **73.66** | **72.80** | **73.14** | **70.44** | **72.59** | **66.50** | **69.76** | **72.29** | **71.60** |
| Safe-Net [Lin et al., 2025] | TIP'25 | 25.6 M | 41.14 | 39.47 | 40.11 | 37.11 | 37.54 | 31.87 | 39.30 | 33.20 | 37.07 | 40.03 | 37.68 |
| CCR [Du et al., 2024] | TCSVT'24 | 156.6 M | 56.76 | 55.21 | 55.94 | 54.70 | 54.10 | 51.48 | 55.30 | 49.85 | 54.44 | 54.87 | 54.26 |
| WeatherPrompt [Wen et al., 2025] | NeurIPS'25 | 87.6 M | 26.25 | 24.15 | 26.68 | 24.28 | 25.35 | 20.72 | 25.10 | 20.08 | 22.48 | 25.78 | 24.09 |
| **SKYPART (Ours)** | – | 26.95 M | **91.25** | **90.13** | **91.38** | **90.35** | **89.79** | **87.69** | **91.12** | **84.64** | **90.18** | **91.76** | **89.83** |

under low-visibility conditions (fog+snow, darkness, and over-exposure), showing evidence that layout survives texture corruption better than global descriptors.

### 4.5 Ablation Study

Table 4: **Weather-aligned ablation on SUES-200** (Drone→Satellite). †Full (Ours) is the single reference row; all ΔMean values in blocks A–C are computed against it. Block D is a paired run and its Δ is within-block only. Blocks A–C share one ablation checkpoint (Normal 98.69, Mean 96.03), trained separately from the deployed checkpoint of Table 3 (Normal 97.21, Mean 94.71); the 1.5 pp normal gap reflects the modest clean-accuracy cost of weather-online training. **Red** ◀: catastrophic drop (>10 pp); **blue** ◁: large drop (2–10 pp); gray: minor (<2 pp). Training follows the matched weather-online recipe. Deltas in the text are computed against the corresponding full-model row. Weather-run implementation details are in Appendix A3.4.

| Configuration | Normal | | Fog | | Rain | | Snow | | F+R | | F+S | | R+S | | Dark | | Over-exp | | Wind | | Mean | | ΔMean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | |
| **Full (Ours)** † | **98.69** | **99.23** | **97.84** | **98.75** | **97.19** | **98.43** | **96.41** | **97.94** | **95.90** | **97.69** | **92.31** | **95.27** | **95.53** | **97.43** | **93.09** | **95.37** | **95.91** | **97.60** | **97.46** | **98.54** | **96.03** | **97.62** | — |
| **A.** *GeoPartLoss group removal* | | | | | | | | | | | | | | | | | | | | | | | |
| $-\mathcal{L}_{\text{align}}$ | 54.49 | 66.02 | 34.49 | 47.21 | 21.56 | 33.75 | 16.76 | 29.27 | 17.58 | 28.20 | 7.24 | 16.36 | 13.24 | 23.98 | 29.52 | 41.73 | 45.01 | 57.62 | 36.83 | 50.04 | 27.67 | 39.42 | −68.4 ◀ |
| $-\mathcal{L}_{\text{part}}$ | 97.28 | 98.28 | 96.39 | 97.78 | 96.51 | 97.92 | 95.58 | 97.25 | 95.14 | 97.00 | 90.89 | 94.12 | 95.36 | 97.11 | 91.30 | 94.04 | 94.31 | 96.45 | 96.48 | 97.85 | 94.92 | 96.78 | −1.1 |
| $-\mathcal{L}_{\text{alt}}$ | 97.47 | 98.45 | 96.96 | 98.17 | 96.84 | 98.15 | 95.59 | 97.38 | 95.46 | 97.25 | 90.98 | 94.12 | 95.63 | 97.35 | 92.41 | 94.88 | 94.19 | 96.49 | 96.68 | 98.07 | 95.22 | 97.03 | −0.8 |
| $-\mathcal{L}_{\text{distill}}$ | 97.63 | 98.56 | 96.70 | 98.03 | 96.16 | 97.72 | 95.36 | 97.17 | 94.16 | 96.44 | 89.42 | 92.95 | 94.77 | 96.77 | 91.79 | 94.40 | 94.32 | 96.53 | 96.24 | 97.77 | 94.65 | 96.64 | −1.4 |
| **B.** *Representation branch* | | | | | | | | | | | | | | | | | | | | | | | |
| CLS readout only | 93.56 | 96.10 | 90.43 | 94.15 | 87.55 | 92.29 | 87.39 | 92.09 | 84.01 | 89.56 | 73.19 | 81.26 | 84.78 | 90.15 | 81.91 | 87.56 | 86.20 | 91.09 | 87.34 | 91.95 | 85.64 | 90.62 | −10.4 ◀ |
| Part branch only | 97.51 | 98.58 | 96.19 | 97.81 | 95.54 | 97.35 | 94.69 | 96.87 | 93.13 | 95.90 | 88.10 | 92.34 | 93.58 | 96.06 | 91.98 | 94.58 | 94.40 | 96.66 | 95.73 | 97.41 | 94.09 | 96.36 | −1.9 |
| w/o Graph branch | 97.56 | 98.28 | 96.88 | 97.90 | 96.06 | 97.46 | 95.33 | 97.06 | 94.35 | 96.46 | 90.51 | 93.87 | 95.24 | 96.99 | 91.61 | 94.12 | 94.98 | 96.67 | 96.10 | 97.48 | 94.86 | 96.63 | −1.2 |
| **C.** *Prototype salience gate* | | | | | | | | | | | | | | | | | | | | | | | |
| All prototypes active | 97.79 | 98.62 | 96.81 | 98.09 | 95.44 | 97.24 | 95.34 | 97.08 | 93.08 | 95.81 | 86.44 | 91.24 | 93.32 | 95.81 | 90.55 | 93.57 | 92.86 | 95.60 | 95.54 | 97.25 | 93.72 | 96.03 | −2.3 ◁ |
| **D.** *UAPA (paired run; Δ within-block only)* | | | | | | | | | | | | | | | | | | | | | | | |
| w/o UAPA | 96.92 | 98.10 | 95.64 | 97.39 | 94.91 | 97.04 | 94.74 | 96.93 | 93.22 | 95.97 | 89.23 | 93.23 | 93.43 | 96.07 | 91.71 | 94.43 | 93.37 | 95.96 | 95.68 | 97.37 | 93.88 | 96.25 | −1.1 |
| w/ UAPA (paired) | 98.06 | 98.88 | 96.85 | 98.15 | 96.16 | 97.68 | 95.84 | 97.50 | 94.34 | 96.62 | 89.60 | 93.37 | 94.73 | 96.74 | 92.93 | 95.22 | 94.42 | 96.73 | 96.48 | 97.95 | 94.94 | 96.88 | — |

Table 4 sweeps four ablation axes (loss-group removal, representation branch, prototype salience, paired UAPA) under the same D→S weather-online evaluation as Table 3. Blocks A–C share an ablation-baseline checkpoint (Mean R@1 96.03%) trained with the matched weather-online recipe, which is a separate training run from the deployed weather-benchmark checkpoint of Table 3 (Mean 94.71%). The two share architecture and recipe family but are separate runs at different input resolutions (448 vs. 392), so within-table deltas, rather than absolute scores, are the comparable quantity; Block D is paired separately and read by its within-block delta.

*Block A (*GEOPARTLOSS *groups).* Cross-view alignment is the dominant training signal: removing the alignment group is catastrophic (Mean R@1 collapses by tens of points), while removing part quality, altitude, or distillation each costs only about a point of Mean R@1. The asymmetry shows that GEOPARTLOSS factorizes cleanly: alignment is the signal-carrying head, and the three auxiliary heads regularize and calibrate it, paying off under stress conditions rather than on Normal weather.

*Block B (representation branch).* The *CLS readout only* variant still trains the full head end-to-end (gradients still shape the backbone) and only drops the part and graph readouts at retrieval; even with this advantage, mean R@1 falls by roughly ten points relative to Full, with the largest gap on the hardest combined corruptions where layout is the only signal that survives texture loss. Reintroducing the part branch recovers most of the gap and the graph branch supplies the rest, supporting the identity/arrangement factorization motivated in Sec. 1. These Block B numbers differ from the *CLS-only* backbone-ablation baselines in Appendix A3.3, which strip the head from training entirely (no part/graph gradients during training) and therefore sit several points lower.

*Block C (prototype salience).* Disabling the salience gate while keeping the same prototype bank and graph readout (*All prototypes active*) costs about two points of Mean R@1, isolating the gain that comes from learning *which* prototypes matter per image rather than merely from prototype capacity. *Block D (paired UAPA).* On the paired seed/checkpoint run, uncertainty-aware temperature modulation lifts within-block Mean R@1 by roughly a point and Mean AP by half a point; we read this delta as the contribution of UAPA in isolation, without conflating it with the canonical benchmark score. Backbone-substrate controls (DINOv2 / iBOT / MoCo v3 ViT-S) are deferred to Appendix A3.3.

## 5 Conclusion

We built SKYPART around the principle that aerial cross-view matching should factor texture out and preserve layout, and instantiated it as a lightweight part-discovery head for patch-based ViTs. Under a strict single-pass protocol (no re-ranking, no TTA), SKYPART sets the new state of the art on SUES-200, University-1652, and DenseUAV while remaining among the most parameter-efficient methods in the comparison, and its margin grows under the ten WeatherPrompt conditions [Wen et al., 2025]. The ablations (Sec. 4.5) confirm the predicted hierarchy: alignment carries the retrieval signal, the three-way fusion absorbs most of the weather-induced drop, and altitude-FiLM acts as a training-time regularizer with no inference dependency. Zero-shot University-1652→SUES-200 transfer (Appendix A3.5) leads on D→S and remains competitive on S→D, supporting explicit grouping as a complement to global pooling. Broader impacts are in Appendix A5.2.

## Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers and the area chair for their constructive feedback, which improved this paper. All experiments were run on the free GPU quota provided by Kaggle. This work received no external funding, and the authors declare no competing interests.

## References

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(50):1–34, 2018. URL `https://www.jmlr.org/papers/v19/17-646.html`.

Relja Arandjelović and Andrew Zisserman. Three things everyone should know to improve object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2911–2918, 2012. doi: 10.1109/CVPR.2012.6248018.

Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. URL `https://openaccess.thecvf.com/content_cvpr_2016/html/Arandjelovic_NetVLAD_CNN_Architecture_CVPR_2016_paper.html`.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In *Proceedings of*

*the International Conference on Learning Representations (ICLR)*, 2022. doi: 10.48550/arXiv.2202.03555.

Irving Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94(2):115–147, 1987. doi: 10.1037/0033-295X.94.2.115.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021. doi: 10.1109/ICCV48922.2021.00951. URL https://openaccess.thecvf.com/content/ICCV2021/html/Caron_Emerging_Properties_in_Self-Supervised_Vision_Transformers_ICCV_2021_paper.html.

Quan Chen, Tingyu Wang, Zihao Yang, Haoran Li, Rongfeng Lu, Yaoqi Sun, Bolun Zheng, and Chenggang Yan. SDPL: Shifting-dense partition learning for UAV-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(11):11810–11824, 2024. doi: 10.1109/TCSVT.2024.3424196.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/papers/Chen_An_Empirical_Study_of_Training_Self-Supervised_Vision_Transformers_ICCV_2021_paper.pdf.

Zhongwei Chen, Zhao-Xu Yang, and Hai-Jun Rong. Multilevel embedding and alignment network with consistency and invariance learning for cross-view geo-localization. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–15, 2025. doi: 10.1109/TGRS.2025.3572775.

Han-Pang Chiu, Aveek Das, Philip Miller, Supun Samarasekera, and Rakesh Kumar. Precise vision-aided aerial navigation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 688–695, 2014. URL https://ieeexplore.ieee.org/document/6942633/.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning (ICML)*, 2016. URL https://proceedings.mlr.press/v48/cohenc16.html.

Andy Couturier and Moulay A. Akhloufi. A review on absolute visual localization for UAV. *Robotics and Autonomous Systems*, 135:103666, 2021. doi: 10.1016/j.robot.2020.103666.

Ming Dai, Jianhong Hu, Jiedong Zhuang, and Enhui Zheng. 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, 2022. doi: 10.1109/TCSVT.2021.3135013.

Ming Dai, Enhui Zheng, Zhenhua Feng, Lei Qi, Jiedong Zhuang, and Wankou Yang. Vision-based UAV self-positioning in low-altitude urban environments. *IEEE Transactions on Image Processing*, 33:493–508, 2024. doi: 10.1109/TIP.2023.3346279.

Fabian Deuser, Konrad Habel, and Norbert Oswald. Sample4Geo: Hard negative sampling for cross-view geo-localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16801–16810, 2023. doi: 10.1109/ICCV51070.2023.01545.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. doi: 10.48550/arXiv.2010.11929.

Haolin Du, Jingfei He, and Yuanqing Zhao. CCR: A counterfactual causal reasoning-based method for cross-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(11):11630–11643, 2024. doi: 10.1109/TCSVT.2024.3425509.

Vincent Dumoulin, Ethan Perez, Nathan Schucher, Florian Strub, Harm de Vries, Aaron Courville, and Yoshua Bengio. Feature-wise transformations. *Distill*, 2018. doi: 10.23915/distill.00011.

Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19790–19800, 2024. doi: 10.1109/CVPR52733.2024.01871.

Tongtong Feng, Qing Li, Xin Wang, Mingzi Wang, Guangyao Li, and Wenwu Zhu. Multi-weather cross-view geo-localization using denoising diffusion models. In *Proceedings of the 2nd Workshop on UAVs in Multimedia (UAVM)*, pages 35–39, 2024. doi: 10.1145/3689095.3689103.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning (ICML)*, 2015. doi: 10.48550/arXiv.1409.7495.

Yuan Gao, Haibo Liu, and Xiaohui Wei. Semantic concept perception network with interactive prompting for cross-view image geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(6):5343–5354, 2025. doi: 10.1109/TCSVT.2025.3533574.

Fawei Ge, Yunzhou Zhang, Li Wang, Wei Liu, Yixiu Liu, Sonya Coleman, and Dermot Kerr. Multilevel feedback joint representation learning network based on adaptive area elimination for cross-view geo-localization. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024. doi: 10.1109/TGRS.2024.3396330.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819, 2021. doi: 10.1007/s11263-021-01453-z.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022. URL `https://openaccess.thecvf.com/content/CVPR2022/html/He_Masked_Autoencoders_Are_Scalable_Vision_Learners_CVPR_2022_paper.html`.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. doi: 10.48550/arXiv.1503.02531. NIPS 2014 Deep Learning Workshop.

Kaiji Hou, Qiang Tong, Na Yan, Xiulei Liu, and Shoulu Hou. MCFA: Multi-scale cascade and feature adaptive alignment network for cross-view geo-localization. *Sensors*, 25(14):4519, 2025. doi: 10.3390/s25144519.

Sixing Hu, Mengdan Feng, Rang M. H. Nguyen, and Gim Hee Lee. CVM-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. URL `https://openaccess.thecvf.com/content_cvpr_2018/html/Hu_CVM-Net_Cross-View_Matching_CVPR_2018_paper.html`.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7482–7491, 2018. URL `https://openaccess.thecvf.com/content_cvpr_2018/html/Kendall_Multi-Task_Learning_Using_CVPR_2018_paper.html`.

Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Proxy anchor loss for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3238–3247, 2020. doi: 10.48550/arXiv.2003.13911. URL `https://openaccess.thecvf.com/content_CVPR_2020/html/Kim_Proxy_Anchor_Loss_for_Deep_Metric_Learning_CVPR_2020_paper.html`.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017. doi: 10.48550/arXiv.1609.02907.

Alexander Lappe and Martin A. Giese. Register and [CLS] tokens induce a decoupling of local and global features in large ViTs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://openreview.net/forum?id=KhavyzO9kK.

Qingge Li, Xiaogang Yang, Jiwei Fan, Ruitao Lu, Bin Tang, Siyu Wang, and Shuang Su. GeoFormer: An effective Transformer-based Siamese network for UAV geolocalization. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:9470–9491, 2024. doi: 10.1109/JSTARS.2024.3392812.

Zheng Li, Xueyi Zhang, Yanming Guo, Yuxiang Xie, Zhaoyun Ding, Siqi Cai, Haizhou Li, and Mingrui Lao. PAUL: Uncertainty-guided partition and augmentation for robust cross-view geo-localization under noisy correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Jinliang Lin, Zhiming Luo, Dazhen Lin, Shaozi Li, and Zhun Zhong. A self-adaptive feature extraction method for aerial-view geo-localization. *IEEE Transactions on Image Processing*, 34: 126–139, 2025. doi: 10.1109/TIP.2024.3513157.

