BEYOND INVARIANCE: LEARNING THE GEOGRAPHIC CORE FOR GENERALIZABLE CROSS-VIEW GEO-LOCALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-view geo-localization (CVGL) identifies locations by matching images across different viewpoints. Existing methods prioritize view-invariant representations, but invariant semantics may exhibit high visual similarity across different locations, limiting geographic discrimination under domain shifts. Although scene semantics can narrow plausible matches, they cannot fully resolve ambiguous places, motivating complementary evidence from fine-grained regional structure. Inspired by sufficient invariant learning, we propose GeoCore to learn a geographic core of complementary semantic and structural evidence. The semantic branch can capture view-invariant scene semantics for coarse discrimination. To resolve the spatial ambiguity, the structural branch learns fine-grained regional structure that remains discriminative across environmental variations. Extensive experiments on UAV and ground-view benchmarks demonstrate that our model consistently outperforms existing approaches, achieving superior performance in cross-area generalization and robustness evaluations.

1 INTRODUCTION

Cross-view geo-localization (CVGL) (Hu et al., 2018; Deuser et al., 2023) retrieves a query's geographic match from geo-referenced imagery captured under different views. Recent UAV and ground-view benchmarks (Zheng et al., 2020; Dai et al., 2024; Workman et al., 2015) span diverse regions and acquisition conditions, making generalizable CVGL essential for deployment. This setting requires transfer to novel scenarios without extra adaptation. However, models trained on a single domain often exploit domain-specific information and degrade in unseen domains (Zhang et al., 2026b;a). Recent approaches employ invariant learning while reducing view-specific noise (Mi et al., 2024; Zhang et al., 2024; Wang et al., 2024a). However, such invariant representations do not necessarily retain transferable discrimination (Wang et al., 2024b; Li et al., 2026).

Figure 1 (a) highlights two limitations of previous methods from the SUES-200 dataset. First, removing buildings causes a larger retrieval performance decline than removing vegetation, indicating that view-invariant semantics contribute unequally to geographic discrimination. Second, the failure case shows that similar scene semantics can confuse the correct match with competing candidates. These observations motivate complementary geographic discrimination: the semantic branch constrains candidate matches through scene-level semantics, while the structural branch can resolve the remaining ambiguity by modeling fine-grained regional structure.

This complementary role has been explored in visual place recognition (Zhang et al., 2025), where semantic knowledge guides the aggregation of local visual patterns to alleviate perceptual aliasing. Beyond resolving ambiguity, complementary evidence may reduce reliance on individual cues under distribution shifts, consistent with sufficient invariant learning (Kim et al., 2025), which emphasizes diverse invariant features for robustness to test-time cue loss. In CVGL, this principle translates into complementary geographic discrimination: scene semantics constrain candidate locations, while regional spatial structure distinguishes semantically similar places. Building on this insight, we define the *geographic core* as a transferable representation that integrates both forms of evidence to support reliable location discrimination across domain shifts and environmental variations. To learn this core, we introduce **GeoCore**, a framework for generalizable and robust CVGL.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

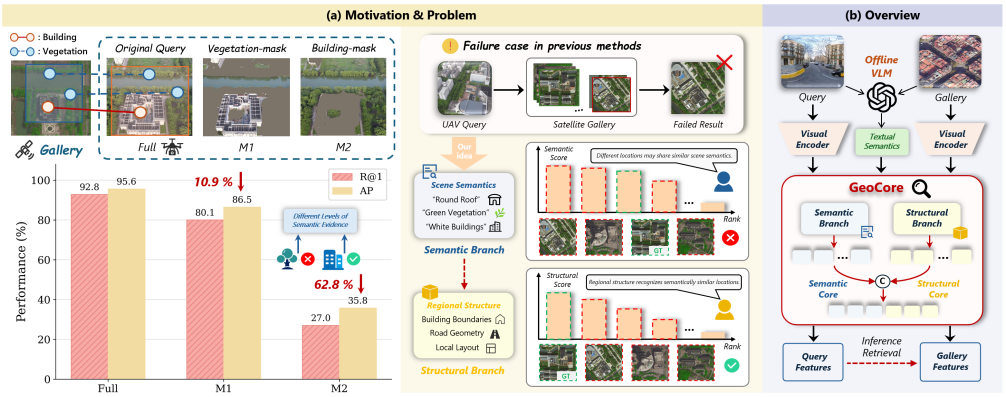


Figure 1: (a) **Motivation:** Semantic-level experiments (left) evaluated using DINOv2. M2 causes a larger retrieval decline than M1, revealing unequal contributions of view-invariant semantics. The failure case (top right) illustrates the difficulty of distinguishing semantically similar locations, motivating a *geographic core* that combines scene semantics with fine-grained regional structure. (b) **Overview:** Two parallel branches capture transferable scene semantics and fine-grained regional structure, which are combined into the geographic core representation for image-only retrieval.

As shown in Figure 1(b), GeoCore employs two branches that capture complementary semantic and structural information, forming the Semantic Core and Structural Core. Specifically, the semantic branch uses view-consistent textual descriptions constructed from the Visual-Language Model (VLM) to learn transferable scene concepts with reduced dependence on domain-specific information. Meanwhile, the structural branch organizes spatial tokens into regional contents and models the corresponding relations through rotation-symmetric refinement, preserving fine-grained geographic distinctions beyond scene semantics. Within this branch, cross-environment regional reliability encourages capturing evidence to remain discriminative under environmental variations. The complementary semantic and structural core features are integrated into a single descriptor for image-only retrieval, without requiring extra textual information at inference. Overall, our main contributions can be summarized as follows:

- We identify a limitation of existing generalizable CVGL methods: view-invariant cues are not equally discriminative, and semantic consistency alone may fail to distinguish different locations. This motivates us to combine complementary semantic and structural evidence for reliable localization across environments.
- We propose GeoCore, a unified framework with two branches. The Textually Anchored Semantic Core learns transferable scene concepts for coarse geographic discrimination, while Reliability-Aware Structural Learning captures fine-grained spatial organization to distinguish semantically similar locations under appearance variations.
- We conduct extensive evaluations on five UAV and ground-view benchmarks. GeoCore consistently outperforms state-of-the-art methods in cross-area generalization and demonstrates strong multi-weather robustness.

2 Related Work

2.1 Cross-View Geo-Localization

Cross-view geo-localization (CVGL) retrieves geo-referenced images across substantially different viewpoints. Ground-view benchmarks introduce realistic Ground-to-Satellite matching and cross-area evaluation (Zhu et al., 2021), while University-1652 introduces UAV-to-Satellite geo-localization (Zheng et al., 2020) and Game4Loc extends this setting to continuous geographic regions under diverse flight conditions (Ji et al., 2025b). Existing methods focus on global-context modeling (Zhu et al., 2022), geometric-layout alignment (Zhang et al., 2024; Zhang & Zhu, 2024), and panorama-to-BEV transformation (Ye et al., 2024). Recent studies further improve robustness

to ground-view variations through contrastive consistency learning (Mi et al., 2024) or exploit complementary multi-modal information for UAV localization (Ji et al., 2025a).

Generalizable CVGL requires transfer to unseen regions and environments. Ouyang et al. (2026) use uncertainty-aware modeling to calibrate negative supervision and improve generalization. InfoGeo uses an information bottleneck to extract object-centric features while suppressing view-specific noise (Zhang et al., 2026b), whereas GeoLink integrates 3D priors into representations for cross-modal semantic consistency (Zhang et al., 2026a). These methods mainly preserve cross-view invariant evidence without explicitly distinguishing its discriminative roles. We learn a geographic core for transferable geographic discrimination beyond cross-view invariance.

2.2 Invariant Learning for Distribution Shift

To address domain shifts, invariant learning seeks stable predictive relationships across environments while reducing domain-specific dependencies. Representative methods enforce shared predictors or minimize risk variation across source domains (Ahuja et al., 2020; Krueger et al., 2021). Recent work further considers structural consistency and high-level dependencies beyond pairwise invariance (Chen et al., 2024). However, invariance learning may retain incomplete signals or discard task-relevant information, motivating predictive sufficiency. PNS-based learning seeks features both necessary and sufficient for prediction (Yang et al., 2023), while INSURE learns compact representations preserving sufficient class-relevant information (Yu et al., 2024). Sufficient invariant learning further emphasizes capturing more invariant features rather than relying on a narrow subset (Kim et al., 2025). Following this principle, GeoCore learns complementary invariant cues through semantic and structural representations for geographic discrimination across domains.

3 Methodology

3.1 Problem Formulation of GeoCore

For each dataset, consider a source-domain training set $\mathcal{D}_{\text{tr}} = \{(x_i^q, x_i^g, t_i, y_i)\}_{i=1}^N$, where $x_i^q \in \mathbb{R}^{C\times H\times W}$ is a query image from UAV or Ground view, $x_i^g \in \mathbb{R}^{C\times H\times W}$ is the corresponding satellite image and t_i is the textual description with location identity y_i. Following InfoGeo (Zhang et al., 2026b), we study generalizable CVGL across datasets without target adaptation. A shared visual encoder $f_\theta(\cdot)$ extracts the global `[CLS]` token $\mathbf{c}_i^v$ and spatial tokens $\mathbf{Z}_i^v$ for each view $v \in \{q, g\}$. During training, GeoCore $g_\theta(\cdot)$ further maps these tokens to a descriptor $\mathbf{f}_i^v = g_\theta(t_i, \mathbf{c}_i^v, \mathbf{Z}_i^v) \in \mathbb{R}^{1\times D}$.

3.2 Learning the Geographic Core

Inspired by sufficient invariant learning (Kim et al., 2025), we aim to reduce reliance on a narrow subset of invariant cues by discovering complementary evidence with reliable discrimination. Generalizable CVGL requires complementary evidence across discriminative levels: transferable scene semantics for coarse narrowing and fine-grained structure for resolving semantic ambiguity. As illustrated in Figure 2, two parallel branches capture the view-invariant semantic core and the fine-grained structure core to improve the generalization and robustness of GeoCore.

3.2.1 Textually Anchored Semantic Core

Recent studies investigate textual guidance for domain generalization (Wen et al., 2025; Kempf et al., 2025) and semantic anchoring for CVGL (Song et al., 2026), showing that textual priors can improve transferability and construct view-consistent correspondences. Therefore, GeoCore uses textual descriptions as view-invariant scene semantics to reduce reliance on domain-specific cues.

For the i-th location, the textual description t_i is obtained from a VLM in an offline manner. t_i is first encoded by a frozen CLIP (Radford et al., 2021) text encoder $E_{\text{text}}(\cdot)$ to extract text embeddings, and then they are projected by the text projector $p_t(\cdot)$:

$$\mathbf{u}_i = p_t(E_{\text{text}}(t_i)) \in \mathbb{R}^{D/2}, \tag{1}$$

Furthermore, `[CLS]` token from each view v is mapped into semantic core by the projector $p_s(\cdot)$:

$$\mathbf{s}_i^v = p_s(\mathbf{c}_i^v) \in \mathbb{R}^{D/2}, \quad v \in \{q, g\}. \tag{2}$$

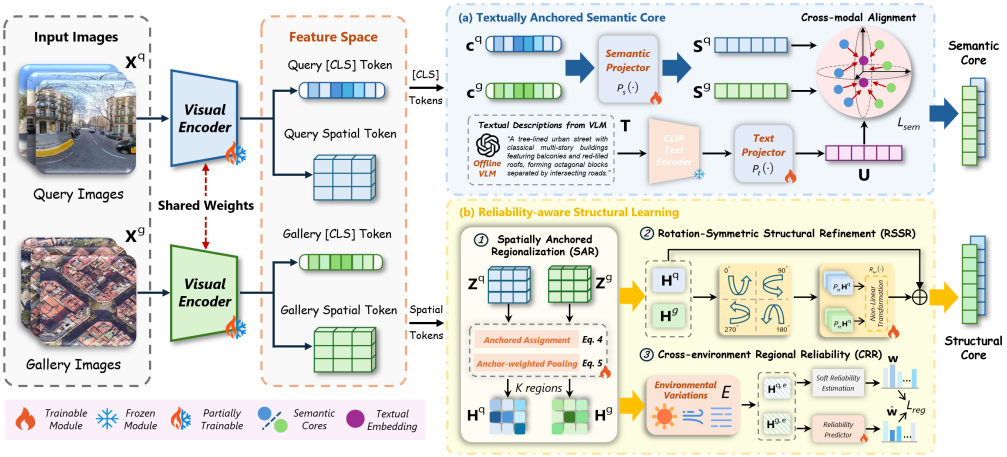


Figure 2: The overview of GeoCore. (a) Textually Anchored Semantic Core is employed to learn transferable scene semantics using VLM's semantic anchors. (b) Reliability-Aware Structural Learning is further proposed to capture fine-grained regional structure that remains geographically discriminative across appearance variations.

The semantic branch aligns both views with the corresponding textual representations:

$$\mathcal{L}_{\text{sem}} = \mathcal{C}(\mathbf{S}^q, \mathbf{U}) + \mathcal{C}(\mathbf{S}^g, \mathbf{U}), \tag{3}$$

where $\mathbf{S}^v = \{\mathbf{s}_i^v\}_{i=1}^N$ and $\mathbf{U} = \{\mathbf{u}_i\}_{i=1}^N$ denote the semantic and textual representations, respectively. $\mathcal{C}(\cdot,\cdot)$ is the cross-modal contrastive objective via InfoNCE (van den Oord et al., 2019).

However, scene-level textual descriptions may overlook differences in local spatial layouts, limiting their ability to distinguish visually similar locations.

3.2.2 Reliability-Aware Structural Learning

To complement the semantic core, we construct the structural core via *Reliability-Aware Structural Learning*, preserving local content and spatial arrangements that scene semantic alignment may overlook. The structural core representations are further regularized through cross-environment regional reliability, encouraging persistent discrimination under environmental variations.

Spatially Anchored Regionalization. GeoCore organizes fine-grained spatial tokens into localized regions through region-level representations (Shlapentokh-Rothman et al., 2024). We softly aggregate $\mathbf{Z}_i^v$ into K coherent regions by content relevance and spatial proximity to form anchored representations for structural refinement and cross-environment reliability learning.