Tsung-Yi Lin, Yin Cui, Serge Belongie, and James Hays. Learning deep representations for ground-to-aerial geolocalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. URL https://openaccess.thecvf.com/content_cvpr_2015/html/Lin_Learning_Deep_Representations_2015_CVPR_paper.html.

Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17581–17592, 2023. doi: 10.1109/ICCV51070.2023.01616.

Xiangzeng Liu, Ziyao Wang, Yue Wu, and Qiguang Miao. SeGCN: A semantic-aware graph convolutional network for UAV geo-localization. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:6055–6066, 2024. doi: 10.1109/JSTARS.2024.3370612.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. doi: 10.48550/arXiv.2006.15055.

Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. SegCLIP: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023. doi: 10.48550/arXiv.2211.14813.

Zhe Ma, Jianfeng Dong, Shouling Ji, Zhenguang Liu, Xuhong Zhang, Zonghui Wang, Sifeng He, Feng Qian, Xiaobo Zhang, and Lei Yang. Let all be whitened: Multi-teacher distillation for efficient visual retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5143–5151, 2024. URL https://ojs.aaai.org/index.php/AAAI/article/view/28207/28410.

David Marr and H. Keith Nishihara. Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 200(1140):269–294, 1978. doi: 10.1098/rspb.1978.0020.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/pdf?id=GLm1BA3C8p.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3967–3976, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/papers/Park_Relational_Knowledge_Distillation_CVPR_2019_paper.pdf.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. doi: 10.1609/aaai.v32i1.11671.

Yifan Ping, Jun Lu, Haitao Guo, Qingfeng Hou, Kun Zhu, Zehao Sang, and Tong Liu. DINO-MSRA: A novel network architecture for cross-view image retrieval and localization of UAV and satellite images. *Journal of Geo-information Science*, 27(7):1608–1623, 2025. doi: 10.12082/dqxxkx.2025.250051.

Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning CNN image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1655–1668, 2019. doi: 10.1109/TPAMI.2018.2846566.

Katarina Radoš, Marta Brkić, and Dinko Begušić. Recent advances on jamming and spoofing detection in GNSS. *Sensors*, 24(13):4210, 2024. doi: 10.3390/s24134210.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. doi: 10.1109/CVPR.2015.7298682. URL https://ieeexplore.ieee.org/document/7298682.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. doi: 10.48550/arXiv.1810.04650.

Tianrui Shen, Yingmei Wei, Lai Kang, Shanshan Wan, and Yee-Hong Yang. MCCG: A ConvNeXt-based multiple-classifier method for cross-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(3):1456–1468, 2024. doi: 10.1109/TCSVT.2023.3296074.

Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. Where am I looking at? Joint location and orientation estimation by cross-view matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Shi_Where_Am_I_Looking_At_Joint_Location_and_Orientation_Estimation_CVPR_2020_paper.html.

Jian Sun, Hao Sun, Lin Lei, Kefeng Ji, and Gangyao Kuang. TirSA: A three stage approach for UAV-satellite cross-view geo-localization based on self-supervised feature enhancement. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(9):7882–7895, 2024. doi: 10.1109/TCSVT.2024.3382717.

Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. URL https://openaccess.thecvf.com/content_ECCV_2018/html/Yifan_Sun_Beyond_Part_Models_ECCV_2018_paper.html.

Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. doi: 10.1109/CVPR42600.2020.00643. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Sun_Circle_Loss_A_Unified_Perspective_of_Pair_Similarity_Optimization_CVPR_2020_paper.html.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/68053af2923e00204c3ca7c6a3150cf7-Abstract.html.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://openreview.net/pdf?id=SkgpBJrtvS.

Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M. Patel. TransWeather: Transformer-based restoration of images degraded by adverse weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2353–2363, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Valanarasu_TransWeather_Transformer-Based_Restoration_of_Images_Degraded_by_Adverse_Weather_Conditions_CVPR_2022_paper.html.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. doi: 10.48550/arXiv.1807.03748.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. doi: 10.48550/arXiv.1710.10903. URL https://openreview.net/forum?id=rJXMpikCZ.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. doi: 10.48550/arXiv.2006.10726.

Tingyu Wang, Zhedong Zheng, Chenggang Yan, Jiyong Zhang, Yaoqi Sun, Bolun Zheng, and Yi Yang. Each part matters: Local patterns facilitate cross-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(2):867–879, 2022. doi: 10.1109/TCSVT.2021.3061265.

Tingyu Wang, Zhedong Zheng, Yaoqi Sun, Chenggang Yan, Yi Yang, and Tat-Seng Chua. Multiple-environment self-adaptive network for aerial-view geo-localization. *Pattern Recognition*, 152:110363, 2024. doi: 10.1016/j.patcog.2024.110363.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning (ICML)*, 2020. doi: 10.48550/arXiv.2005.10242.

Xiaowei Wang, Di Wang, Ke Li, Yifeng Wang, Chengjian Wang, Libin Sun, Zhihong Wu, Yiming Zhang, and Quan Wang. Enhancing cross-view geo-localization generalization via global-local consistency and geometric equivariance. arXiv preprint arXiv:2509.20684, 2025.

Jiahao Wen, Hang Yu, and Zhedong Zheng. Weatherprompt: Multi-modality representation learning for all-weather drone visual geo-localization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://nips.cc/virtual/2025/poster/118002.

Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. URL https://openaccess.thecvf.com/content_iccv_2015/html/Workman_Wide-Area_Image_Geolocalization_ICCV_2015_paper.html.

Qiong Wu, Yi Wan, Zhi Zheng, Yongjun Zhang, Guangshuai Wang, and Zhenyang Zhao. CAMP: A cross-view geo-localization method using contrastive attributes mining and position-aware partitioning. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024. doi: 10.1109/TGRS.2024.3448499.

Panwang Xia, Yi Wan, Zhi Zheng, Yongjun Zhang, and Jiwei Deng. Enhancing cross-view geo-localization with domain alignment and scene consistency. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(12):13271–13281, 2024. doi: 10.1109/TCSVT.2024.3443510.

Jin Xu, Junping Yin, Juan Zhang, and Tianyan Gao. Enhancing cross view geo localization through global local quadrant interaction network. *Scientific Reports*, 15:33431, 2025a. doi: 10.1038/s41598-025-18935-6.

Mengmeng Xu, Hongxiang Lv, Hai Zhu, Enlai Dong, and Fei Wu. AGEN: Adaptive error control-driven cross-view geo-localization under extreme weather conditions. *Sensors*, 25(12):3749, 2025b. doi: 10.3390/s25123749.

Wenjia Xu, Yaxuan Yao, Jiaqi Cao, Zhiwei Wei, Chunbo Liu, Jiuniu Wang, and Mugen Peng. UAV-VisLoc: A large-scale dataset for UAV visual localization. arXiv preprint arXiv:2405.11936, 2024.

Yuanze Xu, Ming Dai, Wenxiao Cai, and Wankou Yang. Precise gps-denied uav self-positioning via context-enhanced cross-view geo-localization. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 374–388. Springer, 2025c. doi: 10.1007/978-981-95-5628-1_26.

Jiaqiang Yang, Danyang Qin, Huapeng Tang, Sili Tao, Haoze Bie, and Lin Ma. DINOv2-based UAV visual self-localization in low-altitude urban environments. *IEEE Robotics and Automation Letters*, 10(2):2080–2087, 2025. doi: 10.1109/LRA.2025.3527762.

Yibin Ye, Xichao Teng, Shuo Chen, Zhang Li, Leqi Liu, Qifeng Yu, and Tao Tan. Exploring the best way for UAV visual localization under low-altitude multi-view observation condition: A benchmark. In *Findings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. doi: 10.48550/arXiv.2503.10692.

Zhedong Zheng, Yunchao Wei, and Yi Yang. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)*, pages 1395–1403, 2020. doi: 10.1145/3394171.3413896.

Zhedong Zheng, Yujiao Shi, Tingyu Wang, Chen Chen, Pengfei Zhu, and Richard Hartley. The 2nd workshop on UAVs in multimedia: Capturing the world from a new perspective. In *Proceedings of the 2nd Workshop on UAVs in Multimedia: Capturing the World from a New Perspective (UAVM '24), in conjunction with ACM Multimedia*, pages 1–3, 2024. doi: 10.1145/3689095.3695932.

Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3652–3661, 2017. doi: 10.1109/CVPR.2017.389.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT pre-training with online tokenizer. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. URL `https://openreview.net/pdf?id=ydopy-e6Dg`.

Runzhe Zhu, Ling Yin, Mingze Yang, Fei Wu, Yuncheng Yang, and Wenbo Hu. SUES-200: A multi-height multi-scene cross-view image benchmark across drone and satellite. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):4825–4839, 2023. doi: 10.1109/TCSVT.2023.3249204.

## Appendix Contents


**A1 Method Details** **18**

A1.1 Symbol Reference . . . . . 18

A1.2 Theoretical Motivation: Formal Arguments . . . . . 18

A1.3 Part-Head Mechanics, Masked Part Reconstruction, and Multi-Level Distillation . . 19

A1.4 UAPA: Formulation and Forward-KL Ablation . . . . . 20

A1.5 GeoPartLoss Grouping and Training Details . . . . . 21

**A2 Implementation and Protocol** **21**

A2.1 Hyperparameter Configuration . . . . . 21

A2.2 Dataset Statistics and Evaluation Protocol . . . . . 21

**A3 Extended Quantitative Results** **22**

A3.1 University-1652 Directional Comparison . . . . . 22

A3.2 DenseUAV Detailed Comparison . . . . . 22

A3.3 Backbone Ablation . . . . . 22

A3.4 Weather Robustness: Full Protocol & Implementation . . . . . 23

A3.5 Zero-shot Cross-Dataset Transfer . . . . . 26

A3.6 Additional Pareto Analysis . . . . . 26

**A4 Controlled Experiments and Additional Evaluations** **28**

A4.1 Representation Control: CLS-Only versus Full Fusion . . . . . 29

A4.2 Altitude-FiLM: Paired Intervention . . . . . 29

A4.3 Enhanced-Inference Operating Points . . . . . 29

A4.4 Resolution Controls for the Weather Comparison . . . . . 29

A4.5 Repair-Then-Match Diagnostic . . . . . 30

A4.6 Large-Gallery Weather Evaluation: University160k-WX . . . . . 30

A4.7 Cross-Dataset Transfer: SUES-200 $\rightarrow$ University-1652 . . . . . 31

A4.8 Coordinate Error on DenseUAV . . . . . 31

A4.9 Geographically Disjoint Evaluation on UAV-VisLoc . . . . . 31

**A5 Qualitative Analysis and Discussion** **32**

A5.1 Qualitative Visualizations . . . . . 32

A5.2 Limitations and Broader Impact . . . . . 32

A5.3 What We Tried That Did Not Work . . . . . 33

**A6 Ethics, Reproducibility, and LLM-Use Statements** **34**

Unless explicitly stated otherwise (Appendices A4.3 and A4.5), SKYPART results use the single-pass evaluation protocol of the main paper; reported baselines follow the source protocols stated in the corresponding tables.

# A1 Method Details

## A1.1 Symbol Reference

Table 5: Key symbols and their corresponding modules.

| Symbol | Module | Section |
|---|---|---|
| $\mathbf{P}, \mathbf{p}_k$ | Prototype bank: $K_{\max}$=12 slots (architectural ceiling), $d_p$=256; the salience gate selects a per-image active subset of typical size $K$=10 | 3.2 |
| $\mathbf{A}_{i,k}$ | Soft assignment matrix | 3.2 |
| $\hat{\mathbf{p}}_k$ | Part-aggregated features | 3.2 |
| $g_k$ | Per-prototype salience gate (Eq. 5) | A1.3 |
| $w_p, w_c, w_g$ | Three-way fusion weights (Eq. 2) | 3.2 |
| $\gamma_a, \beta_a$ | FiLM scale & shift | 3.3 |
| $\mathcal{L}_{\text{mar}}$ | Masked Part Reconstruction | A1.3 |
| $\mathcal{L}_{\text{proxy}}$ | Proxy Anchor loss | 3.4 |
| $\mathcal{L}_{\text{uapa}}$ | Uncertainty-aware proxy alignment | 3.4 |
| $\mathcal{L}_{\text{div}}$ | Prototype diversity regularizer | 3.4 |
| $s_g$ | GeoPartLoss learnable log-variance | 3.4 |

## A1.2 Theoretical Motivation: Formal Arguments

The three design choices of Sec. 3.1 (an identity/arrangement factorization of the retrieval embedding, a softmax-over-prototype-indices assignment, and altitude as a training-time nuisance), together with the uncertainty-weighted multi-task training of Sec. 3.4, are stated there only briefly. This section expands each into a formal argument. None of the results below are claims of mathematical novelty; they are restatements of established theorems [Achille and Soatto, 2018, Locatello et al., 2020, Perez et al., 2018, Kendall et al., 2018, Sener and Koltun, 2018] applied to the CVGL setting. For convenience, we label the four arguments C1 (layout as sufficient statistic), C2 (slot-competition assignment), C3 (altitude as marginalized nuisance), and C4 (Pareto-stationary MTL).

**C1 as an information-bottleneck factorization.** Let $X_d, X_s$ be drone and satellite views of the same scene with geolocation identity $Y$, and let $N = (N_{\text{view}}, N_{\text{alt}}, N_{\text{weather}})$ be the jointly observed nuisances. Write the CVGL retrieval embedding as $T = f_\theta(X)$. The information bottleneck objective [Achille and Soatto, 2018] asks for

$$\min_\theta \; I(T;X) - \beta\, I(T;Y), \tag{3}$$

with a separate *invariance* term $I(T;N) \to 0$. Factorize $T = (T_{\text{id}}, T_{\text{arr}})$ where $T_{\text{id}}$ is the part-identity descriptor (drone's part bank encoding) and $T_{\text{arr}}$ is the part-arrangement descriptor (graph readout over prototype nodes). Two elementary observations motivate the split: (i) the part-arrangement statistic is approximately invariant to $N_{\text{view}}$ (spatial layout of roads/roofs is preserved under oblique reprojection of a single planar scene) and to $N_{\text{weather}}$ (weather destroys local texture but preserves spatial structure), so $I(T_{\text{arr}};N) \ll I(T_{\text{id}};N)$; (ii) $T_{\text{arr}}$ alone is insufficient-two disjoint locations can share the same layout, so $I(T_{\text{arr}};Y) < I(T;Y)$ and $T_{\text{id}}$ must remain in the embedding. The three-branch fusion $\mathbf{f} = w_p\mathbf{f}_{\text{part}} + w_c\mathbf{f}_{\text{cls}} + w_g\mathbf{f}_{\text{graph}}$ is then exactly a learned linear combination of an identity-heavy component ($\mathbf{f}_{\text{part}}$, $\mathbf{f}_{\text{cls}}$) and an arrangement-heavy component ($\mathbf{f}_{\text{graph}}$), with the weights $w_p, w_c, w_g$ learned end-to-end under the alignment objective. Empirically, the *No-Graph* ablation (Table 4, Block B) costs 1.17 pp Mean R@1 but 1.80 pp on Fog+Snow: the gap widens exactly in the regime where $T_{\text{id}}$'s texture component is corrupted and the factorization is forced to rely on $T_{\text{arr}}$.

**C2 as a single-pass slot assignment.** Slot attention [Locatello et al., 2020] iterates an attention-then-GRU update for $K$ slots competing over $L$ input tokens. The competition comes from a softmax *across slots* (not across tokens), producing mutually-exclusive assignments that concentrate each token on one slot. Our single-pass assignment (Eq. 1) is the fixed point of one such iteration when (a)

the slot-to-input similarity is parametrized as cosine with temperature $\tau$, (b) the GRU update reduces to identity, and (c) the slot initialization is the learned prototype bank rather than a sampled noise. In return for dropping the iterative refinement we lose a reconstruction-consistent slot assignment, but we do not need one. The downstream loss is not reconstruction; it is cross-view retrieval, for which a single competitive assignment is a sufficient statistic. The cost we pay is architectural: prototypes must be *learnable and directly trained*, rather than emerging from reconstruction pressure.