Motivated by differentiable spatial assignment (Jampani et al., 2018), we introduce *Anchored Assignment*, which defines the affinity between a normalized spatial token $\hat{\mathbf{z}}_p \in \mathbb{R}^d$ and region k as:

$$A_{pk} = \frac{\exp(s_{pk}/\tau_a)}{\sum_{k'=1}^{K} \exp\left(s_{pk'}/\tau_a\right)}, \quad s_{pk} = \mathbf{w}_k^\top \hat{\mathbf{z}}_p - \frac{\|\mathbf{p}_p - \mathbf{a}_k\|_2^2}{\tau_{\text{s}}}, \tag{4}$$

where $\mathbf{w}_k$ is a learnable prototype, $\mathbf{p}_p$ and $\mathbf{a}_k \in [0,1]^2$ denote normalized patch-level center and anchor coordinates, respectively. K anchors form a regular $n \times n$ square grid (See Appendix A.1.3 for details). τ_a and τ_{s} control assignment sharpness and spatial sensitivity, respectively.

Subsequently, the regional content $\mathbf{h}_k$ is obtained by the Anchor-weighted Pooling:

$$\mathbf{h}_k = p_r\big((\sum_p A_{pk}\hat{\mathbf{z}}_p)/(\sum_p A_{pk} + \epsilon)\big) \tag{5}$$

where $p_r(\cdot)$ denotes a learnable projection and ϵ ensures numerical stability (ϵ is 10^{-8}). The regional contents $\mathbf{H}_i = [\mathbf{h}_1, \ldots, \mathbf{h}_K]^\top \in \mathbb{R}^{K \times D/2}$ preserve the anchor-grid ordering, providing spatially anchored local evidence for subsequent structural refinement and reliability learning.

Rotation-Symmetric Structural Refinement. Spatially Anchored Regionalization produces spatially ordered regional representations. However, prior work (Ouyang & Zhu, 2025) adopts order-agnostic aggregation, which discards the relative arrangement of regions and weakens fine-grained structural modeling. An order-sensitive readout can preserve these arrangements but may depend on a single grid orientation. Therefore, we introduce the Rotation-Symmetric Structural Refinement (RSSR) to reduce this dependence while retaining regional structure.

This module encodes inter-region arrangements while averaging a shared order-sensitive readout over the four quarter-turn permutations of the encoded anchor grid. Using the group averaging strategy (Puny et al., 2022), we construct structural core $\mathbf{r}_i^v$ from regional contents $\mathbf{H}_i^v$:

$$\mathbf{r}_i^v = \frac{1}{K}\sum_{k=1}^{K}\mathbf{h}_{i,k}^v + \frac{\gamma}{|\mathcal{R}|}\sum_{\pi\in\mathcal{R}} R_\omega(P_\pi \mathbf{H}_i^v), \tag{6}$$

where γ is a learnable residual scale, $\mathcal{R} = \{0°, 90°, 180°, 270°\}$, and P_π permutes grid positions according to rotation π. $R_\omega : \mathbb{R}^{K\times D/2} \to \mathbb{R}^{D/2}$ is a nonlinear and order-sensitive transformation that encodes regional arrangements (More details are illustrated in Appendix A.2.3).

Cross-environment Regional Reliability. While RSSR captures rotation-consistent spatial arrangements, it does not determine whether regional evidence remains stable and discriminative under different conditions. Accordingly, we construct reliability supervision from E kinds of geometry-preserving variations (Hendrycks & Dietterich, 2019), sampled independently from locations.

Following soft correspondence estimation in local feature matching (Sarlin et al., 2020), we formulate regional reliability through correspondence-weighted aggregation. For region k in view v ($\bar{v}$ denotes the complementary view), the discriminative margin under variation e is defined as:

$$m_{i,k}^{v,e} = \mathrm{sim}_{\mathrm{soft}}\big(\mathbf{h}_{i,k}^{v,e}, \mathbf{H}_i^{\bar{v},e}\big) - \max_{1\le l\le K}\big\langle \mathbf{h}_{i,k}^{v,e}, \mathbf{h}_{j,l}^{\bar{v},e}\big\rangle, \quad \text{where } j \neq i, \tag{7}$$

where $\mathrm{sim}_{\mathrm{soft}}(\cdot,\cdot)$ aggregates similarities to regions of the matched location in the view $\bar{v}$ using correspondence weights estimated from unperturbed features and fixed across appearance variations (see Appendix A.1.4). This margin measures regional geographic discrimination. Then, we aggregate margins across E environmental variations to estimate soft reliability targets $\mathbf{w}_i^v = \{w_{i,k}^v\}$:

$$\rho_{i,k}^v = -\tau_e \log\Big(\frac{1}{E}\sum_{e=1}^{E}\exp(-m_{i,k}^{v,e}/\tau_e)\Big), \qquad w_{i,k}^v = \sigma\Big(\frac{\rho_{i,k}^v - \mu}{\tau_t}\Big), \tag{8}$$

where $\sigma(\cdot)$ denotes sigmoid, μ is a reference margin, while τ_e, τ_t control margin aggregation and target softness. These reliability targets reflect the geographic discriminability of regional cues.

Finally, regional reliability is predicted by a reliability predictor $G_\psi(\cdot)$ as $\hat{\mathbf{w}}_i^v = \sigma(G_\psi(\mathbf{H}_i^v))$, where $\hat{\mathbf{w}}_i^v = \{\hat{w}_{i,k}^v\}_{k=1}^K$. We propose auxiliary supervision for cross-environment reliability prediction:

$$\mathcal{L}_{\mathrm{reg}} = -\frac{1}{2}\sum_{v\in\{q,g\}}\mathbb{E}_{i,k}\left[\mathrm{sg}(w_{i,k}^v)\log\hat{w}_{i,k}^v + \big(1-\mathrm{sg}(w_{i,k}^v)\big)\log\big(1-\hat{w}_{i,k}^v\big)\right], \tag{9}$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator. This objective encourages regional features to remain geographically discriminative across environments (More Analysis is in Appendix A.1.4).

3.3 Optimization and Inference

GeoCore directly combines semantic and structural core features into a scene-level descriptor:

$$\mathbf{f}_i^v = \mathrm{Concat}(\mathbf{s}_i^v, \alpha\mathbf{r}_i^v), \tag{10}$$

where $\mathrm{Concat}(\cdot,\cdot)$ denotes concatenation and α controls the contribution of the structural branch. This design avoids introducing an extra feature aggregation module, and only images are processed to obtain descriptors $\mathbf{F}^v = \{\mathbf{f}_i^v\}_{i=1}^N$ during inference (See Appendix Algorithm 2 for details).

Following previous CVGL work (Deuser et al., 2023), we apply InfoNCE loss $\mathcal{L}_{\mathrm{align}}$ for cross-view alignment of descriptors. The overall framework is jointly optimized through the objective:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{align}} + \lambda_{\mathrm{sem}}\mathcal{L}_{\mathrm{sem}} + \lambda_{\mathrm{reg}}\mathcal{L}_{\mathrm{reg}}, \tag{11}$$

where λ_{sem} and λ_{reg} are the regularization weights for semantic and structural branch.

Table 1: Comparison (%) with recent state-of-the-art UAV CVGL methods. Dim is the dimension of the descriptors. The best results are marked in **bold** and the second-best ones are underlined. † denotes models equipped with a DINOv3 backbone.

Methods	Dim	University→SUES		University→DenseUAV		DenseUAV→SUES		DenseUAV→University	
		R@1	AP	R@1	AP	R@1	AP	R@1	AP
MCCG	1536	70.31	74.58	17.54	13.01	80.09	83.18	50.52	55.35
Sample4Geo	2048	82.03	85.12	33.46	22.28	78.94	81.86	37.90	42.54
DAC	1024	87.65	89.80	37.81	26.99	86.44	88.70	45.41	50.03
CAMP	1024	88.34	90.49	36.94	26.39	86.79	88.93	44.11	48.60
MFRGN	1024	78.28	81.75	23.60	16.08	67.22	71.14	37.95	41.73
Game4Loc	1024	86.75	89.07	17.76	13.62	67.19	71.40	26.47	31.05
MMGeo	1024	85.01	87.60	22.88	18.54	79.05	82.05	33.92	38.28
CVcities	4096	90.27	91.77	38.87	28.44	89.67	91.34	66.08	70.01
InfoGeo	4096	93.38	94.35	39.88	29.13	93.42	94.55	68.16	72.15
GeoCore	1536	96.61$_{+3.23}$	97.20$_{+2.85}$	**54.70**$_{+14.82}$	**40.37**$_{+11.24}$	95.05$_{+1.63}$	96.12$_{+1.57}$	70.84$_{+2.68}$	74.66$_{+2.51}$
GeoCore†	1536	**98.07**$_{+4.69}$	**98.45**$_{+4.10}$	50.71$_{+10.83}$	38.98$_{+9.85}$	**99.19**$_{+5.77}$	**99.37**$_{+4.82}$	**86.84**$_{+18.68}$	**88.94**$_{+16.79}$
Avg. Δ	-	**+3.96**	**+3.48**	**+12.83**	**+10.55**	**+3.70**	**+3.20**	**+10.68**	**+9.65**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Benchmarks. We evaluate GeoCore on five commonly used benchmarks covering UAV and ground-view CVGL, including University-1652 (Zheng et al., 2020), SUES-200 (Zhu et al., 2023), and DenseUAV (Dai et al., 2024). For ground-view localization, we adopt CVUSA (Workman et al., 2015) and CVACT (Liu & Li, 2019), two large-scale ground-to-satellite benchmarks, with CVACT additionally supporting GPS-based retrieval. These datasets span diverse platforms, scenes, and acquisition conditions, enabling comprehensive cross-domain generalization evaluation. Detailed dataset statistics are provided in the Appendix A.2.1.

Evaluation Protocols. Following standard CVGL protocols, we report Recall@K (R@K) for UAV and ground-view benchmarks, where a query is correctly localized if its reference image appears in the top-K results. We additionally report AP for UAV datasets and Recall@1% (R@1%) for ground-view datasets following prior works.

Implementation Details. GeoCore is implemented with DINOv2 (Oquab et al., 2024) and DINOv3 (Siméoni et al., 2025) ViT-Base visual encoders to assess its applicability, while most experiments in the main text are conducted using DINOv2 for consistency. Textual descriptions are reproduced from GLM-4.6V-Flash (GLM-V Team et al., 2026). A frozen CLIP text encoder is used for semantic supervision. We set the number of spatial regions K to 25, while the structural fusion coefficient α is 1.0 for UAV and 1.5 for Ground. The dimension D of retrieval descriptors is 1536. GeoCore is trained for 9 epochs with a batch size of 16 for UAV-based CVGL and for 18 epochs with a batch size of 64 for ground-view CVGL. The corresponding λ_{sem} and λ_{reg} are $(0.5, 0.95)$ and $(0.3, 1.05)$, respectively. More detailed hyperparameters and network designs are provided in Appendix A.2.2 and A.2.3, while additional ablation studies are illustrated in Appendix A.3.6.

4.2 COMPARING WITH STATE-OF-THE-ART MODELS

UAV-based CVGL Evaluations. Our comparisons cover recent SOTA UAV-based CVGL approaches, including MCCG (Shen et al., 2024), Sample4Geo (Deuser et al., 2023), DAC (Xia et al., 2024), CAMP (Wu et al., 2024), MFRGN (Wang et al., 2024b), Game4Loc (Ji et al., 2025b), MMGeo (Ji et al., 2025a), CVCities (Huang et al., 2024), and InfoGeo (Zhang et al., 2026b). As shown in Table 1, GeoCore consistently outperforms recent CVGL methods across all four cross-dataset UAV transfer settings without adaptation. Averaged over the DINOv2 and DINOv3 variants, GeoCore improves the previous best results by 3.96 and 3.48 points at R@1 and AP on University-1652→SUES-200, 3.70 and 3.20 on DenseUAV→SUES-200. Larger gains are observed under stronger domain shifts, reaching 12.83 and 10.55 on University→DenseUAV and 10.68 and 9.65 on DenseUAV→University-1652. Notably, GeoCore achieves 54.70% R@1 on University-1652→DenseUAV, while GeoCore† reaches 86.84% on DenseUAV→University-1652, surpassing

InfoGeo by 14.82% and 18.68%, respectively. These results demonstrate that the geographic core enables more discriminative and generalizable features across different UAV scenarios. Detailed results of SUES-200 are offered in Appendix A.3.1.

Table 2: Comparison (%) with recent state-of-the-art Ground-to-Satellite CVGL methods

Methods	CVUSA→CVACT$_{val}$		CVUSA→CVACT$_{test}$		CVACT→CVUSA	
	R@1	R@1%	R@1	R@1%	R@1	R@1%
L2LTR (Yang et al., 2021)	47.55	91.39	-	-	-	-
GeoDTR (Zhang et al., 2023)	47.79	92.20	11.24	72.09	-	-
Sample4Geo (Deuser et al., 2023)	56.62	94.69	27.78	94.88	44.95	90.95
EP-BEV (Ye et al., 2024)	59.32	94.69	32.68	95.21	38.32	85.54
CLGT (Ouyang & Zhu, 2025)	64.82	96.16	35.52	96.35	14.96	64.81
GeoCore (Ours)	77.99$_{+13.17}$	97.94$_{+1.78}$	47.92$_{+12.40}$	98.00$_{+1.65}$	72.60$_{+27.65}$	98.20$_{+7.25}$
GeoCore† (Ours)	**82.33**$_{+17.51}$	**98.05**$_{+1.89}$	**54.36**$_{+18.84}$	**98.11**$_{+1.76}$	**75.37**$_{+30.42}$	**98.33**$_{+7.38}$
Avg.Δ	**+15.34**	**+1.84**	**+15.62**	**+1.71**	**+29.04**	**+7.32**

Ground-to-Satellite Matching Evaluations. To validate the generalization capability of GeoCore under distribution shifts, we conduct extensive experiments between CVUSA and CVACT in Table 2. GeoCore consistently achieves substantial improvements over existing CVGL methods across all transfer settings. For example, when trained on CVUSA and evaluated on CVACT$_{test}$, GeoCore achieves 47.92% R@1, outperforming CLGT by 12.40%. With the DINOv3 backbone, GeoCore† further improves the performance to 54.36% R@1, yielding a remarkable gain of 18.84% over previous methods. Moreover, GeoCore† achieves 75.37% R@1 on the CVACT→CVUSA transfer setting, surpassing the strongest competitor by 30.42%. These consistent improvements demonstrate that GeoCore effectively captures transferable geographic cues rather than dataset-specific visual patterns, leading to robust cross-view localization under unseen domain shifts.