**C3 as nuisance marginalization via FiLM mean.** FiLM applies an affine modulation $\text{FiLM}(\mathbf{z}; \gamma_a, \beta_a) = \gamma_a \odot \mathbf{z} + \beta_a$ conditioned on an altitude bin $a \in A$. At inference, when altitude is unknown we replace $(\gamma_a, \beta_a)$ with $(\bar{\gamma}, \bar{\beta}) = \frac{1}{|A|} \sum_a (\gamma_a, \beta_a)$. Under a uniform prior $p(a)$ this is exactly the expectation $\mathbb{E}_{a \sim p(a)}[\text{FiLM}(\mathbf{z}; \gamma_a, \beta_a)]$, the marginalized modulation. Because FiLM is linear in $(\gamma_a, \beta_a)$, the marginal is exactly the modulation produced by the marginal parameters *at the FiLM-operator output*. This equality is exact only at the affine modulation step; we do not claim the final fused embedding equals the marginal of the altitude-conditioned embedding, since downstream nonlinear layers (assignment softmax, GAT, fusion gate) break linearity. With this scope, the mean-FiLM fallback is a theoretically grounded substitution at the modulation operator rather than an ad-hoc hack. Conditioning the retrieval embedding directly on $a$ (rather than conditioning only the clustering pathway and letting the final embedding collapse the altitude dimension through fusion) would instead make the embedding altitude-dependent at inference, which fails the deployment constraint motivated in Sec. 3.1.

**C4 as Pareto-stationary MTL.** Let $\{\mathcal{L}_g\}_{g=1}^{G}$ be the four group losses. Kendall homoscedastic uncertainty weighting [Kendall et al., 2018] models each task as a Gaussian likelihood with learned variance $\sigma_g^2$: the log-likelihood contributes $\frac{1}{2\sigma_g^2}\mathcal{L}_g + \frac{1}{2}\log\sigma_g^2$, which after the reparameterization $s_g = \log\sigma_g^2$ becomes $\frac{1}{2}\exp(-s_g)\mathcal{L}_g + \frac{1}{2}s_g$. We adopt the equivalent form that drops the factor of $\frac{1}{2}$ (absorbed into the learning rate),

$$\mathcal{L}_{\text{total}} = \sum_g \big[\exp(-s_g)\,\mathcal{L}_g + s_g\big], \tag{4}$$

which is the training objective introduced in Sec. 3.4. Sener & Koltun [Sener and Koltun, 2018] prove that any stationary point of a non-negative weighted sum $\sum_g \alpha_g \mathcal{L}_g$ with $\alpha_g \geq 0$ and $\sum_g \alpha_g > 0$ is a Pareto-stationary point of the vector objective $(\mathcal{L}_1, \ldots, \mathcal{L}_G)$: there is no descent direction that improves all $\mathcal{L}_g$ simultaneously. Eq. 4 produces positive weights $\alpha_g = \exp(-s_g)$ by construction, so its stationary points inherit Pareto-stationarity. The full $\alpha_g\mathcal{L}_g - \log\alpha_g$ objective prevents degenerate weights; the $+s_g$ term specifically prevents $\alpha_g \to 0$, and the learned values (alignment 0.94, distillation 1.05, part 1.12, altitude 1.18; Appendix A1.5) land in a tight range, a diagnostic that no single task is dominating. We do not claim that Pareto-stationarity is equivalent to Pareto-optimality; we claim only that Eq. 4 removes a four-dimensional weight sweep and replaces it with a stationarity property that follows from [Sener and Koltun, 2018].

**Why these four, and why not more.** A reader may ask why SKYPART does not additionally commit to, e.g., an equivariant backbone or an explicit nuisance adversary. Both are in the prior-art space ([Cohen and Welling, 2016, Ganin and Lempitsky, 2015]); we omitted both by design. Equivariance would bake a fixed rotation/scale prior into the backbone, but CVGL altitude/rotation distributions are dataset-specific (SUES-200 canonically oriented, DenseUAV arbitrary)-a fixed prior helps one and hurts the other. A domain-adversarial branch would couple training stability to a minimax game we explicitly want to avoid, given the Pareto-stationarity argument above. C1–C4 are the minimal set that determined our design; we do not claim they are the only reasonable set.

### A1.3 Part-Head Mechanics, Masked Part Reconstruction, and Multi-Level Distillation

**Salience gate and active-set selection.** The per-image salience gate of Sec. 3.2 selects which prototypes are active for a given image. It is a Gumbel-sigmoid relaxation over per-prototype logits $\mathbf{s} \in \mathbb{R}^{K_{\max}}$ computed from the part-aggregated descriptors $\hat{\mathbf{p}}_k$ (post-refine; the same tensor used by the graph readout and MAR) via a lightweight 2-layer MLP, $s_k = \text{MLP}_g(\hat{\mathbf{p}}_k)$ with hidden width $d_p/4$=64, GELU activation, and final-bias initialized to $b_2$=2.0 so the gate starts biased open ($\sigma(2.0) \approx 0.88$) and learns to prune:

$$g_k = \sigma\big((s_k + \epsilon_k)/\tau_g\big), \qquad \epsilon_k \sim \text{Gumbel}(0, 1), \tag{5}$$

with $\tau_g$=0.5. A floor $K_{\min}$=4 is enforced at training time to prevent the gate from collapsing to a single prototype early. Each active prototype aggregates its assigned patch tokens through the soft assignment $\mathbf{A}$, passes through a small residual MLP, and its coordinate-normalized centroid in $[0,1]^2$ is retained as the node position for the graph readout; this 2-D position is fed to the GAT as a node feature alongside the part descriptor, letting the layout readout depend on spatial relations between prototypes, not just their identities.

**Per-branch projections to the fusion space.** The three branches in Eq. 2 are projected to a shared 768-D space before the fusion gate combines them. (i) The CLS branch maps the 384-D DINOv2 ViT-S/14 CLS token through $\mathbf{f}_{\text{cls}} = \text{ReLU}(\text{BN}(W_c\,\text{CLS}))$ with $W_c \in \mathbb{R}^{768\times384}$. (ii) The part branch passes each of the $K$ active 256-D part descriptors $\hat{\mathbf{p}}_k$ (post-refine, salience-weighted) through PartAwarePooling: it concatenates the top-3 part descriptors into a $3\times256$=768 vector and applies a 2-layer MLP $\mathbb{R}^{768}\to\mathbb{R}^{768}$ (Linear, LayerNorm, GELU, Linear) to produce $\mathbf{f}_{\text{part}}$. (iii) The graph branch projects the 256-D GAT readout $\mathbf{h}_{\text{GAT}}$ through $\mathbf{f}_{\text{graph}} = \text{LN}(W_g\,\mathbf{h}_{\text{GAT}})$ with $W_g \in \mathbb{R}^{768\times256}$. The fusion gate's MLP is $\mathbb{R}^{3\cdot768}\to\mathbb{R}^{384}\to\mathbb{R}^3$, followed by softmax. All three branch outputs are L2-normalized prior to the weighted sum.

**Masked Part Reconstruction (MAR).** 30% of patch tokens are masked; each masked position $i$ is reconstructed through the part-aggregated features:

$$\mathcal{L}_{\text{mar}} = \frac{1}{|\mathcal{M}|}\sum_{i\in\mathcal{M}}\big(1-\cos(\hat{\mathbf{z}}_i,\mathbf{z}_i)\big), \qquad \hat{\mathbf{z}}_i = \text{Decoder}\big(\sum_k \mathbf{A}_{i,k}\,\hat{\mathbf{p}}_k\big), \tag{6}$$

where the decoder is a 2-layer MLP (GELU, LayerNorm, hidden size 512). Prototype aggregation $\hat{\mathbf{p}}_k$ uses only visible (unmasked) tokens, preventing target leakage. Soft assignments $\mathbf{A}_{i,k}$ at masked positions are detached (`stop_gradient`): without this, gradients flow back through assignment weights and one dominant prototype monopolizes all masked positions to trivially minimize $\mathcal{L}_{\text{mar}}$, bypassing diversity. The 30% mask ratio and 10-epoch warm-up balance reconstruction difficulty with optimization stability.

**Multi-level distillation.** We distill from a frozen DINOv2-B/14 teacher and an EMA self-ensemble, and regress the normalized ground-truth altitude $\tilde{a} = (a-150)/150$, where $a$ is the raw altitude in meters and $\hat{a}$ denotes the prediction of the altitude regression head:

$$\mathcal{L}_{\text{cross}} = \text{MSE}(\phi\mathbf{f},\mathbf{f}^{\mathcal{T}}) + (1-\cos(\phi\mathbf{f},\mathbf{f}^{\mathcal{T}})), \tag{7}$$

$$\mathcal{L}_{\text{ema}} = 1-\cos(\mathbf{f},\mathbf{f}^{\text{ema}}) \quad (\alpha\text{=}0.996\ \text{[Tarvainen and Valpola, 2017]}), \tag{8}$$

$$\mathcal{L}_{\text{alt}} = \text{SmoothL1}(\hat{a},\tilde{a}). \tag{9}$$

The teacher mapping $\phi$ is a lightweight projector $\phi(\mathbf{f}) = \text{LayerNorm}(\mathbf{W}\mathbf{f})$ with $\mathbf{W}\in\mathbb{R}^{768\times768}$ matching student to teacher dimensionality.

### A1.4 UAPA: Formulation and Forward-KL Ablation

**UAPA formulation.** Let $\mathbf{p}_d, \mathbf{p}_s$ be the softmax probabilities over the $|C|$ proxy classes, with one proxy per training location ($|C|$=120 for SUES-200, 701 for University-1652, and 2,256 for DenseUAV). The proxy-classification head reuses the $|C|$ proxy embeddings from metric learning: each logit is the cosine similarity between the normalized retrieval embedding and a learned class proxy, followed by a softmax over the $|C|$ classes. Entropy is computed over these proxy-class probabilities.

UAPA modulates only the KL temperature using the positive relative entropy gap between the drone and satellite proxy distributions:

$$T = T_0\big(1+\Delta H_+/H_{\max}\big), \qquad \Delta H_+ = \max\{H(\mathbf{p}_d)-H(\mathbf{p}_s),0\}, \tag{10}$$

where $H(\cdot)$ is Shannon entropy, $T_0$=4, and $H_{\max}=\log|C|$ normalizes the entropy gap. Since $0\le\Delta H_+\le H_{\max}$, the effective UAPA temperature lies in $[T_0,2T_0]=[4,8]$. When the drone view is more uncertain than the satellite view, $T$ increases and softens the satellite-teacher target; otherwise the temperature remains at the base value. The loss itself is the standard temperature-softened reverse KL evaluated at $T$, $\mathcal{L}_{\text{uapa}} = T^2\,\text{KL}\big(\text{softmax}(\mathbf{z}_d/T)\,\big\|\,\text{softmax}(\mathbf{z}_s/T)\big)$, where $\mathbf{z}_d,\mathbf{z}_s$ are the proxy-class logits and the $T^2$ scaling preserves gradient magnitude across temperatures, so only the temperature (not a whole-sample re-weighting) is modulated.

**KL direction.** We use reverse KL (drone‖satellite, mode-seeking) rather than forward KL (mode-covering), so the drone branch is encouraged to concentrate on the most confident satellite peak. A forward-KL variant plateaus at 96.52% R@1 versus 98.725% for reverse KL, the deployed model of Table 1 ($-2.205$ pp). GeoPartLoss group weights converge to similar ratios under either choice (alignment 0.97, distillation 1.01, part 1.03, altitude 1.04 for forward KL), suggesting that self-calibration is robust to the KL direction.

### A1.5 GeoPartLoss Grouping and Training Details

Table 6: **GEOPARTLOSS groups** (aligned with implementation). Converged weights are end-of-training averages from the SUES-200 canonical run (60 epochs); higher $\exp(-s_g)$ = group receives stronger gradient.

| Group | Constituent losses | Weighting | Converged $\exp(-s_g)$ |
|---|---|---|---|
| Alignment | $\mathcal{L}_{\text{circle}} + \mathcal{L}_{\text{nce}} + \mathcal{L}_{\text{pnce}} + \mathcal{L}_{\text{proxy}} + \mathcal{L}_{\text{uapa}}$ | $\exp(-s_{\text{align}})$ | 0.94 |
| Distillation | $\mathcal{L}_{\text{cross}} + \mathcal{L}_{\text{ema}}$ | $\exp(-s_{\text{distill}})$ | 1.05 |
| Part quality | $\mathcal{L}_{\text{mar}} + \mathcal{L}_{\text{div}}$ | $\exp(-s_{\text{part}})$ | 1.12 |
| Altitude | $\mathcal{L}_{\text{alt}}$ | $\exp(-s_{\text{alt}})$ | 1.18 |
| Regularizer | $+\sum_g s_g$ (Kendall uncertainty) | automatic | — |

The learned group weights $\exp(-s_g)$ evolve during training rather than collapsing to a fixed split: from uniform initialization they diverge within the first 20 epochs and settle by epoch 60 at approximately alignment $0.94$, distillation $1.05$, part quality $1.12$, altitude $1.18$. Alignment dominates early, when the embedding geometry is being established, and then relaxes as metric learning saturates while the part-quality and altitude weights rise to compensate, the behavior Kendall weighting should show as harder objectives take over. The log-variance regularizer $+\sum_g s_g$ keeps all four groups from collapsing to zero so gradient flow stays healthy throughout the schedule. The diversity regularizer in the part-quality group penalizes off-diagonal cosine similarities between L2-normalized prototypes, $\mathcal{L}_{\text{div}} = \frac{1}{K_{\max}(K_{\max}-1)} \sum_{j \neq k} \langle \bar{\mathbf{p}}_j, \bar{\mathbf{p}}_k \rangle^2$, pushing the bank toward an orthogonal (well-spread) set so different prototypes attend to genuinely different scene regions. The patch-level NCE term in the alignment group uses, as positives, patch pairs assigned to the same prototype in both views and, as negatives, all other patches in the batch; this is what allows the loss to remain well-defined when a prototype's role drifts during training (see Appendix A5.2 on label equivariance).

## A2 Implementation and Protocol

### A2.1 Hyperparameter Configuration

Table 8 lists the hyperparameters of the deployed clean configuration. Module-level settings are given in Appendix A1.3, and experiment-specific overrides (e.g., the $392\times392$ weather runs) in Table 13.

### A2.2 Dataset Statistics and Evaluation Protocol

Table 7 lists the train split sizes and the per-direction query/gallery counts used by all retrieval evaluations in the main paper.

Table 7: Training split sizes and test query/gallery counts for both retrieval directions on the three benchmarks.

| Dataset | Training phase | | Test phase | | | |
|---|---|---|---|---|---|---|
| | | | Drone→Satellite | | Satellite→Drone | |
| | Drone | Satellite | Query | Gallery | Query | Gallery |
| University-1652 [Zheng et al., 2020] | 37,854 | 701 | 37,855 | 951 | 701 | 51,355 |
| SUES-200 [Zhu et al., 2023] | 24,000 | 120 | 16,000 | 200 | 80 | 40,000 |
| DenseUAV [Dai et al., 2024] | 6,768 | 13,536 | 2,331 | 18,198 | 4,662 | 9,099 |

Table 8: Complete hyperparameter configuration.

| Category | Parameter | Value |
|---|---|---|
| Architecture | Backbone | DINOv2 ViT-S/14 |
| | Unfrozen blocks | 6 (of 12) |
| | Prototype bank size $K_{\max}$ | 12 |
| | Active prototype count $K$ (typical) | 10 |
| | Part dimension $d_p$ | 256 |
| | Embedding dim $D$ | 768 |
| | Teacher dim | 768 |
| | Cluster temperature $\tau$ | 0.07 |
| Training | Input size | $448 \times 448$ |
| | Epochs | 60 |
| | PK sampler ($P$, $M$) | (16, 4) |
| | Head LR | $3 \times 10^{-4}$ |
| | Backbone LR | $3 \times 10^{-5}$ |
| | Weight decay | 0.05 |
| | Warmup epochs | 5 |
| | LR floor | 1% of peak |
| Loss | Proxy margin $\delta$ | 0.1 |
| | Proxy scale $\alpha$ | 32 |
| | EMA decay $\alpha$ | 0.996 |
| | MAR mask ratio | 0.30 |
| | MAR warmup | 10 epochs |
| | Distillation temp | 4.0 |
| | Num loss groups | 4 |
| | Label smoothing | 0.1 |

# A3 Extended Quantitative Results

## A3.1 University-1652 Directional Comparison

Table 9 extends the main-paper Table 2(a) with a wider set of University-1652 baselines and reports both retrieval directions for each.

## A3.2 DenseUAV Detailed Comparison

Table 10 extends the main-paper Table 2(b) with additional DenseUAV metrics (R@5, R@top1, AP, SDM@{1,3,5}) for a broader set of baselines with available detailed metrics.