UAV Multi-Weather Robustness Evaluations. To evaluate the robust performance of the proposed GeoCore under challenging environmental variations, we further conducted extensive experiments on University-1652→SUES-200 with three adverse weather conditions (Zheng et al., 2024). As shown in Table 3, the reported results are averaged over four altitudes, covering three representative scenarios: **Fog-Rain**, **Fog-Snow**, and **Rain-Snow**.

GeoCore consistently outperforms prior methods across all weather conditions, demonstrating strong robustness to severe appearance shifts. Compared with InfoGeo, GeoCore improves R@1 and AP by an average of 5.53 and 5.10 points across the three weather scenarios. Under the challenging Fog-Snow condition, it further gains 6.76 R@1 and 6.14 AP points over the previous best. GeoCore† with DINOv3 achieves the strongest results on Fog-Snow and Rain-Snow. These results show that GeoCore captures stable geographic structures beyond weather-specific appearance, enabling reliable CVGL under diverse perturbations.

Table 3: Average performance comparison under adverse weather shifts on University-1652→SUES-200.

Methods	Fog-Rain		Fog-Snow		Rain-Snow	
	R@1	AP	R@1	AP	R@1	AP
Sample4Geo	37.40	44.13	26.68	33.05	45.81	51.58
DAC	73.23	76.73	45.03	50.57	68.97	73.07
CV-cities	75.11	79.06	72.33	76.51	85.24	87.79
InfoGeo	79.89	82.99	76.56	80.15	89.19	90.89
GeoCore	**86.18**$_{+6.29}$	**88.76**$_{+5.77}$	83.32$_{+6.76}$	86.29$_{+6.14}$	92.73$_{+3.54}$	94.28$_{+3.39}$
GeoCore†	84.84$_{+4.95}$	87.55$_{+4.56}$	**84.08**$_{+7.52}$	**86.60**$_{+6.45}$	**95.05**$_{+5.86}$	**95.96**$_{+5.07}$

Complexity Analysis. As presented in Table 4, GeoCore maintains favorable efficiency with strong localization performance. The DINOv2 variant requires 99.61M parameters and 106.95G FLOPs, while GeoCore† reduces them to 96.67M and 79.00G. GeoCore also achieves the lowest latency of 12.59 ms, whereas GeoCore† remains competitive at 16.34 ms and is faster than InfoGeo.

Despite relatively high parameter counts and FLOPs, GeoCore maintains low latency with compact retrieval descriptors, while GeoCore† remains competitive in inference efficiency with existing DINOv2-based methods. Both variants support efficient retrieval with strong generalization performance and robustness.

Table 4: Computational cost evaluation of UAV-based CVGL methods.

	Sample4Geo	CVcities	InfoGeo	GeoCore	GeoCore†
Params (M)	87.51	90.50	121.01	99.61	96.67
FLOPs (G)	72.60	91.67	105.63	106.95	79.00
Latency (ms)	13.79	15.22	19.36	12.59	16.34

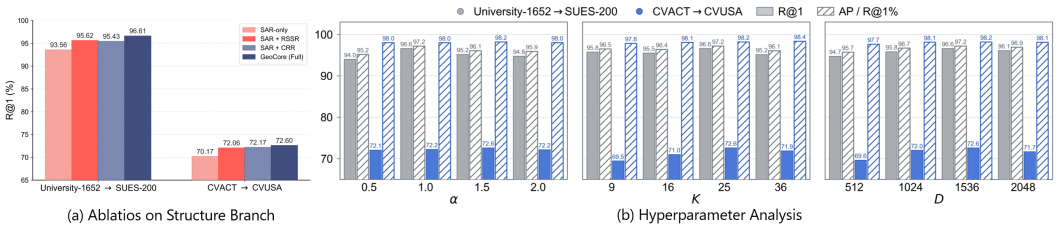


Figure 3: More ablation studies. (a) Ablation of key components in *Reliability-Aware Structural Learning*, including RSSR and CRR, to examine their individual contributions to structural robustness. (b) Hyperparameter analysis of α, K, and D across UAV-based and ground-view CVGL transfer scenarios, evaluating the sensitivity and stability of GeoCore under different settings.

4.3 Ablation Studies

Ablations on Main Components. We conduct ablation studies to evaluate the contributions of two components: Textually Anchored Semantic Core and Reliability-Aware Structural Learning.

Table 5: Ablation study of main components.

Methods	University→SUES		CVACT→CVUSA	
	R@1	AP	R@1	R@1%
GeoCore	**96.61**	**97.20**	**72.60**	**98.20**
w/o Semantic	93.74	94.80	68.93	97.92
w/o Structural	92.41	93.57	71.00	97.38
Baseline	91.28	92.71	68.55	97.30

As shown in Table 5, removing either component consistently degrades performance across both scenarios. Removing the Semantic Branch decreases R@1 by 2.87 % and 3.67 % on University→SUES and CVACT→CVUSA, demonstrating the importance of learning transferable scene semantics. Meanwhile, removing the Structural Branch causes a larger degradation on UAV-based CVGL, with a 4.20 % R@1 drop, highlighting the necessity of spatial structure modeling under viewpoint variations. These results support the complementary contributions of transferable semantics and fine-grained structural evidence.

We also explore the effectiveness of different variants in Reliability-Aware Structural Learning in Figure 3 (a). Starting from SAR, introducing RSSR improves R@1 from 93.56 to 95.62 on University→SUES and from 70.17 to 72.06 on CVACT→CVUSA, while CRR yields comparable gains. Combining both components achieves the best performance of 96.61 and 72.60 R@1, confirming their complementary roles in learning transferable spatial structures.

Hyperparameter Analysis. To examine the effectiveness of key hyperparameters on GeoCore, we study the number of spatial regions K, the structural fusion coefficient α, and the descriptor dimension D. The parameter K controls regional granularity, α balances semantic and structural representations, and D determines representation capacity. As shown in Figure 3 (b), GeoCore performs consistently across a broad range of hyperparameter settings. Increasing the number of regions and descriptor dimensionality improves performance before the gains saturate, suggesting a balance between granularity and compactness, while the preferred fusion weight varies between UAV and ground-view transfer. We therefore use 25 regions and 1536-dimensional descriptors, with fusion weights of 1.0 and 1.5 for UAV and ground-view experiments, respectively.

4.4 Visualization Analysis

Visualization on Multi-Weather Robustness Evaluation. As illustrated in Figure 4 (a), we visualize GeoCore representations using PCA analysis (Abdi & Williams, 2010) under normal and adverse weather conditions. The results show that GeoCore learns compact and consistent cross-view features, where corresponding geographic instances remain well aligned despite severe appearance variations. This demonstrates that the learned geographic core captures transferable semantics beyond weather-specific appearances, supporting robust localization under environmental shifts.

Furthermore, Figure 4 (b) provides spatial affinity comparisons between GeoCore and the DINOv2 baseline. Both methods successfully identify the dominant building regions that provide strong localization evidence. However, GeoCore additionally preserves responses around structural boundaries and surrounding spatial layouts, indicating that the proposed structural learning captures fine-

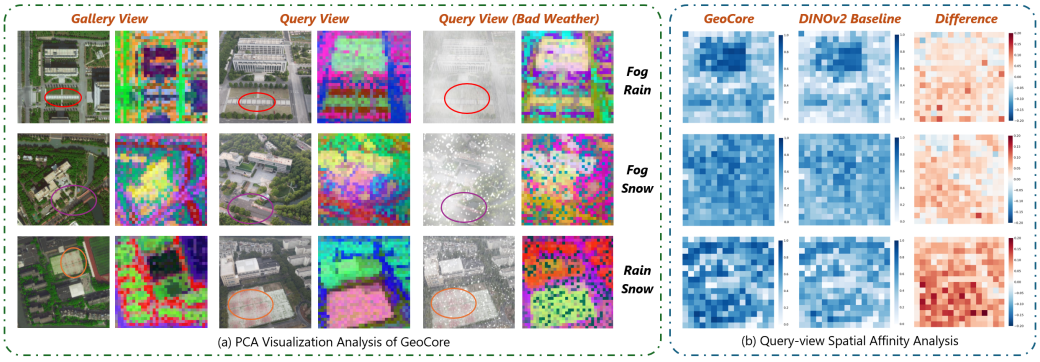


Figure 4: Visualization analysis of GeoCore. (a) PCA visualization of GeoCore on normal and adverse weather scenarios (Circles denote view-consistent visual cues). (b) Spatial affinity comparison between GeoCore and the DINOv2 baseline across normal and adverse weather conditions, with difference maps highlighting detailed structural responses captured by GeoCore.

grained regional structure beyond salient visual regions. These observations verify that GeoCore exploits reliable spatial structures to complement semantic consistency for CVGL.

Attention Maps of Semantic Core and Structural Core. Figure 5 visualizes the attention maps of different cores via Grad-CAM (Selvaraju et al., 2017) on both ground and UAV scenarios.

The Semantic Core generally attends to broad scene-level regions shared across heterogeneous viewpoints, capturing transferable semantic evidence such as buildings, roads, and surrounding context. In contrast, the Structural Core places stronger emphasis on spatially localized patterns, including building boundaries, road geometry, courtyards, and their relative arrangements. This distinction is particularly evident on SUES-200, where the structural representation highlights fine-grained layouts that are less apparent in the semantic response. The final GeoCore representation combines these complementary attention patterns, retaining cross-view semantic consistency and location-specific structural evidence.

These visualizations reveal complementary attention patterns: the semantic guidance captures broad scene-level regions shared across views, while reliability-aware structural learning emphasizes fine-grained regional structure, jointly supporting generalization and robustness of CVGL.

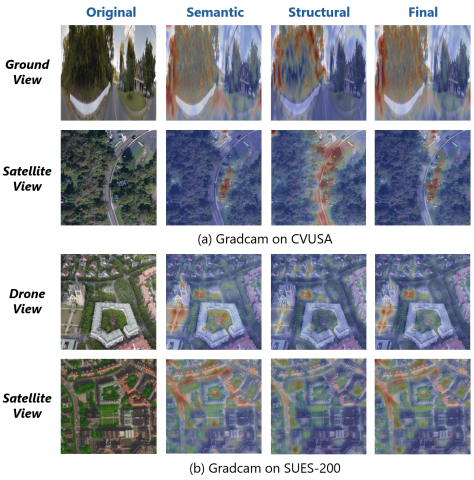


Figure 5: Grad-CAM Visualization of GeoCore. Attention maps on the Semantic Core, Structural Core, and Final descriptor.

5 Conclusion

In this work, we introduce GeoCore to advance generalizable CVGL beyond cross-view invariance toward learning geographically core representations. GeoCore discovers the *geographic core* through Textually Anchored Semantic Core and Reliability-Aware Structural Learning, jointly preserving transferable scene semantics and fine-grained regional structure. Extensive evaluations demonstrate that GeoCore is a practical framework for both UAV and ground-view CVGL, consistently delivering strong generalization and robustness across diverse cross-area and environmental settings. These results validate the complementary roles of hierarchical geographic evidence: transferable semantics narrow candidate locations, while fine-grained structure resolves residual ambiguity. The consistent generalization, robustness, and efficient image-only retrieval further highlight GeoCore's potential to bridge methodological advances and practical deployment, motivating future extensions to broader sensing modalities and more comprehensive real-world scenarios.

AI USE STATEMENT

We used large language models (LLMs) to assist with language polishing, manuscript organization, and exploratory research discussions. All suggestions were critically reviewed by the authors, who take full responsibility for the scientific claims, analyses, and results. The use of a vision-language model for generating training scene-level descriptions is documented separately in the Methodology and Appendix.

ETHICS STATEMENT

This study follows the ethical guidelines and submission requirements of ICLR. All datasets, pre-trained models, and open-source resources are used in accordance with their respective licenses and intended research purposes. The experimental code was independently developed or appropriately adapted with attribution to the original projects, and will be released in accordance with the applicable policies. The authors declare no conflicts of interest and take full responsibility for the methodology, experimental results, and conclusions presented in this work.

REPRODUCIBILITY STATEMENT

The architecture, learning objectives, and key methodological details of GeoCore are provided in Appendix Section A.1, including textual semantic learning, spatially anchored regionalization, rotation-symmetric structural refinement, and cross-environment regional reliability. Appendix Section A.2 further specifies the implementation settings, training procedure, and inference protocol. Extended quantitative evaluations and qualitative analyses are reported in Appendix Sections A.3 and A.4, respectively, while additional discussions are provided in Appendix Section A.5. Overall, these materials provide the methodological and experimental details necessary to facilitate reproduction of our results.

REFERENCES

Hervé Abdi and Lynne J. Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459, 2010.

Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012.

Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 145–155, 2020.

Chaoqi Chen, Luyao Tang, and Hui Huang. Reconstruct and match: Out-of-distribution robustness via topological homogeneity. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

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.

Fabian Deuser, Konrad Habel, and Norbert Oswald. Sample4geo: Hard negative sampling for cross-view geo-localisation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16801–16810. IEEE, 2023.

GLM-V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, et al. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2026. URL https://arxiv.org/abs/2507.01006.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

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 *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7258–7267. IEEE, 2018.

Gaoshuang Huang, Yang Zhou, Luying Zhao, and Wenjian Gan. Cv-cities: Advancing cross-view geo-localization in global cities. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18:1592–1606, 2024.

Varun Jampani, Deqing Sun, Ming-Yu Liu, Ming-Hsuan Yang, and Jan Kautz. Superpixel sampling networks. In *European conference on computer vision*, pp. 363–380. Springer, 2018.

Yuxiang Ji, Boyong He, Zhuoyue Tan, and Liaoni Wu. MMGeo: Multimodal compositional geo-localization for UAVs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 25165–25175, 2025a.

Yuxiang Ji, Boyong He, Zhuoyue Tan, and Liaoni Wu. Game4loc: A uav geo-localization benchmark from game data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3913–3921, 2025b.

Elias Kempf, Simon Schrodi, Max Argus, and Thomas Brox. When and how does CLIP enable domain and compositional generalization? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29706–29734. PMLR, 2025. URL https://proceedings.mlr.press/v267/kempf25a.html.

Taero Kim, Subeen Park, Sungjun Lim, Yonghan Jung, Krikamol Muandet, and Kyungwoo Song. Sufficient invariant learning for distribution shift. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4958–4967. IEEE, 2025.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (REx). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5815–5826, 2021.

Jiaxi Li, Di Lin, Hao Chen, Hongyings Liu, Liang Wan, and Wei Feng. Causal inference via style bias deconfounding for domain generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5617–5626. IEEE, 2019.

Li Mi, Chang Xu, Javiera Castillo-Navarro, Syrielle Montariol, Wen Yang, Antoine Bosselut, and Devis Tuia. Congeo: Robust cross-view geo-localization across ground view variations. In *European Conference on Computer Vision*, pp. 214–230. Springer, 2024.

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, 2024.

Songsong Ouyang and Yingying Zhu. Cvgl: Causal learning and geometric topology. *Advances in Neural Information Processing Systems*, 38:78247–78266, 2025.

Songsong Ouyang, Le Wu, and Yingying Zhu. Must all negatives be pushed away equally? uncertainty-aware cross-view geo-localization via normal inverse gamma distribution. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Omri Puny, Matan Atzmon, Heli Ben-Hamu, Ishan Misra, Aditya Grover, Edward J. Smith, and Yaron Lipman. Frame averaging for invariant and equivariant network design. *arXiv preprint arXiv:2110.03336*, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4938–4947, 2020.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

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.

Michal Shlapentokh-Rothman, Ansel Blume, Yao Xiao, Yuqun Wu, Sethuraman Tv, Heyi Tao, Jae Yong Lee, Wilfredo Torres, Yu-Xiong Wang, and Derek Hoiem. Region-based representations revisited. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17107–17116. IEEE, 2024.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Zixuan Song, Jing Zhang, Di Wang, Zidie Zhou, Wenbin Liu, Haonan Guo, En Wang, and Bo Du. GeoBridge: A semantic-anchored multi-view foundation model bridging images and text for geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27793–27803, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Song_GeoBridge_A_Semantic-Anchored_Multi-View_Foundation_Model_Bridging_Images_and_Text_CVPR_2026_paper.html.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2019. URL https://arxiv.org/abs/1807.03748.

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, 2024a.

Yuntao Wang, Jinpu Zhang, Ruonan Wei, Wenbo Gao, and Yuehuan Wang. Mfrgn: Multi-scale feature representation generalization network for ground-to-aerial geo-localization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 2574–2583, 2024b.

Changsong Wen, Zelin Peng, Yu Huang, Xiaokang Yang, and Wei Shen. Domain generalization in CLIP via learning with diverse text prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9559–9569, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wen_Domain_Generalization_in_CLIP_via_Learning_with_Diverse_Text_Prompts_CVPR_2025_paper.html.

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*, pp. 3961–3969, 2015.

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.

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.

Hongji Yang, Xiufan Lu, and Yingying Zhu. Cross-view geo-localization with layer-to-layer transformer. *Advances in Neural Information Processing Systems*, 34:29009–29020, 2021.

Mengyue Yang, Zhen Fang, Yonggang Zhang, Yali Du, Furui Liu, Jean-Francois Ton, Jianhong Wang, and Jun Wang. Invariant learning via probability of sufficient and necessary causes. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Junyan Ye, Zhutao Lv, Weijia Li, Jinhua Yu, Haote Yang, Huaping Zhong, and Conghui He. Cross-view image geo-localization with panorama-BEV co-retrieval network. In *European Conference on Computer Vision*, pp. 74–90, 2024.

Xi Yu, Huan-Hsin Tseng, Shinjae Yoo, Haibin Ling, and Yuewei Lin. Insure: An information theory inspired disentanglement and purification model for domain generalization. *IEEE Transactions on Image Processing*, 33:3508–3519, 2024.

Hongyang Zhang, Yinhao Liu, Haitao Zhang, Zhongyi Wen, Zhenyu Kuang, Shuxian Liang, and Xiansheng Hua. Geolink: A 3d-aware framework to improve generalization for cross-view geo-localization. *arXiv preprint arXiv:2604.13183*, 2026a.

Hongyang Zhang, Maonan Wang, Ziyao Wang, Hongrui Yin, and Man-On Pun. Infogeo: Information-theoretic object-centric learning for cross-view generalizable uav geo-localization. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026b.

Qingwang Zhang and Yingying Zhu. Aligning geometric spatial layout in cross-view geo-localization via feature recombination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7251–7259, 2024.

Sitao Zhang, Hongda Mao, Qingshuang Chen, and Yelin Kim. Efficient visual place recognition through multimodal semantic knowledge integration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5601–5610, 2025.

Xiaohan Zhang, Xingyu Li, Waqas Sultani, Yi Zhou, and Safwan Wshah. Cross-view geo-localization via learning disentangled geometric layout correspondence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 3480–3488, 2023.

Xiaohan Zhang, Xingyu Li, Waqas Sultani, Chen Chen, and Safwan Wshah. Geodtr+: Toward generic cross-view geolocalization via geometric disentanglement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10419–10433, 2024.

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*, pp. 1395–1403, 2020.

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*, pp. 1–3. Association for Computing Machinery, 2024.

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.

Sijie Zhu, Taojiannan Yang, and Chen Chen. VIGOR: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3640–3649, 2021.

Sijie Zhu, Mubarak Shah, and Chen Chen. TransGeo: Transformer is all you need for cross-view image geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1162–1171, 2022.

A APPENDIX

Appendix Outline. This appendix presents the additional implementation details of the methodology, extended experiments, visualization, and further analysis of the proposed GeoCore. The contents are organized as follows:

- § **Section A.1**: More Details on Methodology.
- § **Section A.2**: More Implementation Details.
- § **Section A.3**: Extended Quantitative Results.
- § **Section A.4**: Qualitative Results.
- § **Section A.5**: Discussion.

A.1 MORE DETAILS ON METHODOLOGY

A.1.1 OVERALL ALGORITHMS OF GEOCORE

Training Pipeline. As summarized in Algorithm 1, GeoCore uses paired query-gallery images and source-location descriptions, encoded by a frozen text encoder and projected into a shared semantic space. For each mini-batch, the shared visual encoder extracts global and spatial tokens. The semantic branch learns textual alignment through $\mathcal{L}_{\text{sem}}$, while the structural branch aggregates tokens into anchored regions and refines their spatial arrangement. Regional reliability supervision is derived from cross-view matching margins across geometry-preserving appearance variations. The two representations are concatenated into retrieval descriptors, and training jointly minimizes cross-view alignment, semantic alignment, and regional reliability losses.

Inference Pipeline. As summarized in Algorithm 2, the shared visual encoder extracts global and spatial tokens from each query and gallery image. The semantic and structural branches produce complementary representations, which are concatenated and normalized into retrieval descriptors. Localization compares each query descriptor with all gallery descriptors and returns the highest-ranked gallery identity.

A.1.2 THEORETICAL ANALYSIS ON GEOCORE

We further provide a theory-oriented analysis of GeoCore to clarify why complementary semantic and structural cores can improve geographic discrimination and why cross-environment reliability favors persistent discriminative cues. Rather than claiming statistical sufficiency of the learned representation, our analysis characterizes the retrieval margin of the fused descriptor and its behavior under environmental variations.

Complementary Geographic Discrimination. Let $\mathbf{s}_i^v$ and $\mathbf{r}_i^v$ denote the semantic and structural core features of location i from view $v \in \{q, g\}$, respectively. For the analysis, we consider branch-wise ℓ_2-normalized representations,

$$\|\mathbf{s}_i^v\|_2 = \|\mathbf{r}_i^v\|_2 = 1. \tag{12}$$

Accordingly, the normalized GeoCore descriptor can be written as:

$$\mathbf{f}_i^v = \frac{[\mathbf{s}_i^v; \alpha \mathbf{r}_i^v]}{\sqrt{1+\alpha^2}}, \tag{13}$$

where $\alpha > 0$ controls the contribution of structural evidence.

Given a query from location i, let g_i^+ denote its correct gallery match and g_j^-, $j \neq i$, an arbitrary competing location. We define the semantic and structural pairwise retrieval margins as below:

$$\Delta_s(i,j) = \left\langle \mathbf{s}_i^q, \mathbf{s}_i^{g^+} \right\rangle - \left\langle \mathbf{s}_i^q, \mathbf{s}_j^{g^-} \right\rangle, \tag{14}$$

$$\Delta_r(i,j) = \left\langle \mathbf{r}_i^q, \mathbf{r}_i^{g^+} \right\rangle - \left\langle \mathbf{r}_i^q, \mathbf{r}_j^{g^-} \right\rangle. \tag{15}$$

Algorithm 1: Training Pipeline of GeoCore

Require : Source-domain pairs $\mathcal{D}_{\text{tr}} = \{(x_i^q, x_i^g, t_i, y_i)\}_{i=1}^N$; region number K; variation number E
Ensure : Image-level retrieval model $\{f_\theta, p_s, p_r, \{\mathbf{w}_k\}_{k=1}^K, R_\omega\}$

Initialize trainable parameters $\Theta = \{\theta, p_s, p_t, p_r, \{\mathbf{w}_k\}_{k=1}^K, R_\omega, G_\psi, \gamma\}$
foreach *source location description* t_i **do**
  $\mathbf{e}_i \leftarrow E_{\text{text}}(t_i)$ using the frozen CLIP text encoder
for *each training epoch* **do**
  foreach *mini-batch* $\mathcal{B} \subset \mathcal{D}_{\text{tr}}$ **do**
    `// Textually Anchored Semantic Core`
    Extract global and spatial tokens $\{(\mathbf{c}_i^v, \mathbf{Z}_i^v)\}_{v\in\{q,g\}} \leftarrow f_\theta(\{x_i^q, x_i^g\})$
    Compute textual references $\mathbf{u}_i \leftarrow p_t(\mathbf{e}_i)$ and semantic features $\mathbf{s}_i^v \leftarrow p_s(\mathbf{c}_i^v)$
    Compute semantic alignment loss $\mathcal{L}_{\text{sem}}$ using Eq. 3
    `// Reliability-Aware Structural Learning`
    foreach $v \in \{q, g\}$ **do**
      Compute anchored assignments $\mathbf{A}_i^v$ and regional contents $\mathbf{H}_i^v$ using Eqs. 4 and 5
      Compute structural representation $\mathbf{r}_i^v \leftarrow \text{RSSR}(\mathbf{H}_i^v)$ using Eq. 6;
    Draw E geometry-preserving appearance variations for each paired observation
    Compute soft regional correspondences from $(\mathbf{H}_i^q, \mathbf{H}_i^g)$ and hold them fixed across variations
    Compute regional margins $\{m_{i,k}^{v,e}\}$ and reliability targets $\{w_{i,k}^v\}$ using Eqs. 7 and 8
    Predict regional reliability $\hat{\mathbf{w}}_i^v \leftarrow \sigma(G_\psi(\mathbf{H}_i^v))$ and compute $\mathcal{L}_{\text{reg}}$ using Eq. 9;
    `// Joint Optimization`
    Form retrieval descriptors $\mathbf{f}_i^v \leftarrow \text{Concat}(\mathbf{s}_i^v, \alpha\mathbf{r}_i^v)$ and batch representations $\mathbf{F}^v \leftarrow \{\mathbf{f}_i^v\}_{i\in\mathcal{B}}$
    Compute cross-view alignment loss $\mathcal{L}_{\text{align}}(\mathbf{F}^q, \mathbf{F}^g)$
    $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{align}} + \lambda_{\text{sem}}\mathcal{L}_{\text{sem}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}$
    Update Θ using $\nabla_\Theta \mathcal{L}_{\text{total}}$
return $\{f_\theta, p_s, p_r, \{\mathbf{w}_k\}_{k=1}^K, R_\omega\}$

Algorithm 2: Inference Pipeline of GeoCore

Require : Query set $\mathcal{Q} = \{x_i^q\}_{i=1}^{N_q}$; gallery set $\mathcal{G} = \{x_j^g\}_{j=1}^{N_g}$; trained image-side model $\{f_\theta, p_s, p_r, \{\mathbf{w}_k\}_{k=1}^K, R_\omega\}$
Ensure : Retrieved gallery identity for each query

foreach *image* $x_i^v \in \mathcal{Q} \cup \mathcal{G}$, $v \in \{q, g\}$ **do**
  Extract global and spatial tokens $(\mathbf{c}_i^v, \mathbf{Z}_i^v) \leftarrow f_\theta(x_i^v)$
  Compute semantic representation $\mathbf{s}_i^v \leftarrow p_s(\mathbf{c}_i^v)$
  Compute anchored assignments $\mathbf{A}_i^v$ and regional contents $\mathbf{H}_i^v$ using Eqs. 4 and 5
  Compute structural representation $\mathbf{r}_i^v \leftarrow \text{RSSR}(\mathbf{H}_i^v)$ using Eq. 6
  Form and normalize the retrieval descriptor $\mathbf{f}_i^v \leftarrow \text{norm}(\text{Concat}(\mathbf{s}_i^v, \alpha\mathbf{r}_i^v))$
foreach *query descriptor* $\mathbf{f}_i^q$ **do**
  Compute similarity scores $\ell_{ij} \leftarrow \langle \mathbf{f}_i^q, \mathbf{f}_j^g \rangle$ for all $\mathbf{f}_j^g \in \mathcal{G}$
  $\hat{j}_i \leftarrow \arg\max_j \ell_{ij}$
  return $y_{\hat{j}_i}$

A positive margin indicates that the corresponding representation ranks the correct location above the competing location.