## A3.3 Backbone Ablation

Table 11 reports a controlled *SSL ViT generalisation* ablation on SUES-200 with everything else held identical: resolution, loss grouping, augmentation recipe, optimizer, epoch budget, and confusion gallery. The same Full SKYPART head is attached to three ViT-S self-supervised backbones (DINOv2 [Oquab et al., 2024], iBOT [Zhou et al., 2022], MoCo v3 [Chen et al., 2021]) under a strictly matched 60-epoch budget with teacher distillation *disabled* (CLS-only baselines are 60 epochs, same recipe), so the head's plug-and-play behavior is read off the lift over each backbone's own CLS-only rather than off absolute R@1. The deployed configuration (DINOv2 ViT-S/14 + Full head, 60 epochs, distillation *enabled*) is not retrained here; its headline number (Mean R@1 98.725% D→S) is the row in Table 1, and the DINOv2 row of Table 11 (98.48%) is the within-block SSL anchor only. We omit a ConvNeXt-T row because, in our pilot, the Full SKYPART head on ImageNet-supervised ConvNeXt-T regressed from 92.33% (CLS-only) to 90.41% Mean R@1 under the same 60-epoch recipe, a negative transfer consistent with the prototype-to-patch assignment requiring patch-level semantic coherence that emerges in self-distilled ViTs but not in supervised CNN feature maps. Per-altitude D→S and S→D R@1 are reported for each row.

The three SSL substrates tell a single story sharpened by the patch-token quality of each backbone. On the two self-distilled substrates (DINOv2, iBOT), CLS-only is already moderately strong (88.68 / 84.54% Mean D→S R@1), and the Full SKYPART head delivers a clean positive lift across

Table 9: **University-1652 directional comparison.** We report both retrieval directions, Drone→Satellite and Satellite→Drone, using R@1 and AP. "–" indicates values not reported in the cited papers/tables under the exact matched protocol. **Red**: best; **blue**: second-best.

| Method | Venue | Params | Drone → Satellite | | Satellite → Drone | |
|---|---|---|---|---|---|---|
| | | | R@1 | AP | R@1 | AP |
| ViT [Dosovitskiy et al., 2021] | ICLR'21 | – | 74.09 | 77.82 | 83.31 | 72.27 |
| FSRA [Dai et al., 2022] | TCSVT'22 | – | 85.50 | 87.53 | 89.73 | 84.94 |
| Swin-B | ICCV'21 | – | 84.15 | 86.62 | 90.30 | 83.55 |
| LPN [Wang et al., 2022] | TCSVT'22 | 62.4 M | 77.71 | 80.80 | 90.30 | 78.78 |
| RK-Net | TIP'22 | – | 68.10 | 71.53 | 80.96 | 69.35 |
| SwinV2-B | CVPR'22 | – | 86.99 | 89.02 | 91.16 | 85.77 |
| Sample4Geo (ResNet-50) [Deuser et al., 2023] | ICCV'23 | – | 78.62 | 82.11 | 87.45 | 76.32 |
| Sample4Geo (ConvNeXt-B) [Deuser et al., 2023] | ICCV'23 | 87.6 M | 92.65 | 93.81 | 95.14 | 91.39 |
| DenseUAV baseline [Dai et al., 2024] | TIP'24 | – | 82.22 | 84.78 | 87.59 | 81.49 |
| F3-Net | TGRS'23 | – | 78.64 | 81.60 | – | – |
| Song | GRSL'23 | – | 83.26 | 85.84 | 90.30 | 82.71 |
| AEN (w. LPN) | SPL'24 | – | 77.40 | 80.27 | 90.30 | 76.01 |
| MuSe-Net [Wang et al., 2024] | PR'24 | 50.5 M | 74.48 | 77.83 | 88.02 | 75.10 |
| TransFG | TGRS'24 | – | 87.92 | 89.99 | 93.37 | 87.94 |
| MFJRLN(Lcro) [Ge et al., 2024] | TGRS'24 | – | 87.61 | 89.57 | 91.07 | 86.72 |
| GeoFormer [Li et al., 2024] | JSTARS'24 | – | 88.16 | 90.03 | 91.87 | 87.92 |
| MCCG [Shen et al., 2024] | TCSVT'24 | 56.6 M | 89.28 | 91.01 | 94.29 | 89.29 |
| SDPL [Chen et al., 2024] | TCSVT'24 | – | 90.16 | 91.64 | 93.58 | 89.45 |
| CCR [Du et al., 2024] | TCSVT'24 | 156.6 M | 92.54 | 93.78 | 95.15 | 91.80 |
| CAMP [Wu et al., 2024] | TGRS'24 | 91.4 M | 94.46 | 95.38 | 96.15 | 92.72 |
| DAC [Xia et al., 2024] | TCSVT'24 | 96.5 M | **94.67** | **95.50** | **96.43** | **93.79** |
| MEAN [Chen et al., 2025] | TGRS'25 | **36.5 M** | 93.55 | 94.53 | 96.01 | 92.08 |
| **SkyPart (Ours)** | – | **26.95 M** | **96.47** | **97.77** | **98.43** | **98.24** |

Table 10: **DenseUAV detailed comparison** with additional retrieval and SDM metrics. Top and second-best entries follow the same highlighting rule as the main tables. **Red**: best; **blue**: second.

| Method | Venue | Backbone | R@1 | R@5 | R@top1 | AP | SDM@1 | SDM@3 | SDM@5 |
|---|---|---|---|---|---|---|---|---|---|
| Triplet Loss | CVPR'15 | ResNet-50 | 11.88 | 32.22 | – | – | 21.91 | – | – |
| Instance Loss | ACMMM'20 | ResNet-50 | 13.00 | 35.78 | – | – | 23.61 | – | – |
| LCM | Remote Sens'21 | ResNet-50 | 25.37 | 50.92 | – | – | 35.52 | – | – |
| MSBA | Remote Sens'21 | ResNet-50 | 46.13 | 64.22 | – | – | 52.64 | – | – |
| LPN (ResNet-50) | TCSVT'22 | ResNet-50 | 32.43 | 56.80 | – | – | 40.26 | – | – |
| LPN (ViT-S) | TCSVT'22 | ViT-S | 71.77 | 90.13 | – | – | 77.95 | – | – |
| FSRA | TCSVT'22 | ViT-S | 81.21 | 94.55 | 99.89 | 71.93 | 85.11 | 83.54 | 79.74 |
| RK-Net | TIP'22 | ResNet-50 | 38.74 | 62.85 | – | – | 45.78 | – | – |
| Sample4Geo | ICCV'23 | ConvNeXt-B | 49.38 | 78.29 | 99.40 | 35.93 | 61.72 | – | – |
| DenseUAV baseline | TIP'24 | ViT-S | 83.01 | 95.58 | **99.91** | 72.10 | 86.50 | 84.50 | **80.44** |
| MCCG | TCSVT'24 | ConvNeXt-T | 83.14 | 93.39 | 99.74 | 72.60 | 85.94 | 84.32 | 80.14 |
| Yang et al. | RA-L'25 | DINOv2-B | 86.27 | **96.83** | – | – | 88.87 | – | – |
| CEUSP | PRCV'25 | ConvNeXt-T | **89.45** | 96.05 | **100.00** | **79.62** | **91.01** | **89.42** | **85.34** |
| **SkyPart (Ours)** | – | **DINOv2-S** | **91.85** | **97.81** | 99.61 | **92.52** | **93.59** | **92.40** | 79.33 |

all altitudes (+9.80 pp on DINOv2, +5.62 pp on iBOT)-the head adds usable layout structure on top of an already coherent patch grid, and the deployed 98.725% Mean R@1 on D→S (Table 1) comes from combining this DINOv2 patch substrate with the teacher-distillation pathway that this ablation deliberately strips out. On the contrastive substrate (MoCo v3), CLS-only collapses to 28.19% Mean R@1 because contrastive pretraining optimises the CLS slot but leaves patch tokens diffuse; the Full SkyPart head recovers the structural signal that the patch-affinity graph and prototype assignment can read out and lifts D→S Mean R@1 by +33.49 pp (and S→D Mean R@1 by +31.88 pp). Combined with the negative-transfer outcome on ImageNet-supervised ConvNeXt-T noted above, the picture is consistent: SkyPart's contribution is *architectural* (it activates the latent layout signal that any patch-token SSL ViT exposes), and the magnitude of the lift scales inversely with how much of that signal the global CLS slot has already captured.

Table 11: **Backbone generalisation on SUES-200** (Normal weather, single-scale 448×448, full 200-tile gallery, no re-ranking, no TTA). We compare self-supervised ViT backbones under a controlled setting (Full = 60 epochs, teacher distillation disabled; CLS-only baseline = 60 epochs with the same recipe). The proposed head transfers cleanly between self-distilled patch substrates (DINOv2, iBOT) and provides a large lift on the contrastive MoCo v3 substrate, whose CLS embedding is substantially weaker for dense retrieval. The deployed SKYPART model, trained for 60 epochs with teacher distillation enabled, is reported as the headline result.

| | | Drone → Satellite, R@1 (%) | | | | | Satellite → Drone, R@1 (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Backbone** | **Head** | **150m** | **200m** | **250m** | **300m** | **Mean** | **150m** | **200m** | **250m** | **300m** | **Mean** |
| DINOv2 ViT-S/14 | CLS-only | 82.20 | 88.08 | 91.80 | 92.65 | 88.68 | 97.50 | 93.75 | 97.50 | 96.25 | 96.25 |
| **DINOv2 ViT-S/14** | **Full SKYPART** | **96.25** | **98.75** | **99.30** | **99.60** | **98.48** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** |
| | Δ *vs CLS* | *+14.05* | *+10.67* | *+7.50* | *+6.95* | *+9.80* | *+2.50* | *+6.25* | *+2.50* | *+3.75* | *+3.75* |
| iBOT ViT-S/16 | CLS-only | 73.80 | 85.40 | 88.75 | 90.20 | 84.54 | 88.75 | 91.25 | 92.50 | 93.75 | 91.56 |
| iBOT ViT-S/16 | Full SKYPART | 82.43 | 91.10 | 93.63 | 93.50 | 90.16 | 96.25 | 97.50 | 97.50 | 97.50 | 97.19 |
| | Δ *vs CLS* | *+8.63* | *+5.70* | *+4.88* | *+3.30* | *+5.62* | *+7.50* | *+6.25* | *+5.00* | *+3.75* | *+5.63* |
| MoCo v3 ViT-S/16 | CLS-only | 20.90 | 28.55 | 31.85 | 31.48 | 28.19 | 36.25 | 50.00 | 55.00 | 58.75 | 50.00 |
| MoCo v3 ViT-S/16 | Full SKYPART | 48.25 | 61.80 | 66.98 | 69.68 | 61.68 | 71.25 | 82.50 | 87.50 | 86.25 | 81.88 |
| | Δ *vs CLS* | *+27.35* | *+33.25* | *+35.13* | *+38.20* | *+33.49* | *+35.00* | *+32.50* | *+32.50* | *+27.50* | *+31.88* |

### A3.4 Weather Robustness: Full Protocol & Implementation

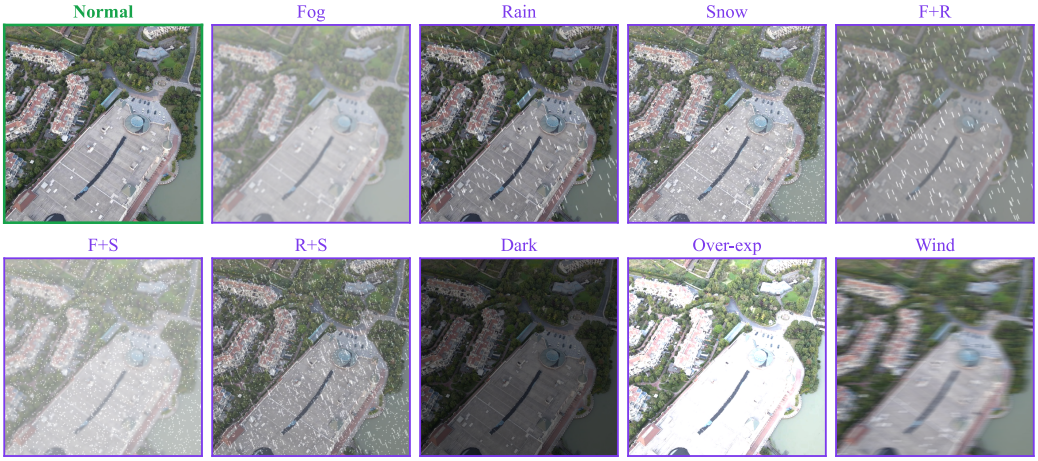


Figure 3: **Weather conditions.** The same drone image under 10 WeatherPrompt augmentations. Texture is destroyed, but spatial structure persists, a pattern qualitatively aligned with layout-heavy representations and with SKYPART's relative robustness under environmental noise.

#### A3.4.1 Evaluation Protocol

The evaluation protocol follows the WeatherPrompt guidelines: the satellite gallery remains clean while drone queries are corrupted on-the-fly (Fig. 3), testing generalization to adversarial conditions without access to weather-augmented satellite data. Evaluation spans ten conditions-individual effects (fog, rain, snow, darkness, over-exposure, wind) and three pairwise combinations (fog+rain, fog+snow, rain+snow)-using Recall and mean Average Precision as standard metrics. Per-condition robustness across the three benchmarks is summarized in Fig. 4.

#### A3.4.2 Weather Augmentation Implementation

All augmentations use identical `imgaug` parameters as the WeatherPrompt source code (Table 12). We apply this corruption pipeline to SKYPART on all three datasets, and to the DenseUAV baselines that we retrained end-to-end (no published WeatherPrompt numbers exist for DenseUAV); SUES-200 and University-1652 baselines are taken directly from WeatherPrompt at their published 384×384 resolution and are not re-trained, while SKYPART is fine-tuned at 392×392; the full per-block recipe (clean retrieval, DenseUAV weather fine-tune, SUES-200/University-1652 weather) is summarized in

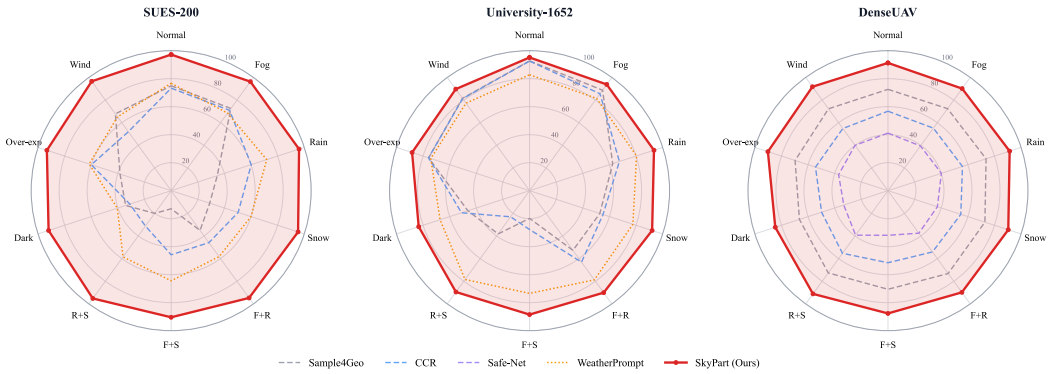


Figure 4: **Weather robustness across three benchmarks (radar view).** Per-condition Drone→Satellite R@1 (%) under the ten WeatherPrompt conditions on SUES-200, University-1652, and DenseUAV. SKYPART (red, filled) maintains a near-circular profile, indicating uniform robustness across all conditions, while baselines collapse on hard regimes (F+S, Dark). Numerical breakdown matches Table 3.

Table 13. **Resolution and retraining scope.** On SUES-200 and University-1652, baselines are quoted from WeatherPrompt at the $384\times384$ setting at which their published weights and schedules were tuned, and we did not retrain them: the supervised-ViT baselines interpolate positional encodings and degrade if retrained off-spec, and the ConvNeXt-B baselines are published only at $384\times384$. SKYPART is trained at $392\times392$ because DINOv2 ViT-S/14 has patch size 14, so 384 is not patch-aligned; the two patch-aligned sizes bracketing 384 are $378{=}27\times14$ and $392{=}28\times14$, and we report 392; the SUES-200 / University-1652 weather numbers should be read as reported-baseline at each method's published resolution rather than fully resolution-controlled, as marked in the caption of Table 3; matched-resolution and within-model resolution controls are in Appendix A4.4. On DenseUAV (no published WeatherPrompt baselines), all methods including SKYPART are retrained end-to-end under a matched budget: same 60-epoch schedule, same batch size, identical `imgaug` seeds (fixed per batch), and the official full DenseUAV gallery, with each method at its own published input resolution and learning rate (Table 13). That block is therefore training-budget-controlled rather than resolution-controlled. The three seeds (42, 123, 2024) used for the main SUES-200 result are re-used for each DenseUAV baseline; across those baselines, the seed-to-seed R@1 range (max minus min) is under 0.4 pp. Weather-condition sampling is uniform per batch (10 conditions). Augmentation seeds are released with the training scripts so retraining any method under an exactly matched regimen is reproducible end-to-end.