Lemma 1 (Retrieval-Margin Decomposition). For the GeoCore descriptor in Eq. 13, its pairwise retrieval margin satisfies the formulation:

$$\Delta_f(i,j) = \frac{\Delta_s(i,j) + \alpha^2 \Delta_r(i,j)}{1+\alpha^2}. \tag{16}$$

Proof. For any query-gallery pair (i, j), the similarity between their fused descriptors is:

$$\begin{aligned} \left\langle \mathbf{f}_i^q, \mathbf{f}_j^g \right\rangle &= \frac{1}{1+\alpha^2} \left\langle [\mathbf{s}_i^q; \alpha\mathbf{r}_i^q], [\mathbf{s}_j^g; \alpha\mathbf{r}_j^g] \right\rangle \\ &= \frac{\left\langle \mathbf{s}_i^q, \mathbf{s}_j^g \right\rangle + \alpha^2 \left\langle \mathbf{r}_i^q, \mathbf{r}_j^g \right\rangle}{1+\alpha^2}. \end{aligned} \tag{17}$$

Subtracting the similarity of a competing gallery location from that of the correct gallery location immediately gives Eq. 16. □

Corollary 1 (Complementary Geographic Disambiguation). Suppose the semantic representation alone fails to distinguish the correct location from a competing location, i.e.,

$$\Delta_s(i,j) \le 0. \tag{18}$$

If the structural representation satisfies:

$$\Delta_r(i,j) > -\frac{\Delta_s(i,j)}{\alpha^2}, \tag{19}$$

then

$$\Delta_f(i,j) > 0, \tag{20}$$

and the fused GeoCore representation correctly ranks the positive location above the competing one.

Proof. Substituting Eq. 19 into Eq. 16 gives:

$$\Delta_s(i,j) + \alpha^2 \Delta_r(i,j) > 0. \tag{21}$$

Since $1 + \alpha^2 > 0$, it follows directly that $\Delta_f(i,j) > 0$. □

Corollary 1 provides a retrieval-level interpretation of complementarity. Structural evidence is useful not merely because it captures information different from semantic evidence, but because it can provide additional positive discrimination against locations that remain ambiguous in the semantic space. This motivates combining coarse transferable semantics with fine-grained regional structure in GeoCore.

Rotation Symmetry of Structural Refinement. Finally, we analyze the rotation-symmetric property of RSSR. The structural refinement term is defined as follows:

$$\mathcal{T}(\mathbf{H}) = \frac{1}{|\mathcal{R}|} \sum_{\pi \in \mathcal{R}} R_\omega(P_\pi \mathbf{H}), \tag{22}$$

where $\mathcal{R} = \{0°, 90°, 180°, 270°\}$ is the quarter-turn rotation group and P_π denotes the corresponding permutation of the anchor-grid positions.

Lemma 3 (Rotation Symmetry of RSSR). For any $\rho \in \mathcal{R}$,

$$\mathcal{T}(P_\rho \mathbf{H}) = \mathcal{T}(\mathbf{H}). \tag{23}$$

Consequently, the complete RSSR representation

$$\mathbf{r}(\mathbf{H}) = \frac{1}{K} \sum_{k=1}^{K} \mathbf{h}_k + \gamma \mathcal{T}(\mathbf{H}) \tag{24}$$

is invariant to quarter-turn permutations of the regional anchor grid.

Proof. Applying a rotation ρ gives:

$$\begin{aligned} \mathcal{T}(P_\rho \mathbf{H}) &= \frac{1}{|\mathcal{R}|} \sum_{\pi \in \mathcal{R}} R_\omega \left(P_\pi P_\rho \mathbf{H}\right) \\ &= \frac{1}{|\mathcal{R}|} \sum_{\pi \in \mathcal{R}} R_\omega \left(P_{\pi \circ \rho} \mathbf{H}\right). \end{aligned} \tag{25}$$

Because $\mathcal{R}$ forms a finite rotation group, right multiplication by any fixed $\rho \in \mathcal{R}$ only permutes its elements, i.e.,

$$\{\pi \circ \rho : \pi \in \mathcal{R}\} = \mathcal{R}. \tag{26}$$

Therefore,

$$\mathcal{T}(P_\rho \mathbf{H}) = \frac{1}{|\mathcal{R}|} \sum_{\pi' \in \mathcal{R}} R_\omega(P_{\pi'} \mathbf{H}) = \mathcal{T}(\mathbf{H}). \tag{27}$$

The mean regional term is also invariant to any permutation of region indices,

$$\frac{1}{K}\sum_{k=1}^{K}(P_\rho \mathbf{H})_k = \frac{1}{K}\sum_{k=1}^{K}\mathbf{h}_k. \tag{28}$$

Hence, $\mathbf{r}(P_\rho \mathbf{H}) = \mathbf{r}(\mathbf{H})$. □

It is worth noting that Lemma 3 establishes rotation symmetry with respect to quarter-turn permutations of an already encoded regional grid. It does not imply end-to-end invariance to arbitrary rotations of the input image. RSSR therefore reduces orientation sensitivity in regional structural modeling without discarding the spatial organization required for geographic discrimination.

Cross-environment Regional Reliability. GeoCore further requires discriminative evidence to remain reliable under appearance variations. For a regional representation k, Cross-Environment Regional Reliability aggregates its discriminative margins $\{m_{i,k}^{v,e}\}_{e=1}^{E}$ as

$$\rho_{i,k}^{v} = -\tau_e \log\left(\frac{1}{E}\sum_{e=1}^{E}\exp\left(-\frac{m_{i,k}^{v,e}}{\tau_e}\right)\right), \tag{29}$$

where $\tau_e > 0$ controls the softness of the aggregation.

Lemma 2 (Soft Worst-Environment Margin). Let

$$m_{\min} = \min_{e\in\{1,\dots,E\}} m_{i,k}^{v,e}. \tag{30}$$

Then the reliability statistic in Eq. 29 satisfies

$$m_{\min} \le \rho_{i,k}^{v} \le m_{\min} + \tau_e \log E. \tag{31}$$

Consequently,

$$\lim_{\tau_e \to 0} \rho_{i,k}^{v} = \min_{e} m_{i,k}^{v,e}. \tag{32}$$

Proof. Since $m_{i,k}^{v,e} \ge m_{\min}$ for every environment e,

$$\exp\left(-\frac{m_{i,k}^{v,e}}{\tau_e}\right) \le \exp\left(-\frac{m_{\min}}{\tau_e}\right). \tag{33}$$

Therefore,

$$\frac{1}{E}\sum_{e=1}^{E}\exp\left(-\frac{m_{i,k}^{v,e}}{\tau_e}\right) \le \exp\left(-\frac{m_{\min}}{\tau_e}\right), \tag{34}$$

which implies

$$\rho_{i,k}^{v} \ge m_{\min}. \tag{35}$$

On the other hand, there exists at least one $e^\star$ such that $m_{i,k}^{v,e^\star} = m_{\min}$. Hence,

$$\frac{1}{E}\sum_{e=1}^{E}\exp\left(-\frac{m_{i,k}^{v,e}}{\tau_e}\right) \ge \frac{1}{E}\exp\left(-\frac{m_{\min}}{\tau_e}\right), \tag{36}$$

which gives

$$\rho_{i,k}^{v} \le m_{\min} + \tau_e \log E. \tag{37}$$

Combining the two inequalities proves Eq. 31. Since the approximation gap is upper-bounded by $\tau_e \log E$, Eq. 32 follows as $\tau_e \to 0$. □

Lemma 2 shows that the reliability statistic provides a smooth approximation to the worst-environment regional discriminative margin. Therefore, a high reliability target requires a regional cue to remain discriminative across appearance variations rather than achieving a large margin only under a favorable condition. This property provides a theoretical interpretation of the *persistence* targeted by Cross-environment Regional Reliability.

Robust Retrieval with Complementary Evidence. We next connect complementary representation learning with cross-environment discrimination. Let $\mathcal{E}$ denote a finite collection of environmental variations. For each $e \in \mathcal{E}$, define the semantic and structural retrieval margins as

$$\Delta_s^e(i,j) = \left\langle \mathbf{s}_i^{q,e}, \mathbf{s}_i^{g^+,e} \right\rangle - \left\langle \mathbf{s}_i^{q,e}, \mathbf{s}_j^{g^-,e} \right\rangle, \tag{38}$$

$$\Delta_r^e(i,j) = \left\langle \mathbf{r}_i^{q,e}, \mathbf{r}_i^{g^+,e} \right\rangle - \left\langle \mathbf{r}_i^{q,e}, \mathbf{r}_j^{g^-,e} \right\rangle. \tag{39}$$

We further define their worst-environment margins as

$$\underline{\Delta}_s(i,j) = \min_{e\in\mathcal{E}} \Delta_s^e(i,j), \qquad \underline{\Delta}_r(i,j) = \min_{e\in\mathcal{E}} \Delta_r^e(i,j). \tag{40}$$

Theorem 1 (Robust Retrieval Margin of GeoCore). For any query i and competing location j, the fused GeoCore descriptor satisfies

$$\min_{e\in\mathcal{E}} \Delta_f^e(i,j) \geq \frac{\underline{\Delta}_s(i,j) + \alpha^2 \underline{\Delta}_r(i,j)}{1+\alpha^2}. \tag{41}$$

Consequently, if

$$\underline{\Delta}_s(i,j) + \alpha^2 \underline{\Delta}_r(i,j) > 0, \tag{42}$$

then the correct gallery location is ranked above location j under every environment $e \in \mathcal{E}$.

Proof. According to Lemma 1, for each environment e,

$$\Delta_f^e(i,j) = \frac{\Delta_s^e(i,j) + \alpha^2 \Delta_r^e(i,j)}{1+\alpha^2}. \tag{43}$$

Since

$$\Delta_s^e(i,j) \geq \underline{\Delta}_s(i,j), \qquad \Delta_r^e(i,j) \geq \underline{\Delta}_r(i,j), \tag{44}$$

for every $e \in \mathcal{E}$, we obtain

$$\Delta_f^e(i,j) \geq \frac{\underline{\Delta}_s(i,j) + \alpha^2 \underline{\Delta}_r(i,j)}{1+\alpha^2}. \tag{45}$$

Taking the minimum over e gives Eq. 41. If Eq. 42 holds, the right-hand side of Eq. 41 is positive, and hence $\Delta_f^e(i,j) > 0$ for all $e \in \mathcal{E}$. □

Theorem 1 establishes a sufficient condition for robust pairwise retrieval over the considered environmental variations. In particular, complementary structural evidence can compensate for weak semantic discrimination, while persistent structural margins improve the lower bound of the fused retrieval margin. Lemma 2 further shows that the reliability target used by GeoCore is a region-level smooth surrogate for worst-environment discrimination. Together, these results explain why GeoCore emphasizes evidence that is both *complementary* across representations and *persistent* across appearance variations.

Contrastive Alignment. We employ bidirectional InfoNCE for both cross-view retrieval and cross-modal alignment. Given two batches of normalized embeddings $\mathbf{A} = \{\mathbf{a}_i\}_{i=1}^B$ and $\mathbf{B} = \{\mathbf{b}_i\}_{i=1}^B$, we define the formulation as:

$$\begin{aligned} \mathcal{C}(\mathbf{A},\mathbf{B}) &= \frac{1}{2}\left(\ell_{\mathbf{A}\to\mathbf{B}} + \ell_{\mathbf{B}\to\mathbf{A}}\right), \\ \ell_{\mathbf{A}\to\mathbf{B}} &= -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\mathbf{a}_i^\top \mathbf{b}_i/\tau)}{\sum_{j=1}^{B}\exp(\mathbf{a}_i^\top \mathbf{b}_j/\tau)}, \end{aligned} \tag{46}$$

where τ is a learnable temperature. Accordingly, the final cross-view descriptors are aligned by $\mathcal{L}_{\text{align}} = \mathcal{C}(\mathbf{F}^q, \mathbf{F}^g)$. The same objective anchors the query and gallery semantic representations to the corresponding textual embeddings:

$$\mathcal{L}_{\text{sem}} = \mathcal{C}(\mathbf{S}^q, \mathbf{U}) + \mathcal{C}(\mathbf{S}^g, \mathbf{U}), \tag{47}$$

where $\mathbf{F}^q$ and $\mathbf{F}^g$ are the final GeoCore descriptors, while $\mathbf{S}^q$, $\mathbf{S}^g$, and $\mathbf{U}$ denote the query, gallery, and textual semantic embeddings, respectively.

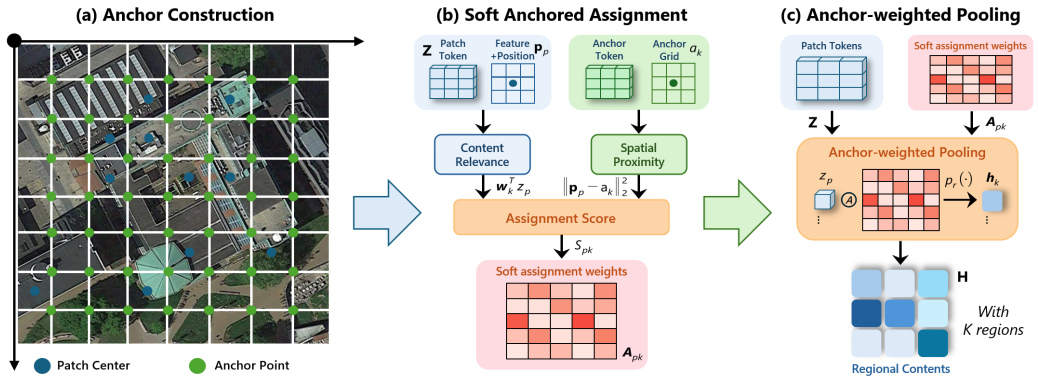


Figure 6: The overview of Spatially Anchored Regionalization.

A.1.3 Details of Spatial Anchors in Spatially Anchored Regionalization

As shown in Figure 6, SAR consists of three stages: constructing fixed spatial anchors, assigning patch tokens to anchors through content-spatial affinities, and aggregating them via anchor-weighted pooling to obtain spatially organized regional representations.