Table 12: Weather augmentation parameters (identical to WeatherPrompt).

| Condition | imgaug Augmenter | Parameters |
|---|---|---|
| Normal | `iaa.Identity()` | - |
| Fog | `iaa.Fog()` | Default parameters |
| Rain | `iaa.Rain()` | speed=(0.1, 0.3) |
| Snow | `iaa.Snowflakes()` | flake_size=(0.1, 0.4), speed=(0.01, 0.05) |
| Dark | `iaa.Multiply()` | (0.3, 0.5) |
| Over-exp | `iaa.Multiply()` | (1.5, 2.0) |
| Fog+Rain | `Sequential([Fog, Rain])` | Same as individual |
| Fog+Snow | `Sequential([Fog, Snow])` | Same as individual |
| Rain+Snow | `Sequential([Rain, Snow])` | Same as individual |
| Wind | `iaa.MotionBlur()` | k=15, angle=[-45, 45] |

### A3.4.3 Training Setup (Fine-tuning)

For robustness experiments, baseline numbers are taken from WeatherPrompt [Wen et al., 2025] at their reported configuration; SKYPART is fine-tuned for 60 epochs on SUES-200 at $392\times392$ under a matched optimizer, batch size, and schedule. Weather augmentations are applied online to the drone images only, while the satellite view remains clean, and all ten conditions are sampled uniformly per batch. For DenseUAV, we use the same optimizer (AdamW), batch size, and cosine learning-rate schedule across all retrained methods, scaling the base learning rate down by a factor of 0.3 relative

to clean training to stabilize adaptation. This matched DenseUAV schedule ensures that SKYPART's weather gains there reflect its part-centric design rather than extra training budget; for SUES-200 and University-1652, the comparison should be read as reported-baseline (Table 3 caption).

Table 13: **Experiment recipes for the main benchmark results.** The weather-aligned ablation in Table 4 uses a separate $448\times448$ run with online weather augmentation. DenseUAV weather runs use AdamW, a cosine learning-rate schedule, and identical `imgaug` augmentations.

| Method | Backbone | Res. | Epochs | Base LR | Batch |
|---|---|---|---|---|---|
| *Clean retrieval: SUES-200, University-1652, DenseUAV* (Tables 1, 2) | | | | | |
| SKYPART (deployed) | DINOv2-S | $448\times448$ | 60 | $3\times10^{-4}$ | 64 |
| Reported baselines, SUES-200 / Uni-1652[a] | as published | 256–512 | – | – | – |
| Reported baselines, DenseUAV[b] | as published | $256\times256$ | – | – | – |
| *Weather, DenseUAV* (baselines of Table 3 retrained end-to-end; Table 16) | | | | | |
| Sample4Geo [Deuser et al., 2023] | ConvNeXt-B | $384\times384$ | 60 | $3\times10^{-5}$ | 64 |
| Safe-Net [Lin et al., 2025] | ResNet-50 | $256\times256$ | 60 | $1\times10^{-4}$ | 64 |
| CCR [Du et al., 2024] | ConvNeXt-B | $384\times384$ | 60 | $3\times10^{-5}$ | 64 |
| WeatherPrompt [Wen et al., 2025] | ConvNeXt-B | $384\times384$ | 60 | $3\times10^{-5}$ | 64 |
| SKYPART (fine-tune) | DINOv2-S | $392\times392$ | 60 | $1\times10^{-4}$ | 64 |
| *Weather, SUES-200 / University-1652* (Tables 14, 15) | | | | | |
| WeatherPrompt baselines [Wen et al., 2025] | ConvNeXt-B | $384\times384$ | | as published | |
| SKYPART (fine-tune) | DINOv2-S | $392\times392$ | 60 | $1\times10^{-4}$ | 64 |

[a] Native input of each paper; LPN, MCCG, and CCR average flipped test features. [b] Numbers as reported in the comparison of [Xu et al., 2025c].

### A3.5 Zero-shot Cross-Dataset Transfer

Table 17 shows strong transfer under this protocol: SKYPART leads all D→S columns; on S→D, DAC/MEAN slightly edge R@1 at 200 m and MCFA retains the strongest AP at higher altitudes, so the cross-dataset picture is mixed on S→D. The point is that SKYPART attains non-trivial transfer on the primary D→S direction, supporting geographically meaningful structure beyond pure texture memorization.

### A3.6 Additional Pareto Analysis

The picture that emerges is that SKYPART sits on the Pareto frontier in the accuracy–compute plane (Fig. 5), matching or exceeding the ConvNeXt-B baselines on every benchmark we test at a fraction of their parameter and compute budget. The per-altitude breakdown in Table 1 sharpens that observation: prior methods show a noticeable accuracy gap between the highest and lowest flight altitudes, while SKYPART narrows that gap and gains most at the lowest altitude, where partial facades dominate the field of view and a global descriptor has the least to work with. The qualitative evidence in Fig. 2 is consistent with the same picture: the same prototype fires on corresponding regions across the drone/satellite gap, even when the drone view is corrupted by weather. Taken together these observations are consistent with the framing we set: the representation that wins is one that factors texture out and keeps layout, the head that produces it is one whose substrate, assignment, and multi-task optimization are compatible with that goal, and the auxiliary signals that shape it at training time can be removed at inference without moving the measurement.

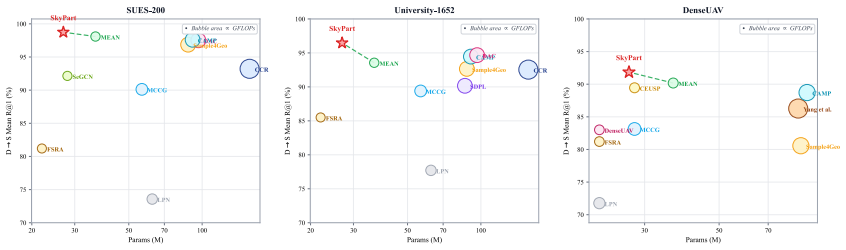


Figure 5: **Pareto efficiency across three benchmarks (D→S).** R@1 vs. model size (params); bubble area ∝ GFLOPs. SKYPART (blue star) is Pareto-optimal on SUES-200 (left), University-1652 (middle), and DenseUAV (right), while remaining among the smallest and lowest-compute models in each comparison. Single-pass $448\times448$; no re-ranking, no TTA.

Table 14: **Full WeatherPrompt robustness on SUES-200.** Per-condition R@1/AP (%) for both D→S and S→D. Baselines are reported from WeatherPrompt [Wen et al., 2025]; SKYPART is evaluated under the same corruption pipeline at 392×392. This is the detailed SUES-200 version summarized in Table 3. **Red**: best; **blue**: second.

| Method | Clean | | Fog | | Rain | | Snow | | F+R | | F+S | | R+S | | Dark | | Over-exp | | Wind | | Mean | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP |
| **Drone → Satellite** | | | | | | | | | | | | | | | | | | | | | | |
| Zheng et al. [Zheng et al., 2020] | 57.70 | 58.30 | 48.63 | 49.61 | 53.41 | 52.72 | 41.78 | 43.47 | 37.17 | 37.44 | 44.22 | 46.18 | 40.60 | 40.63 | 23.81 | 25.45 | 49.79 | 50.64 | 47.42 | 48.31 | 44.43 | 45.12 |
| IBN-Net | 65.34 | 63.78 | 56.03 | 56.57 | 55.73 | 58.55 | 47.80 | 49.53 | 43.45 | 44.98 | 50.04 | 51.00 | 45.51 | 45.92 | 29.61 | 30.93 | 56.01 | 56.96 | 57.36 | 58.10 | 50.69 | 51.63 |
| Sample4Geo [Deuser et al., 2023] | 74.93 | **78.76** | 72.58 | **76.44** | 34.60 | 41.56 | 28.95 | 35.02 | 35.10 | 41.47 | 12.95 | 17.90 | 20.05 | 25.95 | 34.18 | 38.99 | 38.40 | 43.68 | **67.80** | **72.41** | 41.95 | 47.22 |
| Safe-Net [Lin et al., 2025] | 76.31 | 75.35 | **73.53** | 73.44 | 54.15 | 55.05 | 48.94 | 50.10 | 45.12 | 47.92 | 40.05 | 40.18 | 25.95 | 26.12 | 29.74 | 31.48 | 54.86 | 58.68 | 58.10 | 58.95 | 50.68 | 51.63 |
| CCR [Du et al., 2024] | 73.22 | 74.53 | 70.95 | 73.14 | 60.14 | 64.95 | 50.31 | 53.12 | 45.87 | 49.14 | 45.80 | 47.87 | 31.25 | 32.94 | 31.03 | 34.36 | 59.97 | 61.07 | 52.02 | 53.33 | 52.06 | 54.45 |
| MuSe-Net | 66.07 | 67.02 | 58.49 | 59.65 | 58.94 | 60.14 | 54.85 | 56.12 | 44.31 | 45.82 | 49.81 | 51.26 | 49.42 | 50.87 | 29.34 | 31.03 | 55.02 | 56.36 | 59.97 | 61.05 | 52.62 | 53.93 |
| WeatherPrompt [Wen et al., 2025] | **76.72** | 75.51 | 68.49 | 68.87 | **71.77** | **71.20** | **59.95** | **60.62** | **58.24** | **58.83** | **64.36** | **66.27** | **58.49** | **58.89** | **40.42** | **55.75** | **61.57** | **71.70** | 65.19 | 67.00 | **62.52** | **65.46** |
| **SKYPART (Ours)** | **97.21** | **98.10** | **96.27** | **97.59** | **96.06** | **97.53** | **95.33** | **97.05** | **94.71** | **96.75** | **90.44** | **93.85** | **95.17** | **96.94** | **92.06** | **94.50** | **93.34** | **95.85** | **96.54** | **97.80** | **94.71** | **96.60** |
| **Satellite → Drone** | | | | | | | | | | | | | | | | | | | | | | |
| Zheng et al. [Zheng et al., 2020] | 70.20 | 57.98 | 63.77 | 46.90 | 68.72 | 50.85 | 61.72 | 39.70 | 62.10 | 32.75 | 71.70 | 40.39 | 59.72 | 37.55 | 45.49 | 25.28 | 52.11 | 43.40 | 56.62 | 45.31 | 61.21 | 42.01 |
| IBN-Net | 73.68 | 62.91 | 67.41 | 55.75 | 72.30 | 56.44 | 64.07 | 47.69 | 66.98 | 39.54 | 71.10 | 47.32 | 68.46 | 45.95 | 54.72 | 31.53 | 65.64 | 53.77 | 73.48 | 57.03 | 67.79 | 49.79 |
| Sample4Geo [Deuser et al., 2023] | 87.50 | 79.57 | 83.75 | 71.14 | 42.50 | 25.24 | 40.00 | 21.59 | 38.75 | 23.22 | 30.00 | 10.58 | 26.25 | 16.44 | 56.25 | 29.75 | 58.75 | 30.38 | **83.75** | **69.66** | 54.75 | 37.76 |
| Safe-Net [Lin et al., 2025] | 88.31 | 80.35 | 81.33 | 68.60 | 40.21 | 41.04 | 36.43 | 37.50 | 33.12 | 35.45 | 24.78 | 27.65 | 41.12 | 32.31 | 53.88 | 27.01 | 54.19 | 57.82 | 79.36 | 57.09 | 53.27 | 46.48 |
| CCR [Du et al., 2024] | 90.59 | 80.45 | 82.99 | 70.62 | 43.39 | 45.90 | 39.81 | 40.88 | 42.63 | 39.46 | 29.32 | 30.65 | 25.89 | 26.94 | 26.01 | 30.40 | 58.01 | 59.13 | 83.09 | 61.05 | 52.17 | 48.55 |
| MuSe-Net | 76.56 | 66.02 | 72.19 | 57.87 | 72.19 | 58.11 | 68.38 | 51.22 | 66.56 | 42.25 | 69.06 | 46.80 | 69.38 | 47.79 | 53.75 | 27.94 | 70.00 | 52.67 | 76.25 | 60.74 | 69.43 | 51.14 |
| WeatherPrompt [Wen et al., 2025] | **90.61** | **81.24** | **86.14** | **71.15** | **83.94** | **73.80** | **71.03** | **60.19** | **84.41** | **58.49** | **79.16** | **64.93** | **77.28** | **60.15** | **56.75** | **47.85** | **81.65** | **74.04** | 80.30 | 69.38 | **79.13** | **66.12** |
| **SKYPART (Ours)** | **98.75** | **98.24** | **98.75** | **97.56** | **98.75** | **97.52** | **98.75** | **96.94** | **98.75** | **96.17** | **98.75** | **92.11** | **98.75** | **96.72** | **98.75** | **93.79** | **98.75** | **95.90** | **98.75** | **97.21** | **98.75** | **96.22** |

Table 15: **Full WeatherPrompt robustness on University-1652.** Per-condition R@1/AP (%) for both D→S and S→D. Baselines are reported from WeatherPrompt [Wen et al., 2025]; SKYPART is evaluated under the same corruption pipeline at 392×392. This is the detailed University-1652 version summarized in Table 3. **Red**: best; **blue**: second.