Anchor Construction. Spatially Anchored Regionalization (SAR) introduces a fixed spatial reference frame to organize patch-level features into consistently indexed regions. Given $K = n^2$ regions, we uniformly place K anchors on the normalized image plane $[0,1]^2$. For the anchor located at the r-th row and c-th column of an $n \times n$ grid, where $r, c \in \{0, \dots, n-1\}$, its spatial coordinate is defined as:

$$\mathbf{a}_{r,c} = \left[\frac{c + \frac{1}{2}}{n}, \frac{r + \frac{1}{2}}{n} \right]^\top \in [0,1]^2. \tag{48}$$

The corresponding region index can be written as $k = rn + c + 1$. In our default configuration, $K = 25$, resulting in a 5×5 anchor grid. The anchor coordinates are fixed throughout training and inference, providing a common spatial reference across images without introducing additional learnable positional parameters.

Similarly, suppose the visual encoder produces a spatial feature map of size $H_p \times W_p$. For a patch token located at row u and column v, its normalized center coordinate is:

$$\mathbf{p}_{u,v} = \left[\frac{v + \frac{1}{2}}{W_p}, \frac{u + \frac{1}{2}}{H_p} \right]^\top. \tag{49}$$

Using normalized coordinates makes the spatial assignment independent of the original image resolution and directly compatible with different visual backbones and patch-grid sizes.

Soft Anchored Assignment. SAR assigns spatial tokens to K anchored regions by jointly considering learnable content affinity and spatial proximity. Unlike hard grid partitioning, the anchors provide a soft spatial prior rather than fixed boundaries, allowing region assignments to adapt to scene content while maintaining spatial coherence. This design relates to appearance-aware superpixels (Achanta et al., 2012), differentiable superpixel sampling (Jampani et al., 2018), and recent region-level representations (Shlapentokh-Rothman et al., 2024).

Anchor-Weighted Pooling. Regional features are obtained by normalized affinity-weighted pooling followed by a learnable projection. The resulting regions are stacked according to their fixed anchor-grid indices, forming a consistently ordered representation across images and viewpoints for subsequent structural refinement and reliability learning.

Why Spatial Anchoring? The anchor design serves two purposes: (1) It provides a consistent spatial reference for region indexing. Purely content-based grouping does not explicitly associate

region indices with spatial positions, making it difficult to capture layout differences between semantically similar locations. Fixed anchors associate each region with a persistent spatial reference, while $\mathbf{w}_k$ and A_{pk} allow its content to adapt to individual scenes. This spatial indexing does not imply direct regional correspondence across heterogeneous views. (2) It preserves the grid topology required by Rotation-Symmetric Structural Refinement (RSSR). A consistent association between region indices and anchor positions allows P_π to represent quarter-turn rotations of the regional grid rather than arbitrary feature permutations. RSSR can thus encode spatial arrangements relevant to location discrimination while remaining invariant to these discrete grid rotations.

SAR differs from conventional grid pooling, which assigns patches to predefined spatial cells. SAR fixes the anchor positions but determines regional membership through differentiable affinities that combine content relevance and spatial proximity. It therefore combines a consistent spatial organization with adaptive regional content, providing localized evidence for structural modeling and cross-environment reliability learning.

A.1.4 Details of Soft Regional Correspondence in Cross-Environment Regional Reliability

To account for imperfect spatial alignment across heterogeneous viewpoints, we compute a soft correspondence distribution between regional representations instead of enforcing one-to-one matching. For region k in view v, the correspondence weight to region l in the complementary view $\bar{v}$ is estimated from the original, unperturbed regional features:

$$Q^v_{i,kl} = \frac{\exp\left(\left\langle \mathbf{h}^v_{i,k}, \mathbf{h}^{\bar{v}}_{i,l} \right\rangle / \tau_c\right)}{\sum_{l'=1}^{K} \exp\left(\left\langle \mathbf{h}^v_{i,k}, \mathbf{h}^{\bar{v}}_{i,l'} \right\rangle / \tau_c\right)}, \tag{50}$$

where τ_c controls the sharpness of the correspondence distribution. The resulting matrix $\mathbf{Q}^v_i \in \mathbb{R}^{K\times K}$ provides a soft cross-view correspondence for each regional token. Unlike hard region matching, this formulation allows multiple candidate regions to contribute according to their visual similarity, thereby accommodating viewpoint-induced spatial misalignment.

The complete procedure of Cross-Environment Regional Reliability (CRR) is summarized in Algorithm 3. CRR first estimates cross-view correspondence from the original regional representations, and then evaluates whether the corresponding regional evidence remains geographically discriminative under multiple geometry-preserving appearance variations. The resulting environment-wise margins are aggregated into soft reliability targets, which supervise a lightweight regional reliability predictor.

Given an environmental variation e, the correspondence-weighted positive similarity used in the regional discriminative margin is computed as

$$\mathrm{sim}_{\mathrm{soft}}\left(\mathbf{h}^{v,e}_{i,k}, \mathbf{H}^{\bar{v},e}_i\right) = \sum_{l=1}^{K} Q^v_{i,kl} \left\langle \mathbf{h}^{v,e}_{i,k}, \mathbf{h}^{\bar{v},e}_{i,l} \right\rangle. \tag{51}$$

Importantly, $\mathbf{Q}^v_i$ is estimated once from the original regional representations and kept fixed across all environmental variations. This decouples cross-view correspondence estimation from appearance perturbations, so variations in the resulting margin primarily reflect changes in regional discriminability rather than unstable matching assignments.

For each region and appearance variation, CRR measures geographic discrimination as the margin between its soft similarity to the matched location and its strongest similarity to competing locations. The margins across variations are aggregated through a temperature-controlled soft minimum, emphasizing challenging environments with low discrimination. The aggregated margin is then converted into a soft reliability target.

A reliability predictor estimates these targets from original regional representations using binary cross-entropy with detached targets. Fixing cross-view correspondences across variations makes the supervision primarily reflect changes in regional discrimination induced by appearance variations. CRR serves as an auxiliary training objective, encouraging persistent geographic evidence without regional reweighting or additional matching operations at inference.

Algorithm 3: Cross-environment Regional Reliability (CRR)

Require: Regional representations $\{\mathbf{H}_i^q, \mathbf{H}_i^g\}_{i=1}^B$; number of appearance variations E
Ensure : Reliability loss $\mathcal{L}_{\text{reg}}$

Apply E geometry-preserving appearance variations and obtain $\{\mathbf{H}_i^{v,e}\}_{e=1}^E$ for $v \in \{q, g\}$
for $v \in \{q, g\}$ **do**
  Let $\bar{v}$ denote the complementary view
  for $i = 1, \ldots, B$ **do**
    Compute $\mathbf{Q}_i^v$ from the original pair $(\mathbf{H}_i^v, \mathbf{H}_i^{\bar{v}})$
    Fix $\mathbf{Q}_i^v$ across all environmental variations
    for $k = 1, \ldots, K$ **do**
      for $e = 1, \ldots, E$ **do**
        $s_{i,k}^{v,e,+} \leftarrow \text{sim}_{\text{soft}}(\mathbf{h}_{i,k}^{v,e}, \mathbf{H}_i^{\bar{v},e})$
        $s_{i,k}^{v,e,-} \leftarrow \max_{\substack{j \neq i \\ 1 \le l \le K}} \left\langle \mathbf{h}_{i,k}^{v,e}, \mathbf{h}_{j,l}^{\bar{v},e} \right\rangle$
        $m_{i,k}^{v,e} \leftarrow s_{i,k}^{v,e,+} - s_{i,k}^{v,e,-}$
      $\rho_{i,k}^v \leftarrow \text{SoftMin}_{\tau_e}(\{m_{i,k}^{v,e}\}_{e=1}^E)$
      $w_{i,k}^v \leftarrow \sigma((\rho_{i,k}^v - \mu)/\tau_t)$
    $\hat{\mathbf{w}}_i^v \leftarrow \sigma(G_\psi(\mathbf{H}_i^v))$
$\mathcal{L}_{\text{reg}} \leftarrow \text{BCE}(\text{sg}(\mathbf{w}), \hat{\mathbf{w}})$
return $\mathcal{L}_{\text{reg}}$

A.2 More Implementation Details

A.2.1 Details on Benchmarks

University-1652. University-1652 contains drone, satellite, and ground-view images of 1,652 buildings. For the UAV–satellite setting, the training split contains 701 locations with 37,854 drone images and 701 satellite images. During evaluation, drone-to-satellite retrieval uses 37,855 drone queries and 951 satellite gallery images, while satellite-to-drone retrieval uses 701 satellite queries and 51,355 drone gallery images. The training and test locations are disjoint.

SUES-200. SUES-200 contains 200 locations observed from satellite and UAV platforms at four altitudes: 150 m, 200 m, 250 m, and 300 m. The training split contains 120 locations with 24,000 UAV images and 120 satellite images. The remaining 80 locations are used as queries at test time, while all 200 locations form the gallery. This results in 16,000 UAV queries, 40,000 UAV gallery images, 80 satellite queries, and 200 satellite gallery images.

DenseUAV. DenseUAV is a low-altitude UAV geo-localization benchmark. Its training split contains 2,256 locations with 6,768 UAV and 13,536 satellite images. The query split contains 777 locations with 2,331 UAV and 4,662 satellite images, while the gallery contains 3,033 locations with 9,099 UAV and 18,198 satellite images.

CVUSA. CVUSA contains aligned panoramic street-view and satellite image pairs across the United States. Following the standard split, we use 35,532 pairs for training and 8,884 pairs for evaluation.

CVACT. CVACT contains aligned street-view panoramas and satellite images collected in Canberra, Australia. It provides 35,532 training pairs and 8,884 validation pairs, together with a larger test split of 92,802 pairs for large-scale retrieval evaluation.

Unless otherwise specified, we follow the official training and evaluation splits of each benchmark.

A.2.2 Details on Hyperparameters

UAV-view Settings. For University-1652→SUES-200, the patch embedding layer and transformer blocks 0–7 are frozen, while the remaining four DINOv2 blocks and GeoCore modules are optimized with learning rates of 1.15×10^{-5} and 2.25×10^{-4}, respectively. We train the model for 9 epochs

Table 6: Statistics of the UAV-based CVGL benchmarks used in our experiments. Only the UAV and satellite modalities are reported.

Dataset	Split / View	# Locations	# Images
University-1652	Train / Drone	701	37,854
	Train / Satellite	701	701
	Test / Drone Query	701	37,855
	Test / Satellite Gallery	951	951
SUES-200	Train / Drone	120	24,000
	Train / Satellite	120	120
	Test / Drone Query	80	16,000
	Test / Satellite Gallery	200	200
DenseUAV	Train / UAV	2,256	6,768
	Train / Satellite	2,256	13,536
	Query / UAV	777	2,331
	Gallery / Satellite	3,033	18,198

Table 7: Statistics of the ground-to-satellite CVGL benchmarks.

Dataset	Train Pairs	Val Pairs	Test Pairs
CVUSA	35,532	–	8,884
CVACT	35,532	8,884	92,802

using a batch size of 16, 0.75 warm-up epochs, a cosine-scheduler horizon of 34 epochs, and weight decay 0.01.

Ground-view Settings. For CVACT→CVUSA, the pretrained backbone and GeoCore modules are assigned learning rates of 1.75×10^{-5} and 3.0×10^{-4}, respectively. We train the model for 18 epochs using a batch size of 64, 0.9 warm-up epochs, a cosine-scheduler horizon of 34 epochs, and weight decay 0.005. After epoch 9, the visual backbone is frozen and the learning rate of the GeoCore modules is reduced to 1.2×10^{-4} for stable late-stage optimization.

For both settings, we set the spatial temperature τ_s and assignment temperature τ_a to 0.25 and 0.10, respectively. The correspondence temperature τ_c is set to 0.07, while the environmental aggregation temperature τ_e and reliability-target temperature τ_t are both set to 0.10. We use two environmental variations ($E = 2$) and set the reference margin μ to 0 in CRR.

A.2.3 Details on Network Designs

We provide the details of network designs in the trainable modules: the *Semantic Projector* $p_s(\cdot)$, the *Text Projector* $p_t(\cdot)$, the *Non-linear Transformation* $R_\omega(\cdot)$ in RSSR, and the *Reliability Predictor* $G_\psi(\cdot)$. Let D denote the final descriptor dimension. The Semantic and Structural Cores each produce a $D/2$-dimensional representation.

Semantic Projector $p_s(\cdot)$. The Semantic Projector maps the global `[CLS]` token $\mathbf{c}^v \in \mathbb{R}^{d_v}$ from view $v \in \{q, g\}$ into the semantic embedding space:

$$\mathbf{s}^v = \mathrm{Norm}(p_s(\mathbf{c}^v)), \qquad p_s(\mathbf{x}) = \mathbf{W}_s\,\mathrm{LN}(\mathbf{x}) + \mathbf{b}_s, \tag{52}$$

where $\mathrm{Norm}(\cdot)$ denotes ℓ_2 normalization. The linear layer projects the backbone dimension d_v to $D/2$. Its parameters are shared between the query and gallery branches.

Text Projector $p_t(\cdot)$. Textual descriptions are encoded offline by the frozen CLIP text encoder. Given a textual embedding $\mathbf{t} \in \mathbb{R}^{d_t}$, the trainable Text Projector produces

$$\mathbf{u} = \mathrm{Norm}(p_t(\mathbf{t})), \qquad p_t(\mathbf{t}) = \mathbf{W}_t\,\mathrm{LN}(\mathbf{t}) + \mathbf{b}_t. \tag{53}$$

Thus, $p_s(\cdot)$ and $p_t(\cdot)$ map visual and textual features into the same $D/2$-dimensional space. Their cross-modal alignment is optimized using $\mathcal{L}_{\mathrm{sem}}$.

Non-linear Transformation $R_\omega(\cdot)$. The Non-linear Transformation in RSSR operates on the ordered regional representation $\mathbf{H}^v \in \mathbb{R}^{K\times D/2}$. Since the region order is defined by the fixed anchor grid, $R_\omega(\cdot)$ models spatial relationships while retaining the regional arrangement.