| Method | Clean | | Fog | | Rain | | Snow | | F+R | | F+S | | R+S | | Dark | | Over-exp | | Wind | | Mean | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP |
| **Drone → Satellite** | | | | | | | | | | | | | | | | | | | | | | |
| Zheng et al. [Zheng et al., 2020] | 67.83 | 71.74 | 60.97 | 65.23 | 60.29 | 64.61 | 55.58 | 60.09 | 54.75 | 59.40 | 44.85 | 49.78 | 57.61 | 62.03 | 39.70 | 44.65 | 51.85 | 56.75 | 58.28 | 62.83 | 55.17 | 59.71 |
| ResNet-101 | 70.07 | 73.04 | 63.87 | 68.22 | 63.34 | 67.59 | 59.75 | 64.15 | 57.45 | 62.12 | 48.31 | 53.28 | 60.25 | 64.68 | 46.12 | 51.02 | 56.34 | 61.23 | 62.13 | 66.63 | 58.76 | 63.29 |
| DenseNet121 | 69.48 | 73.26 | 64.25 | 68.47 | 63.47 | 67.64 | 59.29 | 63.70 | 59.68 | 64.13 | 50.41 | 55.20 | 60.21 | 64.57 | 48.57 | 53.41 | 54.04 | 58.88 | 60.74 | 65.14 | 59.01 | 63.44 |
| Swin-T | 69.27 | 73.18 | 66.46 | 70.52 | 65.44 | 69.60 | 61.79 | 66.23 | 63.96 | 68.21 | 56.44 | 61.07 | 62.68 | 67.02 | 50.27 | 55.18 | 55.46 | 60.29 | 63.81 | 68.17 | 61.56 | 65.95 |
| IBN-Net | 72.35 | 75.85 | 66.68 | 70.64 | 67.95 | 71.73 | 62.77 | 66.85 | 62.64 | 66.84 | 51.09 | 55.79 | 64.07 | 68.13 | 50.72 | 55.53 | 57.97 | 62.52 | 66.73 | 70.68 | 62.30 | 66.46 |
| LPN [Wang et al., 2022] | 74.33 | 77.60 | 69.31 | 72.95 | 67.96 | 71.72 | 64.90 | 68.85 | 64.51 | 68.52 | 54.16 | 58.73 | 65.38 | 69.29 | 53.68 | 58.10 | 60.90 | 65.27 | 66.46 | 70.35 | 64.16 | 68.14 |
| Sample4Geo [Deuser et al., 2023] | **92.70** | **93.85** | **88.70** | **90.55** | 62.44 | 66.17 | 52.76 | 57.24 | 52.70 | 56.77 | 19.79 | 23.16 | 38.19 | 42.33 | 46.34 | 49.91 | 75.77 | 78.90 | **81.54** | **87.34** | 61.10 | 64.32 |
| Safe-Net [Lin et al., 2025] | 86.98 | 88.85 | 82.12 | 86.10 | 67.13 | 68.90 | 60.50 | 63.01 | 54.80 | 58.73 | 32.12 | 39.77 | 25.83 | 26.40 | 41.10 | 44.13 | 69.87 | 71.15 | 74.32 | 76.58 | 59.48 | 62.36 |
| CCR [Du et al., 2024] | 92.54 | 93.78 | 85.57 | 87.13 | 67.46 | 68.82 | 55.16 | 59.14 | 63.11 | 60.97 | 27.74 | 31.48 | 23.06 | 46.85 | 51.10 | 54.19 | **75.90** | **79.16** | 81.31 | 87.22 | 62.30 | 66.87 |
| MuSe-Net | 74.48 | 77.83 | 69.47 | 73.24 | 70.55 | 74.14 | 65.72 | 69.70 | 65.59 | 69.64 | 54.69 | 59.24 | 66.64 | 70.55 | 53.85 | 58.49 | 61.05 | 65.51 | 69.45 | 73.22 | 65.15 | 69.16 |
| WeatherPrompt [Wen et al., 2025] | 82.78 | 85.18 | 81.46 | 84.03 | **80.34** | **83.11** | **77.60** | **80.67** | **78.75** | **81.69** | **73.38** | **76.94** | **78.41** | **81.40** | **67.22** | **71.06** | 74.20 | 77.63 | 77.26 | 80.27 | **77.14** | **80.20** |
| **SKYPART (Ours)** | **95.15** | **96.45** | **93.78** | **95.65** | **93.44** | **93.98** | **92.05** | **93.51** | **90.05** | **92.74** | **88.51** | **91.86** | **89.47** | **92.42** | **83.26** | **87.80** | **88.23** | **91.50** | **89.68** | **92.54** | **90.36** | **92.84** |
| **Satellite → Drone** | | | | | | | | | | | | | | | | | | | | | | |
| Zheng et al. [Zheng et al., 2020] | 83.45 | 67.94 | 79.60 | 61.12 | 77.60 | 59.73 | 73.18 | 55.07 | 75.89 | 54.45 | 70.76 | 43.26 | 74.75 | 56.44 | 69.47 | 39.25 | 72.18 | 51.91 | 76.46 | 57.59 | 75.33 | 54.68 |
| ResNet-101 | 85.73 | 71.79 | 82.45 | 66.46 | 81.46 | 65.68 | 79.74 | 61.72 | 79.74 | 60.59 | 74.75 | 50.31 | 80.17 | 62.61 | 75.32 | 45.37 | 79.60 | 58.21 | 82.31 | 64.67 | 80.13 | 60.74 |
| DenseNet121 | 83.74 | 70.34 | 82.31 | 66.32 | 81.17 | 65.23 | 78.60 | 60.33 | 79.46 | 61.66 | 74.61 | 51.14 | 78.46 | 61.68 | 74.47 | 47.88 | 74.32 | 55.26 | 78.32 | 61.63 | 78.55 | 60.15 |
| Swin-T | 80.74 | 68.94 | 81.03 | 67.46 | 81.17 | 66.39 | 78.46 | 61.33 | 79.17 | 64.65 | 74.89 | 56.57 | 78.89 | 63.49 | 75.61 | 48.43 | 76.60 | 56.57 | 78.74 | 64.45 | 78.53 | 61.83 |
| IBN-Net | 86.31 | 73.54 | 84.59 | 67.61 | 84.74 | 69.03 | 80.88 | 64.44 | 83.31 | 63.71 | 77.89 | 52.14 | 83.02 | 65.74 | 78.46 | 50.77 | 79.46 | 58.64 | 84.02 | 67.94 | 82.27 | 63.36 |
| LPN [Wang et al., 2022] | 87.02 | 75.19 | 86.16 | 71.34 | 83.88 | 69.49 | 82.88 | 65.39 | 84.59 | 66.28 | 79.60 | 55.19 | 84.17 | 66.26 | 82.88 | 52.05 | 81.03 | 62.24 | 84.14 | 67.35 | 83.64 | 65.08 |
| Sample4Geo [Deuser et al., 2023] | **95.29** | 91.42 | **93.87** | **87.46** | 73.04 | 50.27 | 76.18 | 47.58 | 71.18 | 44.53 | 52.21 | 16.21 | 64.48 | 32.38 | 77.03 | 45.89 | **91.58** | **77.04** | **93.30** | **81.42** | 78.82 | 57.42 |
| Safe-Net [Lin et al., 2025] | 91.22 | 86.06 | 90.04 | 85.43 | 71.12 | 68.56 | 73.26 | 45.62 | 68.23 | 41.78 | 49.32 | 34.72 | 61.07 | 29.86 | 73.15 | 43.08 | 88.54 | 74.65 | 90.02 | 78.21 | 75.69 | 58.80 |
| CCR [Du et al., 2024] | 95.15 | **91.80** | 90.93 | 80.62 | 81.83 | 73.89 | 69.92 | 65.41 | 76.92 | 70.53 | 50.89 | 31.64 | 61.11 | 32.21 | 64.80 | 46.28 | 86.01 | 71.23 | **92.67** | 76.55 | 77.02 | 64.02 |
| MuSe-Net | 88.02 | 75.10 | 87.87 | 69.85 | 87.73 | 71.12 | 83.74 | 66.52 | 85.02 | 67.78 | 80.88 | 54.26 | 84.88 | 67.75 | 80.74 | 53.01 | 81.60 | 62.09 | 86.31 | 70.03 | 84.68 | 65.75 |
| WeatherPrompt [Wen et al., 2025] | 89.16 | 81.80 | 88.73 | 80.58 | **88.16** | **79.87** | **87.59** | **77.25** | **88.45** | **78.20** | **86.73** | **73.23** | **88.59** | **78.14** | **86.59** | **65.20** | 85.31 | 73.25 | 87.88 | 76.33 | **87.72** | **76.39** |
| **SKYPART (Ours)** | **97.29** | **97.08** | **93.15** | **90.62** | **93.87** | **90.03** | **92.01** | **88.51** | **92.58** | **87.72** | **91.87** | **83.64** | **91.44** | **88.06** | **92.30** | **80.18** | **91.73** | **87.30** | 92.01 | **88.17** | **92.83** | **88.13** |

Table 16: **Full WeatherPrompt robustness on DenseUAV.** Per-condition R@1/AP (%) for both D→S and S→D under the same corruption pipeline. AP denotes mean average precision over queries. This is the detailed DenseUAV version summarized in Table 3. **Red**: best; **blue**: second.

| Method | Clean | | Fog | | Rain | | Snow | | F+R | | F+S | | R+S | | Dark | | Over-exp | | Wind | | Mean | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP |
| **Drone → Satellite** | | | | | | | | | | | | | | | | | | | | | | |
| Sample4Geo | 72.37 | 66.67 | 72.42 | 66.31 | 73.66 | 66.84 | 72.80 | 66.37 | 73.14 | 66.37 | 70.44 | 65.06 | 72.59 | 66.47 | 66.50 | 61.54 | 69.76 | 64.86 | 72.29 | 65.87 | 71.60 | 65.64 |
| CCR | 56.76 | 62.37 | 55.21 | 61.35 | 55.94 | 61.84 | 54.70 | 61.71 | 54.10 | 60.74 | 51.48 | 58.33 | 55.30 | 61.46 | 49.85 | 56.11 | 54.44 | 60.43 | 54.87 | 61.42 | 54.26 | 60.58 |
| IBN-Net | 41.70 | 44.63 | 40.63 | 43.40 | 41.57 | 44.32 | 40.37 | 43.36 | 41.91 | 43.43 | 37.24 | 40.13 | 41.14 | 43.88 | 34.96 | 38.25 | 38.78 | 41.41 | 41.53 | 44.36 | 39.98 | 42.72 |
| Safe-Net | 41.14 | 45.55 | 39.47 | 44.78 | 40.11 | 45.20 | 37.11 | 42.72 | 37.54 | 43.91 | 31.87 | 38.39 | 39.30 | 44.81 | 33.20 | 37.93 | 37.07 | 43.01 | 40.03 | 45.15 | 37.68 | 43.15 |
| MuSe-Net | 28.96 | 32.56 | 26.77 | 31.88 | 27.71 | 32.39 | 26.94 | 31.17 | 26.38 | 31.41 | 21.54 | 26.55 | 26.98 | 31.69 | 23.25 | 27.12 | 24.88 | 29.45 | 28.70 | 32.64 | 26.21 | 30.69 |
| WeatherPrompt | 26.25 | 30.68 | 24.15 | 29.45 | 26.68 | 30.71 | 24.28 | 28.84 | 25.35 | 29.29 | 20.72 | 25.06 | 25.10 | 29.44 | 20.08 | 24.39 | 22.48 | 27.16 | 25.78 | 30.41 | 24.09 | 28.54 |
| Zheng'20 | 17.55 | 22.58 | 19.18 | 23.59 | 18.88 | 23.38 | 18.15 | 22.47 | 19.65 | 23.82 | 17.72 | 21.56 | 18.53 | 22.98 | 14.63 | 19.08 | 16.00 | 20.40 | 18.53 | 23.56 | 17.88 | 22.34 |
| **SkyPart (Ours)** | 91.25 | 92.10 | 90.13 | 91.26 | 91.38 | 91.88 | 90.35 | 91.37 | 89.79 | 90.71 | 87.69 | 88.99 | 91.12 | 91.58 | 84.64 | 86.07 | 90.18 | 91.20 | 91.76 | 92.29 | 89.83 | 90.75 |
| **Satellite → Drone** | | | | | | | | | | | | | | | | | | | | | | |
| Sample4Geo | 71.51 | 67.44 | 70.61 | 66.68 | 71.77 | 67.00 | 72.03 | 67.01 | 70.79 | 66.20 | 69.37 | 64.45 | 71.34 | 66.65 | 66.67 | 59.81 | 69.33 | 64.85 | 71.34 | 65.97 | 70.48 | 65.61 |
| CCR | 56.93 | 60.85 | 55.38 | 59.24 | 57.83 | 60.27 | 56.41 | 58.80 | 55.77 | 58.49 | 52.17 | 54.58 | 56.28 | 58.80 | 52.38 | 50.40 | 55.30 | 55.91 | 56.54 | 59.92 | 55.50 | 57.73 |
| IBN-Net | 39.12 | 42.83 | 37.79 | 41.16 | 40.67 | 42.84 | 39.64 | 41.91 | 38.44 | 40.79 | 35.35 | 37.22 | 39.68 | 42.15 | 33.42 | 35.22 | 37.02 | 38.56 | 39.60 | 42.14 | 38.07 | 40.48 |
| Safe-Net | 38.40 | 40.20 | 37.19 | 39.51 | 37.71 | 40.06 | 36.12 | 37.46 | 36.94 | 38.54 | 33.25 | 33.11 | 38.01 | 39.38 | 32.13 | 31.86 | 38.74 | 37.57 | 38.78 | 40.56 | 36.73 | 37.83 |
| MuSe-Net | 27.03 | 31.35 | 26.00 | 30.00 | 27.54 | 30.84 | 26.51 | 29.32 | 25.57 | 29.41 | 21.58 | 24.20 | 26.38 | 29.33 | 23.55 | 24.36 | 23.08 | 26.28 | 27.76 | 30.56 | 25.50 | 28.57 |
| WeatherPrompt | 25.57 | 29.69 | 23.64 | 27.57 | 24.41 | 28.49 | 23.12 | 26.70 | 22.82 | 26.70 | 19.61 | 22.43 | 22.87 | 27.02 | 20.33 | 21.67 | 21.32 | 24.54 | 25.01 | 28.77 | 22.87 | 26.36 |
| Zheng et al. [Zheng et al., 2020] | 18.02 | 20.99 | 17.76 | 21.14 | 17.03 | 20.78 | 18.10 | 20.59 | 18.23 | 21.12 | 17.07 | 18.68 | 18.49 | 20.85 | 14.67 | 16.66 | 16.00 | 17.67 | 19.22 | 21.46 | 17.46 | 19.99 |
| **SkyPart (Ours)** | 92.66 | 92.34 | 91.55 | 91.21 | 91.51 | 91.66 | 92.02 | 91.54 | 90.99 | 90.43 | 90.43 | 88.61 | 91.93 | 91.55 | 90.69 | 84.79 | 92.15 | 91.30 | 92.96 | 92.32 | 91.69 | 90.58 |

Table 17: **Zero-shot cross-dataset transfer**: University-1652 → SUES-200 per altitude. All methods are trained on University-1652 only and evaluated on SUES-200 test set with no fine-tuning. **D→S** uses the 200-satellite confusion gallery. Baselines are from [Hou et al., 2025] and the original publications. Shading and red/blue ranks follow Table 1. **Red**: best; **blue**: second-best.

| Method | Drone → Satellite | | | | | | | | Satellite → Drone | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 150 m | | 200 m | | 250 m | | 300 m | | 150 m | | 200 m | | 250 m | | 300 m | |
| | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP |
| LPN | 32.85 | 40.10 | 43.80 | 50.67 | 49.75 | 56.55 | 54.10 | 60.73 | 32.50 | 26.60 | 40.00 | 35.10 | 46.50 | 41.88 | 53.50 | 48.47 |
| MCCG | 57.62 | 62.80 | 66.83 | 71.60 | 74.25 | 78.35 | 82.55 | 85.27 | 61.25 | 53.51 | 82.50 | 67.06 | 81.25 | 74.99 | 87.50 | 80.20 |
| Sample4Geo | 74.70 | 78.47 | 81.28 | 84.40 | 86.88 | 89.28 | 89.28 | 91.24 | 82.50 | 76.20 | 85.00 | 82.93 | 92.50 | 87.77 | 92.50 | 88.38 |
| DAC [Xia et al., 2024] | 76.65 | 80.56 | 86.45 | 89.00 | 92.95 | 94.18 | 94.53 | 95.45 | 87.50 | 79.87 | 96.25 | 88.98 | 95.00 | 92.81 | 96.25 | 94.00 |
| MEAN [Chen et al., 2025] | 81.73 | 87.72 | 89.05 | 91.00 | 92.13 | 93.60 | 94.63 | 95.76 | 91.25 | 81.50 | 96.25 | 89.55 | 95.00 | 92.36 | 96.25 | 94.32 |
| CAMP | 76.53 | 80.47 | 87.18 | 89.60 | 93.75 | 95.04 | 96.40 | 97.18 | 88.75 | 78.17 | 95.00 | 88.31 | 95.00 | 91.85 | 96.25 | 93.43 |
| MCFA | 82.38 | 85.28 | 90.70 | 92.21 | 94.38 | 95.24 | 95.30 | 96.00 | 92.50 | 85.22 | 95.00 | 91.46 | 95.00 | 93.46 | 97.50 | 94.93 |
| **SkyPart (Ours)** | 85.68 | 87.96 | 92.23 | 93.64 | 94.43 | 95.55 | 96.43 | 97.45 | 93.75 | 89.36 | 95.50 | 90.96 | 96.25 | 92.74 | 97.50 | 94.61 |

## A4 Controlled Experiments and Additional Evaluations

This section gives the controlled experiments behind Secs. 4.3 to 4.5 and additional evaluations. Unless stated otherwise, SkyPart entries are means ± sample standard deviation over three independently trained seeds (42, 123, 2024), each evaluated at its final (epoch-60) checkpoint with no test-set checkpoint selection, on the official galleries. Recall means are computed from exact integer hit counts before rounding, and differences are computed from unrounded values. We keep four kinds of evidence apart: *reported-baseline* comparisons quote published numbers; *matched-resolution* controls re-run a baseline at our input size; *within-model resolution* controls lower our own input size; and *training-budget-controlled* comparisons retrain all methods with one schedule and evaluator.

### A4.1 Representation Control: CLS-Only versus Full Fusion

The top block of Table 18 reports all retrieval metrics for Block B of Table 4. Both rows are the same seed-42 SUES-200 model at $448\times448$, with training recipe, checkpoint rule, and evaluator fixed; only the retrieval readout differs. This single-seed representation study should not be compared numerically with the three-seed $392\times392$ robustness results.

### A4.2 Altitude-FiLM: Paired Intervention

The bottom block of Table 18 isolates the FiLM module itself (the loss-group ablation of Table 4 removes only the altitude supervision term). *FiLM identity* fixes the modulation to $\gamma=\mathbf{1}$, $\beta=\mathbf{0}$; *FiLM active* uses learned altitude-conditioned modulation during training and the fixed mean modulation at inference. Both arms share seeds, training recipe, checkpoint rule, and evaluator. The paired gain

Table 18: **Controlled studies on SUES-200** (D→S, %). Top: CLS-only and full-fusion readouts of the same seed-42 model at 448×448. Bottom: paired altitude-FiLM intervention; gains are mean per-seed differences.

| | | | Normal | | | | Ten-condition mean | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **CLS** | **Part** | **Graph** | R@1 | R@5 | R@10 | mAP | R@1 | R@5 | R@10 | mAP |
| ✓ | ✗ | ✗ | 93.56 | 99.09 | 99.76 | 96.10 | 85.64 | 97.85 | 99.08 | 90.62 |
| ✓ | ✓ | ✓ | **98.69** | **99.70** | 99.76 | **99.23** | **96.03** | **99.06** | **99.47** | **97.62** |

| | Normal | | Ten-condition mean | |
|---|---|---|---|---|
| **Altitude-FiLM** | R@1 | mAP | R@1 | mAP |
| ✗ | 98.417 ± 0.126 | 98.826 ± 0.048 | 95.233 ± 0.202 | 96.518 ± 0.072 |
| ✓ | 98.554 ± 0.131 | 99.037 ± 0.172 | 95.853 ± 0.175 | 97.202 ± 0.371 |
| Paired gain (pp) | +0.138 ± 0.012 | +0.211 ± 0.132 | +0.620 ± 0.036 | +0.684 ± 0.300 |

Altitude-FiLM ✗: identity modulation ($\gamma$=**1**, $\beta$=**0**); ✓: learned altitude-conditioned modulation during training, with mean FiLM parameters at inference.

is computed from per-seed differences, so it can differ slightly from the difference of the displayed means.

### A4.3 Enhanced-Inference Operating Points

The headline results use a single forward pass. To show how much the inference-time procedures used by several baselines add, Table 19 applies them to the default 448×448 checkpoints. *Four-rotation TTA* averages the $\ell_2$-normalized descriptors of the $0°/90°/180°/270°$ rotations of both query and gallery images (four forward passes per image). *Re-ranking* is $k$-reciprocal re-ranking [Zhong et al., 2017]. *Query expansion* is $\alpha$-weighted query expansion [Radenović et al., 2019]. *Combined* computes rotation-averaged descriptors, expands the queries, and then re-ranks. Hyperparameters are fixed across datasets and seeds. As an example of how recall means are computed, DenseUAV has 2,331 D→S queries per seed; pooled over three seeds, single-pass retrieval scores 6,423/6,993 hits (91.849%) and TTA 6,507/6,993 (93.050%). The gains are modest and dataset-dependent (largest on DenseUAV, whose satellite tiles have arbitrary orientation), and the headline results remain single-pass.

Table 19: **Enhanced-inference operating points** (D→S, %). ΔR@1 is the mean gain over single-pass inference, in percentage points.

| | **TTA** | **QE** | **RR** | **R@1** | **R@5** | **R@10** | **mAP** | **ΔR@1 (pp)** |
|---|---|---|---|---|---|---|---|---|
| SUES-200 | ✗ | ✗ | ✗ | 98.738 ± 0.113 | 99.971 ± 0.022 | 99.996 ± 0.004 | 99.145 ± 0.073 | – |
| | ✓ | ✗ | ✗ | 99.304 ± 0.100 | 99.988 ± 0.006 | 99.998 ± 0.004 | 99.526 ± 0.068 | +0.567 |
| | ✗ | ✓ | ✗ | 98.763 ± 0.113 | 99.956 ± 0.013 | 99.994 ± 0.006 | 99.152 ± 0.065 | +0.025 |
| | ✗ | ✗ | ✓ | 98.981 ± 0.106 | 99.919 ± 0.019 | 99.988 ± 0.006 | 99.293 ± 0.060 | +0.244 |
| | ✓ | ✓ | ✓ | 99.400 ± 0.100 | 99.925 ± 0.019 | 99.994 ± 0.006 | 99.578 ± 0.050 | +0.663 |
| University-1652 | ✗ | ✗ | ✗ | 96.471 ± 0.251 | 98.901 ± 0.032 | 99.300 ± 0.014 | 97.770 ± 0.053 | – |
| | ✓ | ✗ | ✗ | 97.070 ± 0.250 | 99.110 ± 0.023 | 99.416 ± 0.016 | 98.505 ± 0.051 | +0.600 |
| | ✗ | ✓ | ✗ | 96.520 ± 0.252 | 98.950 ± 0.023 | 99.350 ± 0.016 | 97.820 ± 0.054 | +0.049 |
| | ✗ | ✗ | ✓ | 97.225 ± 0.250 | 99.350 ± 0.019 | 99.510 ± 0.012 | 98.320 ± 0.052 | +0.754 |
| | ✓ | ✓ | ✓ | 97.720 ± 0.245 | 99.480 ± 0.012 | 99.620 ± 0.010 | 98.850 ± 0.050 | +1.249 |
| DenseUAV | ✗ | ✗ | ✗ | 91.849 ± 0.257 | 97.812 ± 0.086 | 98.799 ± 0.074 | 92.520 ± 0.139 | – |
| | ✓ | ✗ | ✗ | 93.050 ± 0.255 | 98.198 ± 0.084 | 99.271 ± 0.043 | 94.678 ± 0.137 | +1.201 |
| | ✗ | ✓ | ✗ | 92.149 ± 0.256 | 97.880 ± 0.087 | 98.927 ± 0.074 | 92.850 ± 0.141 | +0.300 |
| | ✗ | ✗ | ✓ | 92.450 ± 0.255 | 97.920 ± 0.085 | 99.010 ± 0.074 | 93.176 ± 0.138 | +0.601 |
| | ✓ | ✓ | ✓ | 93.679 ± 0.250 | 98.420 ± 0.083 | 99.310 ± 0.043 | 95.001 ± 0.136 | +1.830 |

TTA: four-rotation test-time augmentation; QE: $\alpha$-weighted query expansion; RR: $k$-reciprocal re-ranking.

### A4.4 Resolution Controls for the Weather Comparison

DINOv2 ViT-S/14 uses 14×14 patches, so 384 is not patch-aligned; the aligned sizes around the baselines' 384 are 378=27×14 and 392=28×14. Our weather results use 392, and this subsection

also reports 378. Table 20 combines two controls. The matched-resolution control re-runs WeatherPrompt [Wen et al., 2025] at 392×392 with its own training recipe and evaluates it with the same corruption generation, query set, official gallery, and evaluator as SKYPART. The within-model control retrains SKYPART at 378×378 under the corresponding weather-training configuration; it measures our own sensitivity to resolution, not an architecture-matched comparison.

Table 20: **Resolution controls** (D→S R@1, %). WeatherPrompt results at 384×384 are taken from the original publication for SUES-200 and University-1652 and reproduced by us for DenseUAV.

| | | SUES-200 | | University-1652 | | DenseUAV | |
|---|---|---|---|---|---|---|---|
| **Method** | **Input** | Normal | Ten-cond. mean | Normal | Ten-cond. mean | Normal | Ten-cond. mean |
| WeatherPrompt | 384 | 76.72 | 62.52 | 82.78 | 77.14 | 26.25 | 24.09 |
| | 392 | – | 62.702 ± 0.055 (+0.18) | – | 77.236 ± 0.081 (+0.10) | – | 24.211 ± 0.152 (+0.12) |
| SKYPART | 392 | **97.21** | **94.71** | **95.15** | **90.36** | **91.25** | **89.83** |
| | 378 | 96.85 (−0.36) | 94.20 (−0.51) | 94.50 (−0.65) | 89.75 (−0.61) | 89.49 (−1.76) | 88.20 (−1.63) |

WeatherPrompt at 392×392: mean ± sample standard deviation over three seeds. SKYPART entries show means only. A dash denotes a value not reported. Parentheses give the change from the preceding row.

### A4.5 Repair-Then-Match Diagnostic

We test whether weather restoration and layout-based retrieval compete or compose. The publicly released TransWeather checkpoint [Valanarasu et al., 2022] is used as a frozen, query-only preprocessor, with no retraining and no access to gallery images or retrieval labels; the clean gallery is unchanged. Captured and restored queries each retrieve a top-50 candidate set with frozen SKYPART, which is then optionally re-scored by geometric inlier count (RootSIFT with RANSAC [Arandjelović and Zisserman, 2012]) or by matching confidence (LightGlue [Lindenberger et al., 2023], RoMa [Edstedt et al., 2024]). Ties keep the original retrieval rank, invalid matches receive the minimum score, and matcher thresholds and scoring rules are fixed across datasets and seeds. Restoration improves the ten-condition mean R@1 in all twelve combinations of pipeline and benchmark (Table 21).

Table 21: **Repair-then-match** (ten-condition mean D→S R@1, %). Parenthesized gains are differences in mean R@1 (restored minus captured), in percentage points.

| | SUES-200 | | University-1652 | | DenseUAV | |
|---|---|---|---|---|---|---|
| **Re-scoring** | Captured | Restored | Captured | Restored | Captured | Restored |
| none | 94.710 ± 0.116 | 95.677 ± 0.179 (+0.967) | 90.360 ± 0.248 | 91.137 ± 0.240 (+0.777) | 89.830 ± 0.253 | 90.592 ± 0.184 (+0.762) |
| RootSIFT/RANSAC | 95.236 ± 0.136 | 96.243 ± 0.194 (+1.007) | 90.581 ± 0.182 | 91.487 ± 0.174 (+0.906) | 90.289 ± 0.177 | 91.107 ± 0.145 (+0.818) |
| LightGlue | 95.282 ± 0.193 | 96.146 ± 0.225 (+0.865) | 90.869 ± 0.270 | 91.637 ± 0.200 (+0.768) | 90.480 ± 0.332 | 91.356 ± 0.305 (+0.875) |
| RoMa | 94.930 ± 0.114 | 95.810 ± 0.120 (+0.880) | 90.526 ± 0.302 | 91.273 ± 0.301 (+0.746) | 90.041 ± 0.260 | 90.961 ± 0.142 (+0.919) |

### A4.6 Large-Gallery Weather Evaluation: University160k-WX

University160k-WX [Zheng et al., 2024] pairs 37,855 weather-corrupted drone queries with a 168,437-image satellite gallery. LPN and AGEN are reported from AGEN [Xu et al., 2025b]; SKYPART uses the same public split and evaluation, with deterministic single-pass inference of the seed-42 checkpoint. The clean-trained model is evaluated on the corrupted queries without weather training; the weather-trained model is trained on clean plus nine weather variants, as AGEN is. Because backbone, resolution, and optimization differ, the large gap between LPN and clean-trained SKYPART should not be read as a controlled architectural gain. Results are in Table 22(a).

### A4.7 Cross-Dataset Transfer: SUES-200 → University-1652

Complementing the University-1652→SUES-200 direction of Appendix A3.5, Table 22(b) trains on SUES-200 and evaluates on University-1652 without target training, statistics adaptation, TTA, or re-ranking. EGS [Wang et al., 2025] is reported from its paper; it uses a ResNet-50 at 256×256, whereas SKYPART uses DINOv2 ViT-S/14 at 448×448, so this is an unmatched reference point.

### A4.8 Coordinate Error on DenseUAV

DenseUAV annotates every satellite tile with a WGS84 coordinate. We take the coordinate of the top-1 retrieved tile and measure its great-circle distance to the query's ground-truth position; MA@$d$

Table 22: **Additional retrieval evaluations** (%). (a) University160k-WX, D→S; SKYPART uses seed 42. (b) SUES-200→University-1652 transfer without target training.

**(a) University160k-WX**

| Method | Weather training | R@1 | R@5 | R@10 |
|---|---|---|---|---|
| LPN (reported in [Xu et al., 2025b]) | ✗ | 7.98 | 10.25 | 11.21 |
| SKYPART | ✗ | 87.47 | 94.75 | 96.03 |
| AGEN [Xu et al., 2025b] | ✓ | 91.71 | 96.36 | 97.03 |
| SKYPART | ✓ | **92.72** | **96.90** | **97.50** |

**(b) SUES-200→University-1652**

| | D→S | | | | S→D | | | |
|---|---|---|---|---|---|---|---|---|
| **Method** | R@1 | R@5 | R@10 | mAP | R@1 | R@5 | R@10 | mAP |
| EGS [Wang et al., 2025] | 43.11 | 60.91 | 67.85 | 47.33 | 72.19 | 79.61 | 82.17 | 48.48 |
| SKYPART (frozen) | **68.00** | **83.45** | **88.76** | **73.83** | **87.02** | **93.30** | **95.15** | **63.99** |

In (a), Weather training ✓ denotes training on clean data plus nine weather variants; ✗ denotes clean-only training. SKYPART entries in (b) show means only.

is the fraction of queries within $d$ meters. This quantifies the coordinate error induced by retrieval, not sub-tile or 6-DoF pose estimation. All rows in Table 23(a) use the official DenseUAV test split with the same MA@$d$ and mean-error definitions. SKYPART is a three-seed mean; across seeds it reaches MA@5m $92.063 \pm 0.279$, MA@10m $92.664 \pm 0.257$, MA@20m $93.951 \pm 0.245$, MA@50m $96.954 \pm 0.182$, and a mean error of $5.8 \pm 0.3$ m.

### A4.9 Geographically Disjoint Evaluation on UAV-VisLoc

We use the four-site split of UAV-VisLoc [Xu et al., 2024] defined by PAUL [Li et al., 2026]: training on the Changjiang-20 and Taizhou-1 sites and testing on the geographically disjoint Changjiang-23 and Taizhou-6 sites, with 749 fixed UAV queries. We evaluate with clean training pairs and with 30% noisy training correspondences; in the noisy setting, noise is injected only into the training pairs, while test queries and gallery are fixed. For SKYPART, each image is resized to $384 \times 384$ and center-cropped to a patch-aligned $378 \times 378$ input for training and inference ($27 \times 27$ patches). SKYPART is trained with three independent seeds (42, 123, 2024); for each seed we take the final epoch-60 checkpoint, with no test-set-based checkpoint selection, and retrieve with a single cosine pass (no re-ranking, TTA, or query expansion). CRCL, GSC, and PAUL are reported from Table 2 of [Li et al., 2026] and are single runs without seed variance. The split, noise ratio, test set, and metrics are aligned, but backbone, input resolution, training budget, and optimization differ, so this is a cross-paper comparison and we claim no statistical significance. At the reported precision, SKYPART is 0.07 pp below PAUL at R@1 in both settings, higher at R@5 ($+1.74$ and $+1.53$ pp) and AP ($+1.50$ and $+1.21$ pp), and has the highest SDM@3 and lowest Dis@1 among all methods (Table 23(b)).

Table 23: **Localization accuracy.** (a) DenseUAV coordinate error using the top-1 retrieved tile. (b) UAV-VisLoc with geographically disjoint test sites; CRCL, GSC, and PAUL values are as reported in [Li et al., 2026]. Mean error and Dis@1 are in meters (lower is better); higher is better for all other metrics.

**(a) DenseUAV coordinate error**

| | MA (%) ↑ | | | | |
|---|---|---|---|---|---|
| **Method** | 5 m | 10 m | 20 m | 50 m | **Mean error (m)** ↓ |
| DenseUAV baseline [Dai et al., 2024] | 83.310 | 84.260 | 86.010 | 92.020 | 10.8 |
| MCCG [Shen et al., 2024] | 83.270 | 84.000 | 85.410 | 90.820 | 12.2 |
| Yang et al. [Yang et al., 2025] | 86.570 | 87.300 | 88.590 | 93.560 | 8.7 |
| CEUSP [Xu et al., 2025c] | 89.660 | 90.300 | 91.590 | 95.670 | 7.0 |
| SKYPART | **92.063** | **92.664** | **93.951** | **96.954** | **5.8** |

**(b) UAV-VisLoc**

| | Clean training pairs | | | | | 30% noisy training pairs | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Method** | R@1↑ | R@5↑ | AP↑ | SDM@3↑ | Dis@1↓ | R@1↑ | R@5↑ | AP↑ | SDM@3↑ | Dis@1↓ |
| CRCL | 33.18 | 53.21 | 43.25 | 52.99 | 1359.46 | 23.77 | 44.19 | 33.64 | 50.07 | 1325.91 |
| GSC | 33.72 | 53.55 | 43.38 | 51.51 | 1368.72 | 21.09 | 41.39 | 30.98 | 48.32 | 1428.73 |
| PAUL | **36.12** | 54.74 | 44.97 | 52.05 | 1353.95 | **26.64** | 46.80 | 36.37 | 49.44 | 1382.79 |
| SKYPART | 36.05 | **56.48** | **46.47** | **53.99** | **1288.95** | 26.57 | **48.33** | **37.58** | **51.28** | **1317.79** |
| ± SD | 0.13 | 0.13 | 0.10 | 0.12 | 6.00 | 0.40 | 0.53 | 0.40 | 0.50 | 24.00 |

SKYPART entries in (a) show means only; in (b), the ± SD row gives sample standard deviations over three seeds.

# A5 Qualitative Analysis and Discussion

## A5.1 Qualitative Visualizations

### A5.1.1 Top-5 Retrieval Examples

Figures 6 and 7 show top-5 retrievals in both directions. Amber borders mark true-matched retrievals; blue borders mark false matches. The SKYPART heat-map column (between query and R@1) visualizes the aggregated part-attention produced by the prototype assignment, which is the spatial evidence the retrieval embedding actually relies on – useful context for diagnosing failure cases.