We first apply a residual MLP along the region axis:

$$\widetilde{\mathbf{H}}^v = (\mathbf{H}^v)^\top + f_{\mathrm{mix}}\big((\mathbf{H}^v)^\top\big), \tag{54}$$

where $f_{\mathrm{mix}}(\cdot)$ consists of LayerNorm, two linear layers, GELU, and dropout. The mixed features are compressed into S ordered summaries, concatenated, and projected to $D/2$ dimensions:

$$R_\omega(\mathbf{H}^v) = \mathrm{Norm}\Big[\mathbf{W}_o\,\mathrm{Dropout}\Big(\mathrm{LN}\Big(\mathrm{vec}\,\Big(f_{\mathrm{sum}}(\widetilde{\mathbf{H}}^v)\Big)\Big)\Big)\Big]. \tag{55}$$

We use a hidden dimension of 32, $S = 4$, and a dropout rate of 0.1.

To reduce sensitivity to quarter-turn rotations, the shared $R_\omega(\cdot)$ processes four rotated regional grids:

$$\mathbf{r}^v = \mathrm{Norm}\left[\frac{1}{K}\sum_{k=1}^{K}\mathbf{h}_k^v + \frac{\gamma}{|\mathcal{R}|}\sum_{\pi\in\mathcal{R}} R_\omega(P_\pi\mathbf{H}^v)\right], \tag{56}$$

where $\mathcal{R} = \{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$, P_π denotes the corresponding grid permutation, and γ is a learnable residual scale.

Reliability Predictor $G_\psi(\cdot)$. The Reliability Predictor estimates whether individual regions remain discriminative under environmental variations. The semantic representation is prepended to the regional sequence to provide global context:

$$\mathbf{X}^v = [\mathbf{s}^v; \mathbf{h}_1^v; \ldots; \mathbf{h}_K^v], \qquad \widetilde{\mathbf{X}}^v = \mathrm{Transformer}_\psi(\mathbf{X}^v). \tag{57}$$

Each contextualized regional token is mapped to a scalar reliability prediction:

$$\widehat{w}_k^v = \sigma(G_\psi(\mathbf{H}^v)_k) = \sigma\big(\mathbf{w}_\psi^\top\,\mathrm{LN}(\widetilde{\mathbf{x}}_k^v) + b_\psi\big). \tag{58}$$

The predictor uses a two-layer Transformer block with four attention heads, a feed-forward dimension of D, GELU activation, and 0.1 dropout. It is optimized by the cross-environment soft targets through $\mathcal{L}_{\mathrm{reg}}$.

The environmental variations and Reliability Predictor are required only during training. At inference, the Structural Core is constructed directly from SAR and RSSR without environmental augmentation or explicit regional reweighting.

A.2.4 Details on the Generation of Textual Descriptions

Offline Text Generation. GeoCore assigns each source-domain training location a compact description of its cross-view-invariant semantics. We generate descriptions offline with GLM-4.6V-Flash and keep them fixed during training. Generation uses only paired source-training images; no target-domain images or annotations are involved.

For UAV-based CVGL, we uniformly sample three UAV observations from each University-1652 location and pair them with its satellite image. For ground-view CVGL, we use one 360° street-view panorama and its matched satellite image from CVACT.

Retrieval-Oriented Prompt Design. Rather than captioning each image independently, we prompt the vision-language model to produce one location-level description. It first identifies persistent scene elements and coarse spatial relations, consolidating evidence across UAV viewpoints when applicable. It then verifies these cues against the satellite image and removes unsupported or view-specific evidence. Finally, it summarizes the verified cues according to their likely value for distinguishing locations.

The prompt excludes transient or appearance-dependent information, such as color, illumination, weather, shadows, vehicles, pedestrians, facade details, and readable text. It also avoids camera pose, image orientation, exact counts, precise geometry, and geographic names. The resulting descriptions focus on stable elements, such as buildings, road topology, open spaces, vegetation, and water—and their coarse spatial relations.

Structured Semantic Output. The generator returns a JSON record containing view-level evidence, cross-view verification, uncertainty, and the final semantics. We retain three compact fields for supervision: distinctive elements $\mathcal{E}_i$, distinctive relations $\mathcal{R}_i$, and coarse scene semantics $\mathcal{O}_i$. These fields are serialized deterministically as

$$\begin{aligned} t_i = \texttt{Scene elements: } & \mathcal{E}_i\,. \\ \texttt{Coarse spatial relations: } & \mathcal{R}_i\,. \\ \texttt{Overall scene: } & \mathcal{O}_i\,. \end{aligned} \tag{59}$$

The prompt keeps each field concise so that t_i fits within the CLIP text encoder's context length. We encode each textual description once with the frozen CLIP text encoder and cache its ℓ_2-normalized embedding $\mathbf{e}_i^t \in \mathbb{R}^{512}$. These embeddings provide training-time semantic supervision; neither the generator nor the descriptions are needed at inference.

UAV-Satellite Example.

UAV–Satellite Prompt

```
SYSTEM
You are an expert analyst for generalizable cross-view geo-localization
    . The input contains multiple UAV images and one satellite image of
     the same location.

Generate one location-level semantic description rather than
    independent image captions.

USER
Input order:
[UAV View 1] [UAV View 2] [UAV View 3] [Satellite View]

1. Multi-view UAV consensus
Identify persistent scene elements and coarse spatial relations
    supported across the UAV viewpoints.

2. Satellite verification
Retain only cues reasonably supported by the satellite image.

3. Retrieval-oriented abstraction
Prioritize distinctive elements and coarse relations useful for
    identifying the location.

Ignore weather, illumination, color, transient objects, facade details,
     camera pose, orientation, precise geometry, and geographic names.

Return JSON with the fields:
distinctive_elements, distinctive_relations, coarse_scene_semantics,
    uncertainty, and confidence.
```

UAV–Satellite Output

```
{
  "distinctive_elements": [
    "railway corridor",
    "dense building complex",
    "landscaped green courtyard"
  ],
  "distinctive_relations": [
    "railway runs alongside the building complex",
    "green courtyard adjoins the main building"
  ],
  "coarse_scene_semantics": [
    "urban complex with railway and green space"
  ]
```

```
}
```

Street-Satellite Example.

Street–Satellite Prompt

```
SYSTEM
You are an expert analyst for generalizable cross-view geo-localization
    . The input contains one 360-degree street-view panorama and its
    matching satellite image.

Produce one orientation-invariant, location-level semantic description
    supported by both views.

USER
Input order:
[Street-view Panorama] [Satellite View]

1. Street-view observation
Identify stable road structure, land use, major scene elements, and
    coarse layout.

2. Satellite verification
Remove street-only cues and retain evidence that can also be verified
    from the overhead view.

3. Retrieval-oriented abstraction
Prioritize distinctive road topology, building and land-cover
    organization, and their coarse relations.

Ignore vehicles, signs, facade details, weather, colors, camera heading
    , exact counts, precise geometry, and geographic names.

Return the result as structured JSON with the fields:
distinctive_elements, distinctive_relations, coarse_scene_semantics,
    uncertainty, and confidence.
```

Street–Satellite Output

```
{
  "distinctive_elements": [
    "major road intersection",
    "triangular traffic islands",
    "tree-lined roads"
  ],
  "distinctive_relations": [
    "roads converge at a junction",
    "traffic islands separate directions"
  ],
  "coarse_scene_semantics": [
    "tree-lined urban road junction"
  ]
}
```

A.3 Extended Quantitative Results

A.3.1 Detailed Experimental Results of SUES-200

We provide detailed evaluations on SUES-200 across different UAV altitudes and adverse weather conditions to further examine the robustness of GeoCore under viewpoint-scale and appearance

Table 8: Robustness evaluation under adverse weather shifts on University-1652→SUES-200 at different UAV altitudes.

Methods	150m		200m		250m		300m	
	R@1	AP	R@1	AP	R@1	AP	R@1	AP
Fog-Rain								
Sample4Geo	26.63	33.78	35.95	42.65	41.67	48.23	45.33	51.87
DAC	61.48	66.38	72.15	75.85	78.68	81.46	80.62	83.20
CV-cities	65.10	70.34	75.17	79.08	80.33	83.52	79.85	83.30
InfoGeo	77.90	81.32	79.90	82.95	81.12	83.30	80.65	83.80
GeoCore	**82.70**$_{+4.80}$	**85.95**$_{+4.63}$	**86.75**$_{+6.85}$	**89.27**$_{+6.32}$	**87.93**$_{+6.81}$	**90.14**$_{+6.62}$	**87.33**$_{+6.68}$	**89.66**$_{+5.86}$
GeoCore†	82.60$_{+4.70}$	85.54$_{+4.22}$	84.93$_{+5.03}$	87.56$_{+4.61}$	86.25$_{+5.13}$	88.62$_{+5.10}$	85.28$_{+4.63}$	87.78$_{+3.98}$
Avg. Δ	**+4.75**	**+4.43**	**+5.94**	**+5.46**	**+5.97**	**+5.86**	**+5.65**	**+4.92**
Fog-Snow								
Sample4Geo	21.40	27.66	25.65	31.99	29.25	35.64	30.42	36.92
DAC	39.40	45.17	45.25	50.60	46.53	52.21	48.92	54.31
CV-cities	63.43	68.71	72.15	76.05	77.30	81.16	76.42	80.11
InfoGeo	74.60	78.92	76.63	80.16	77.50	81.52	76.50	80.00
GeoCore	80.35$_{+5.75}$	83.85$_{+4.93}$	83.78$_{+7.15}$	86.78$_{+6.62}$	84.85$_{+7.35}$	87.54$_{+6.02}$	84.30$_{+7.80}$	**87.00**$_{+6.89}$
GeoCore†	**81.25**$_{+6.65}$	**84.34**$_{+5.42}$	**84.35**$_{+7.72}$	**86.82**$_{+6.66}$	**86.03**$_{+8.53}$	**88.07**$_{+6.55}$	**84.68**$_{+8.18}$	86.97$_{+6.86}$
Avg. Δ	**+6.20**	**+5.17**	**+7.44**	**+6.64**	**+7.94**	**+6.29**	**+7.99**	**+6.88**
Rain-Snow								
Sample4Geo	35.55	41.77	44.50	50.38	50.27	55.95	52.90	58.22
DAC	59.78	64.62	67.40	71.57	74.05	77.72	74.65	78.37
CV-cities	75.60	79.54	85.00	87.53	89.77	91.68	90.57	92.42
InfoGeo	85.75	87.95	88.53	90.43	90.73	92.20	91.73	92.98
GeoCore	87.43$_{+1.68}$	90.00$_{+2.05}$	92.33$_{+3.80}$	93.97$_{+3.54}$	95.35$_{+4.62}$	96.39$_{+4.19}$	95.80$_{+4.07}$	96.75$_{+3.77}$
GeoCore†	**92.68**$_{+6.93}$	**93.95**$_{+6.00}$	**95.25**$_{+6.72}$	**96.11**$_{+5.68}$	**96.00**$_{+5.27}$	**96.77**$_{+4.57}$	**96.25**$_{+4.52}$	**96.97**$_{+3.99}$
Avg. Δ	**+4.31**	**+4.02**	**+5.26**	**+4.61**	**+4.94**	**+4.38**	**+4.29**	**+3.88**

Table 9: Comparison with recent state-of-the-art UAV-based CVGL methods on University-1652→SUES-200 at different UAV altitudes.

Methods	Venue	University→SUES (Avg)		150 m		200 m		250 m		300 m	
		R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP
MCCG	TCSVT'23	70.31	74.58	57.62	62.80	66.83	71.60	74.25	78.35	82.55	85.57
Sample4Geo	ICCV'23	82.03	85.12	70.05	74.93	80.68	83.90	87.35	89.72	90.03	91.91
DAC	TCSVT'24	87.65	89.80	76.65	80.56	86.45	89.00	92.95	94.18	94.53	95.45
CAMP	TGRS'24	88.34	90.49	78.90	82.38	86.83	89.28	91.95	93.63	95.68	96.65
MFRGN	ACM MM'24	78.28	81.75	65.90	70.93	74.28	78.38	85.05	87.61	87.88	90.06
Game4Loc	AAAI'25	86.75	89.07	79.03	82.50	86.03	88.37	90.25	92.08	91.70	93.32
MMGeo	ICCV'25	85.01	87.60	73.08	77.45	82.73	85.92	90.33	92.03	93.90	95.01
CVcities	JSTARS'24	90.27	91.77	82.73	85.60	90.08	91.63	94.35	95.12	93.90	94.74
InfoGeo	ICML'26	93.38	94.35	89.93	91.53	92.78	93.96	95.32	95.75	95.50	96.15
GeoCore	Ours	96.61$_{+3.23}$	97.20$_{+2.85}$	93.30$_{+3.37}$	94.08$_{+2.55}$	96.83$_{+4.05}$	97.52$_{+3.56}$	97.65$_{+2.33}$	98.15$_{+2.40}$	98.68$_{+3.00}$	99.05$_{+2.40}$
GeoCore†	Ours	**98.07**$_{+4.69}$	**98.45**$_{+4.10}$	**96.68**$_{+6.75}$	**97.27**$_{+5.74}$	**97.98**$_{+5.20}$	**98.35**$_{+4.39}$	**98.58**$_{+3.26}$	**98.90**$_{+3.15}$	**99.05**$_{+3.37}$	**99.27**$_{+2.62}$

shifts. We report multi-weather results on University-1652→SUES-200 and altitude-wise transfer performance from both University-1652 and DenseUAV. Across these settings, GeoCore consistently maintains strong cross-domain localization performance.

As shown in Tables 8–10, GeoCore consistently maintains strong performance across UAV altitudes, source domains, and adverse weather conditions. Under multi-weather evaluation, GeoCore outperforms previous methods across all altitude-weather combinations, with particularly clear gains under the challenging Fog-Snow condition, where the average R@1 improvement ranges from 6.20 to 7.99 points. For University-1652→SUES-200, the two GeoCore variants improve upon the previous best altitude-averaged results by 3.96 points in R@1 and 3.48 in AP on average, with the largest mean gains at 150 m. Similar trends are observed for DenseUAV→SUES-200, with average gains of 3.70/3.20 R@1/AP over the previous best, increasing to 5.41/4.38 at 150 m. GeoCore† further improves performance across all altitudes, reaching 99.93%/99.94% R@1/AP at 250 m and 100.00% at 300 m. These results demonstrate robust cross-domain generalization across substantial variations in UAV altitude and observation scale.