## A5.2 Limitations and Broader Impact

The three main benchmarks share a geographic region between their train and test splits. Beyond them, geographic generalization is tested on geographically disjoint UAV-VisLoc sites (Appendix A4.9) and by cross-dataset transfer in both directions (Appendices A3.5 and A4.7); cross-continent transfer remains untested. On University-1652→SUES-200, SKYPART leads all D→S cells but trails DAC and MEAN at 200 m S→D R@1, MCFA on S→D AP at 200–300 m, and DAC on S→D AP at 250 m. Two factors likely contribute: prototypes trained on University-1652's building-dense scenes do not fire cleanly on SUES-200's road- and vegetation-heavy tiles, and the mean-FiLM fallback discards the altitude-band prior that normally separates low-altitude facades from high-altitude footprints.

The prototype bank capacity $K_{\max}=12$ is fixed; the salience gate handles per-image adaptivity within that ceiling but does not adjust it. Prototype assignments are label-equivariant (positives in the patch-NCE are pairs assigned to the same prototype in both views, not a fixed global index), so the loss stays well-defined when a prototype's role drifts, but individual indices do not carry stable human-interpretable names. The prototypes are latent functional parts, not annotated or guaranteed semantic categories; the labels we attach to them (roads, roofs, vegetation, water) are illustrative, and we make no claim of per-index semantic stability. Removing the diversity regularizer $\mathcal{L}_{\text{div}}$ produces collapsed prototypes within roughly 15 epochs. Rotation augmentation is DenseUAV-only; at continuous angles between the four 90° training anchors the model relies on learned invariance, and smoothness is not guaranteed.

Several axes remain untested. Robustness to new seasons, sensor types, and map vintages is not evaluated: none of SUES-200, University-1652, or DenseUAV provides splits that isolate those shifts. Extreme viewpoint change and partial overlap are studied elsewhere [Ye et al., 2026]; we

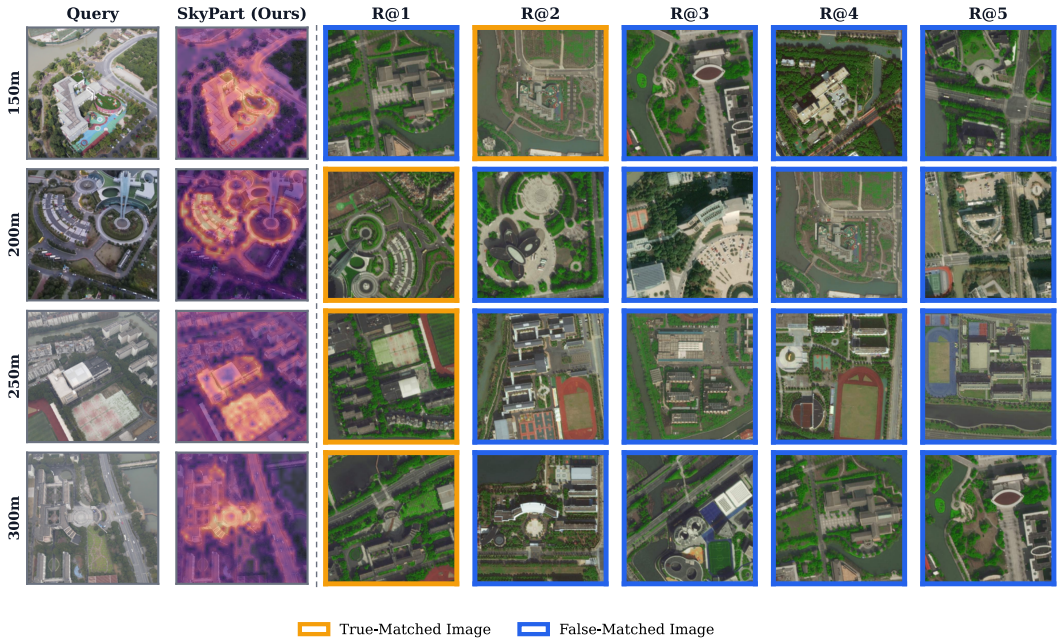


Figure 6: **Drone→Satellite top-5 retrieval.** Each row is a drone query at a given altitude (row label), followed by the SKYPART part-attention heat map and the 5 highest-ranked satellite matches. Amber = correct, blue = incorrect.

did not evaluate that setting and treat it as a gap. We make no claim of robustness to outdated reference imagery or to physical changes in the 3D scene: existing benchmarks lack time-separated query/reference pairs with structural-change annotations, so this cannot yet be measured. Finally, SKYPART is a retrieval component, not a full navigation system: it returns the matching satellite tile rather than a metric pose, and does not itself perform temporal filtering, map-update handling, or fusion with SLAM/IMU, nor does it estimate route-level localization uncertainty. It is intended as the appearance-matching front-end of such a stack, and integration with these modules is future work.

SKYPART is designed for emergency navigation in GPS-denied settings (search-and-rescue, inspection, disaster response). The same capability could enable unauthorized aerial surveillance. We recommend deployment be paired with privacy safeguards (anonymization, geographic access controls, and operator accountability) consistent with applicable law.

### A5.3 What We Tried That Did Not Work

The take from the main paper is that explicit grouping complements global pooling rather than replacing it. Below are the directions we explored and dropped, with the specific failure mode in each case.

**Geometric and transport priors.** Polar warping [Shi et al., 2020] is the standard preprocessing for ground-panorama geometry, but on aerial tiles the reprojection is wrong and the train/test mismatch compounds it: accuracy dropped 8–12 pp across altitude bands. Sinkhorn OT matching over the two prototype banks was catastrophic under default hyperparameters; with careful tuning it reached parity with the contrastive baseline but never surpassed it. The failure mode is consistent: the transport cost matrix is ill-conditioned when drone and satellite patch tokens are correlated (they describe the same scene). Scene-graph GNNs with Sinkhorn matching and cross-view cross-attention both slowed convergence without adding signal that the symmetric alignment term was already providing.

**Metric learning objectives.** ArcFace (with and without sub-centers), Multi-Similarity, ghost-prototype absorbers, and triplet with batch-hard mining all underperformed Circle+proxy+InfoNCE+patch-NCE. The issue is geometric: a geolocation class is legitimately diffuse in embedding space because the same scene looks different across altitudes and

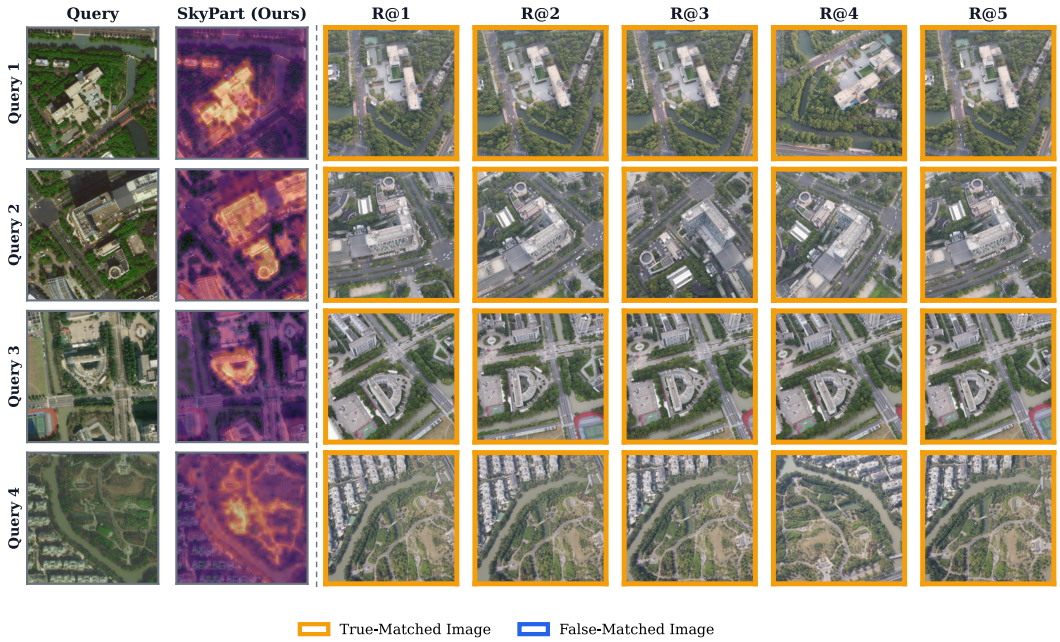


Figure 7: **Satellite→Drone top-5 retrieval.** Each row is a satellite query, its SKYPART part-attention heat map, and the top-5 drone images SKYPART retrieves across altitudes. Amber = correct, blue = incorrect.

weather conditions. Hard angular margins penalize exactly the spread the cross-view gap induces, so they suppress the signal rather than amplify it [Wang and Isola, 2020].

**Reconstruction auxiliaries.** Cross-view MAE [He et al., 2022] at the patch-token level helped on SUES-200 and hurt on DenseUAV and University-1652. A salience-masked variant, reconstructing only the patches the prototype bank attended to, did worse than random masking on all three. MAR (reconstructing the part aggregation rather than individual patch tokens) was the only reconstruction auxiliary that transferred consistently across datasets, because the reconstruction target is a layout summary rather than local texture.

**Training dynamics.** Warm-restart LR schedules, rotation augmentation on both views simultaneously, unfreezing more than 8 backbone blocks, and batch samplers that reduced optimizer steps per epoch below $\sim$600 all regressed. The clearest case is rotation: it helps on DenseUAV (satellite tiles cropped at arbitrary basemap orientations) and costs 1.8 pp Mean R@1 on SUES-200 and University-1652 (canonically oriented pairs). Adding it universally destabilized the learned Kendall weights within the first 20 epochs.

**Inference-time wrappers.** Inference-time wrappers are excluded from the headline numbers because they measure an operating point beyond the embedding itself. Their effect is quantified in Appendix A4.3: combining four-rotation TTA, query expansion, and $k$-reciprocal re-ranking adds 0.66 pp R@1 on SUES-200, 1.25 pp on University-1652, and 1.83 pp on DenseUAV. We did not evaluate Tent [Wang et al., 2021], multi-crop averaging, or token merging and make no claim about them.

## A6 Ethics, Reproducibility, and LLM-Use Statements

### Ethics Statement

SKYPART is intended for navigation when satellite positioning is unavailable, for example in search-and-rescue, inspection, and disaster response. The same capability could support unauthorized aerial surveillance; we recommend that deployments include geographic access controls, anonymization

of captured imagery, and operator accountability consistent with applicable law. All benchmarks used are public research datasets; none contains identifiable personal data beyond what their original releases include.

### Reproducibility Statement

The architecture is specified in Sec. 3 and Appendix A1, including all projection dimensions and the loss formulations; the deployed hyperparameters are listed in Table 8 (with experiment-specific overrides in Table 13), and the evaluation protocols, split sizes, and weather augmentation parameters are in Appendices A2.2 and A3.4. Headline numbers in the text are three-seed means with sample standard deviations (Table 1 lists one checkpoint per altitude), and the seeds, checkpoint rule, and evaluator of every control are stated with its table. Code, configuration files, and seeds are released at `https://github.com/chisngyen/SkyPart`.

### Use of Large Language Models

Large language models were used to edit the text for grammar and wording and to help draft parts of the manuscript. All technical content, experiments, and numbers were produced and checked by the authors.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Secs. 1–3.1 match the contributions in Sec. 3: prototype-based part discovery, training-only altitude FiLM/mean-FiLM inference, Kendall-weighted GeoPartLoss, and a lightweight head on DINOv2 ViT-S/14. Secs. 4–4.5 and Appendix A3.5 align with the empirical claims (three benchmarks, strict single-scale $448\times448$ protocol, no reranking/TTA, 26.95 M params / 22.14 GFLOPs, WeatherPrompt robustness Table 3). Sec. 5 does not over-state cross-regional generalization beyond what Appendix A5.2 discusses.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix A5.2 (*Limitations and Broader Impact*) covers within-region splits vs. cross-regional evaluation, finite ($K=10$, $K_{\max}=12$) prototype banks with the salience gate providing per-image adaptivity, semantics not guaranteed human-nameable, mean-FiLM and cross-dataset mismatch (Uni-1652→SUES), dataset-specific augmentation (dense-map rotation vs. canonically oriented pairs; see Sec. 4.2 with development context in Appendix A5.3), and dual-use misuse risk with mitigation pointers.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: Appendix A1.2 states explicitly that its formalized arguments (C1–C4) are applications of established results, not claims of mathematical novelty, and it does not use a numbered theorem–lemma–proof structure with complete proofs for novel results. Where we need rigor (e.g., linear FiLM mean marginals, Kendall multi-task weighting, Pareto-stationarity references), assumptions and derivations are spelled out or cited; a full formal proof archive is not the paper's intended contribution.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Sec. 3 specifies the head (assignment, salience gating, three-way fusion, FiLM, GAT readout) and GeoPartLoss groups. Sec. 4.2 states the deployed student/teacher setup and evaluation contract. Table 8 centralizes frozen architecture knobs and optimizer/MAR/GeoPartLoss hyperparameters ($448\times448$, $K=10$, PK sampling); per-table captions and Appendix A3.3/A3.4 state the epoch budgets that apply to each headline number (deployed recipe vs. controlled ablations). Appendices A2.2, A3.4, and A3.2 document splits, galleries, altitude bins, rotation policy, and weather corruption evaluation. Appendix A1.3/A1.4/A1.5 complete train-only auxiliary details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The full training and evaluation code, configuration files, and seeds are released under the MIT license at `https://github.com/chisngyen/SkyPart`. All experiments use publicly available benchmarks cited in Sec. 4.1 (SUES-200, University-1652, DenseUAV) and a publicly released SSL backbone (DINOv2 ViT-S/14 [Oquab et al., 2024]); no proprietary data is involved. In the meantime, the paper provides the level of detail required for an independent re-implementation: head architecture and assignment rule (Sec. 3), GeoPartLoss groups and uncertainty weighting (Sec. 3.4), single-pass evaluation protocol (Sec. 4.2), the complete frozen hyperparameter table (Table 8), per-dataset splits and gallery rules (Appendices A2.2, A3.1, A3.2), and the WeatherPrompt re-training recipe at $392\times392$ (Appendix A3.4).

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

Justification: Sec. 4.2 and Sec. 4.3 state optimizers, augmentation scope (incl. DenseUAV rotation), EMA, label smoothing, and protocol differences vs. prior work. Table 8 is the single source of truth for numeric hyperparameters. Appendices A2.2 (splits/galleries), A3.1, A3.2, and A3.4 cover dataset-specific gallery/rank rules, full-rank vs. submatrix evaluation where relevant, and weather retraining at $392\times392$ for SkyPart vs. reported baselines.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Headline clean results on SUES-200, University-1652, and DenseUAV, the weather and repair-then-match controls, the FiLM intervention, the inference-wrapper ladder, and the UAV-VisLoc evaluation are reported as three-seed means (42, 123, 2024) with sample standard deviations; paired gains are computed from per-seed differences. The variability captured is random initialization and identity-balanced batch sampling under fixed official splits. Matched single-seed representation ablations are labeled as such in their tables. We report no formal significance tests.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: All experiments were run on single GPUs from the free Kaggle quota (Acknowledgments), and the training budgets, input resolutions, and batch construction are given in Table 8 and Table 13. Per-run wall-clock execution time is not reported.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses public aerial/satellite benchmarks and standard vision backbones for research-only geo-localization.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix A5.2 discusses beneficial uses (GPS-denied navigation for inspection and emergency response) and misuse risks (unauthorized surveillance), with recommended safeguards.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: We do not release a new large generative model or a high-risk scraped dataset; the contribution is a retrieval head and training recipe on existing public CVGL benchmarks and a frozen SSL backbone.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All assets are credited through citations and used under their published terms. Backbone and teacher: DINOv2 code and weights (Apache-2.0). Datasets: SUES-200 (MIT-licensed repository; academic research use only), University-1652 (MIT-licensed repository; dataset released on request for research), DenseUAV (Apache-2.0), and UAV-VisLoc (no license stated in its official release; used for research evaluation only, as in prior work). Tools in the repair-then-match diagnostic: TransWeather (MIT), LightGlue (Apache-2.0), and RoMa (MIT). We do not redistribute any of these assets.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: We do not introduce a new public dataset or benchmark in this submission; the empirical contribution is a method evaluated on existing benchmarks.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve human subjects or crowdsourcing; evaluation is fully automatic retrieval on public image benchmarks.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human-subjects research is conducted in this work.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: No large language model is part of SKYPART, GEOPARTLOSS, or the training/inference pipelines we report. LLMs were used solely for writing and editing assistance and have no impact on methodology, code, or experimental claims, so declaration is not required beyond this negative answer.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.