Table 10: Comparison with recent state-of-the-art UAV-based CVGL methods on DenseUAV→SUES-200 at different UAV altitudes.

Methods	Venue	DenseUAV→SUES (Avg)		150 m		200 m		250 m		300 m	
		R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP
MCCG	TCSVT'23	80.09	83.18	64.72	69.64	78.95	82.27	87.38	89.54	89.30	91.27
Sample4Geo	ICCV'23	78.94	81.86	67.00	71.54	84.18	86.62	84.93	87.41	79.65	81.86
DAC	TCSVT'24	86.44	88.70	82.43	85.19	87.23	89.35	89.70	91.52	86.38	88.72
CAMP	TGRS'24	86.79	88.93	72.30	76.12	89.63	91.59	90.53	92.35	94.68	95.66
MFRGN	ACM MM'24	67.22	71.14	61.50	66.09	62.30	66.97	72.33	75.58	72.73	75.90
Game4Loc	AAAI'25	67.19	71.40	53.20	58.94	66.53	70.85	73.85	77.25	75.18	78.55
MMGeo	ICCV'25	79.05	82.05	76.15	79.48	77.03	80.23	80.02	82.90	82.98	85.59
CVcities	JSTARS'24	89.67	91.34	80.20	83.24	89.73	91.54	93.83	94.87	94.90	95.65
InfoGeo	ICML'26	93.42	94.55	88.18	90.33	93.25	94.35	96.00	96.61	96.25	96.90
GeoCore	Ours	$\underline{95.05}_{+1.63}$	$\underline{96.12}_{+1.57}$	$\underline{89.78}_{+1.60}$	$\underline{91.48}_{+1.39}$	$\underline{95.20}_{+1.95}$	$\underline{96.16}_{+1.81}$	$\underline{97.43}_{+1.43}$	$\underline{98.15}_{+1.54}$	$\underline{97.80}_{+1.55}$	$\underline{98.46}_{+1.56}$
GeoCore†	Ours	$\mathbf{99.19}_{+5.77}$	$\mathbf{99.37}_{+4.82}$	$\mathbf{97.40}_{+9.22}$	$\mathbf{97.94}_{+7.61}$	$\mathbf{99.45}_{+6.20}$	$\mathbf{99.58}_{+5.23}$	$\mathbf{99.93}_{+3.93}$	$\mathbf{99.94}_{+3.33}$	$\mathbf{100.00}_{+3.75}$	$\mathbf{100.00}_{+3.10}$

Table 11: Ablation studies of different regional readout strategies in RSSR.

Components	University-1652→SUES-200		CVACT→CVUSA	
	R@1	AP	R@1	R@1%
Pooling	95.18	96.03	72.02	<u>98.12</u>
Ordered readout	<u>95.92</u>	<u>96.62</u>	<u>72.35</u>	98.05
RSSR + shuffled regions	95.14	96.02	71.42	98.01
RSSR (Full)	**96.61**	**97.20**	**72.60**	**98.20**

A.3.2 Effect of Rotation-Symmetric Structural Refinement

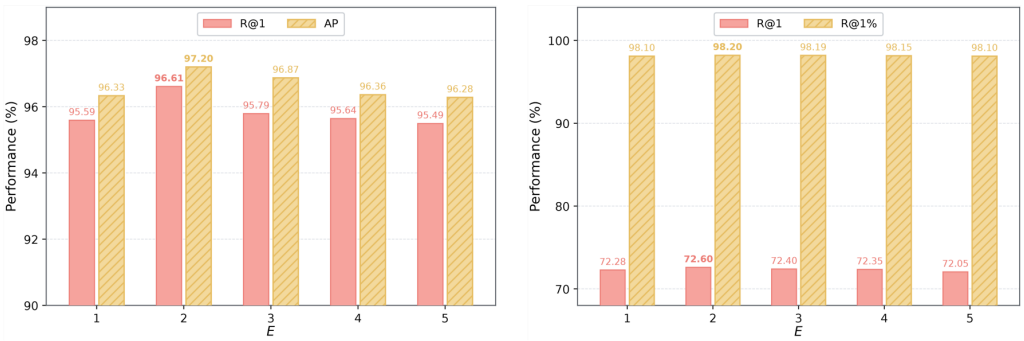
Table 11 evaluates how regional readout affects geographic discrimination in the structural branch. We compare mean pooling, ordered readout without rotation-symmetric aggregation, RSSR with shuffled regions, and the full RSSR. These variants examine the contributions of regional spatial organization and rotation-symmetric refinement.

Compared with mean pooling, ordered readout improves R@1 by 0.74 and 0.33 percentage points on University-1652→SUES-200 and CVACT→CVUSA, respectively, suggesting that regional arrangements provide useful location cues beyond aggregated regional content. Conversely, shuffling regions before RSSR reduces R@1 by 1.47 and 1.18 points relative to the full model. This degradation supports the importance of preserving spatial organization when learning structural evidence for location discrimination.

The full RSSR achieves the best results across all reported metrics, improving over ordered readout by 0.69/0.58 points in R@1/AP on UAV-based transfer and 0.25/0.15 points in R@1/R@1% on ground-view transfer. These results support modeling regional arrangements while reducing sensitivity to quarter-turn orientations of the anchor grid. By retaining spatial information beyond pooled regional content, RSSR strengthens the structural evidence available for cross-area geo-localization.

A.3.3 Effect of Cross-environment Regional Reliability

We study the number of geometry-preserving appearance variations E used to construct regional reliability targets. As shown in Figure 7, increasing E from 1 to 2 consistently improves performance in both transfer settings. In particular, $E = 2$ achieves the best results, reaching 96.61% R@1 and 97.20% AP on University-1652→SUES-200, and 72.60% R@1 and 98.20% R@1% on CVACT→CVUSA. These results indicate that multiple variations provide more reliable evidence of whether regional cues remain geographically discriminative under appearance changes. Further increasing E yields no additional benefit and causes a modest performance decline. A larger number of variations may introduce redundant or excessively challenging perturbations, weakening the stability of the resulting reliability targets. Nevertheless, performance remains relatively stable across the evaluated range, demonstrating that GeoCore is not highly sensitive to this hyperparameter. We therefore set $E = 2$, which provides sufficient environmental diversity while maintaining effective and computationally efficient reliability supervision.



(a) University-1652→SUES-200 (b) CVACT→CVUSA

Figure 7: Effect of the number of environmental variations E on Cross-Environment Regional Reliability conducted on UAV and Ground-view transfer scenarios.

Table 12: Descriptor-level analysis (%) of the Semantic and Structural Cores under both UAV and ground-view cross-dataset transfer settings.

Backbone	Core	University-1652→SUES-200				CVACT→CVUSA			
		R@1	R@5	R@10	AP	R@1	R@5	R@10	R@1%
DINOv2	Semantic Core	89.41	98.36	99.46	91.36	44.08	62.69	69.10	87.11
	Structural Core	90.54	98.17	99.54	92.27	71.81	86.41	90.45	98.12
	Full Model	**96.61**	**99.43**	**99.84**	**97.20**	**72.60**	**86.91**	**90.81**	**98.20**
DINOv3	Semantic Core	97.48	99.59	99.84	97.94	59.79	77.22	83.08	95.35
	Structural Core	83.80	96.84	99.05	86.56	56.37	74.18	80.74	94.55
	Full Model	**98.07**	**99.72**	**99.94**	**98.45**	**75.37**	**88.29**	**91.67**	**98.33**

A.3.4 Effect of Semantic Core and Structural Core

Table 12 analyzes the retrieval capability of the Semantic and Structural Cores independently. On University-1652→SUES-200, the Semantic Core already achieves strong performance, particularly with DINOv3, reaching 97.48% R@1. This is consistent with the characteristics of UAV imagery: although candidate locations can be geographically close, the elevated viewpoint captures a broader scene context, allowing high-level concepts such as building complexes, vegetation, and surrounding land cover to provide effective coarse geographic discrimination. The Structural Core further contributes complementary local organization, and combining both representations consistently yields the best performance across backbones.

A different behavior emerges on CVACT→CVUSA. Ground-view benchmarks contain densely and continuously sampled neighboring locations, whose street scenes often share nearly identical semantic compositions. Consequently, semantic evidence alone is considerably less discriminative, achieving only 44.08% and 59.79% R@1 with DINOv2 and DINOv3, respectively. Fine-grained structural evidence becomes particularly important in this setting, as local geometry, regional layout, and spatial organization provide additional cues for distinguishing semantically similar locations. Accordingly, the full GeoCore improves R@1 to 72.60% and 75.37%, substantially exceeding the corresponding single-core representations. These results support our coarse-to-fine perspective: semantic cues effectively narrow candidate locations, while structural cues provide the complementary evidence required to resolve residual geographic ambiguity.

A.3.5 Comparisons on Different Visual Backbones

We further evaluate GeoCore with different visual backbones to examine whether its generalization gains depend on a particular feature extractor. Specifically, we compare DINOv2 and DINOv3 baselines with GeoCore instantiated using the corresponding backbones under the same cross-dataset protocols on University-1652→SUES-200 and CVACT→CVUSA. As shown in Ta-

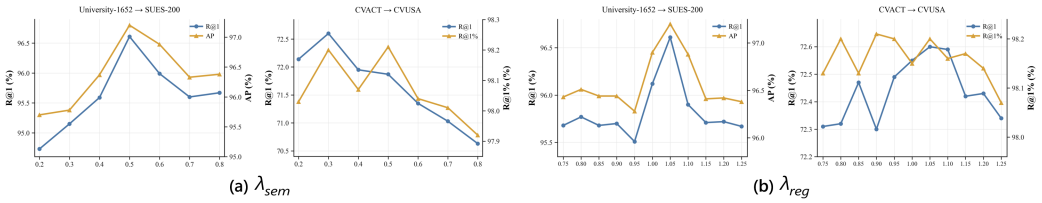


Figure 8: Additional Hyperparameter Studies on λ_{sem} and λ_{reg}.

ble 13, DINOv3 generally provides stronger baseline representations than DINOv2, particularly on University-1652→SUES-200, where R@1 increases from 91.28% to 95.63%.

More importantly, GeoCore consistently improves both backbones across all retrieval metrics. With DINOv2, R@1 increases by 5.33 and 4.05 points on University-1652→SUES-200 and CVACT→CVUSA, respectively, while DINOv3 yields gains of 2.44 and 5.97 points. Similar improvements are observed for R@5, R@10, AP, and R@1%. These results show that GeoCore provides consistent cross-domain benefits beyond backbone strength, indicating that its semantic and structural geographic-core learning contributes additional discriminative power across pretrained DINO encoders.

Table 13: Comparison (%) across different visual backbones on UAV and ground-view cross-dataset transfer settings.

Method	Backbone	University-1652→SUES-200				CVACT→CVUSA			
		R@1	R@5	R@10	AP	R@1	R@5	R@10	R@1%
Baseline	DINOv2	91.28	96.69	99.50	92.71	68.55	83.60	88.02	97.30
Baseline	DINOv3	95.63	98.80	99.34	96.30	69.40	85.47	90.11	97.71
GeoCore	DINOv2	96.61	99.43	99.84	97.20	72.60	86.91	90.81	98.20
GeoCore	DINOv3	98.07	99.72	99.94	98.45	75.37	88.29	91.67	98.33

A.3.6 Additional Hyperparameter Studies

We further investigate the sensitivity of GeoCore to the regularization weights λ_{sem} and λ_{reg}, which control semantic alignment and cross-environment regional reliability, respectively. As shown in Figure 8, we vary each coefficient while keeping the other fixed and evaluate two representative transfer settings, University-1652→SUES-200 and CVACT→CVUSA. This analysis examines whether GeoCore requires careful tuning across UAV- and ground-view scenarios.

For semantic regularization weight λ_{sem}, the preferred strength differs slightly across scenarios: University-1652→SUES-200 achieves the best performance at $\lambda_{\text{sem}} = 0.5$ with 96.61 R@1 and 97.20 AP, whereas CVACT→CVUSA favors a smaller value around 0.3, reaching 72.60 R@1. Nevertheless, performance remains relatively stable over a broad range, indicating limited sensitivity to semantic regularization. A similar trend is observed for λ_{reg}. The UAV setting peaks around $\lambda_{\text{reg}} = 1.05$ with 96.61 R@1 and 97.20 AP, while the ground-view setting varies only marginally across most tested values and reaches 72.60 R@1 and 98.20 R@1% near the optimum. Overall, GeoCore exhibits stable performance around moderate regularization strengths, suggesting that its gains do not rely on narrowly tuned hyperparameters.

A.4 Qualitative Results

A.4.1 Failure Case Analysis

Failure Case Analysis on UAV-based CVGL Figure 9 illustrates representative localization results at different flight altitudes. At 150 m, the query images cover a relatively limited spatial extent and exhibit stronger scale and perspective discrepancies with satellite references. Consequently, only partial buildings, roads, or vegetation layouts may be visible, making geographically similar

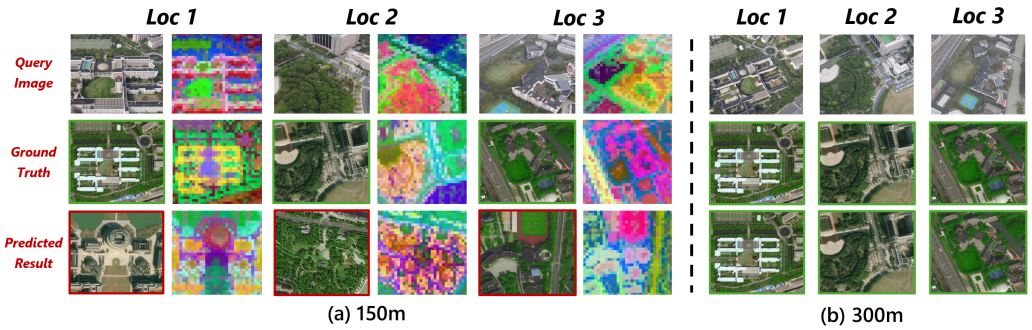


Figure 9: Visualization Analysis of Failed Case on UAV-based CVGL at 150m and 300m.

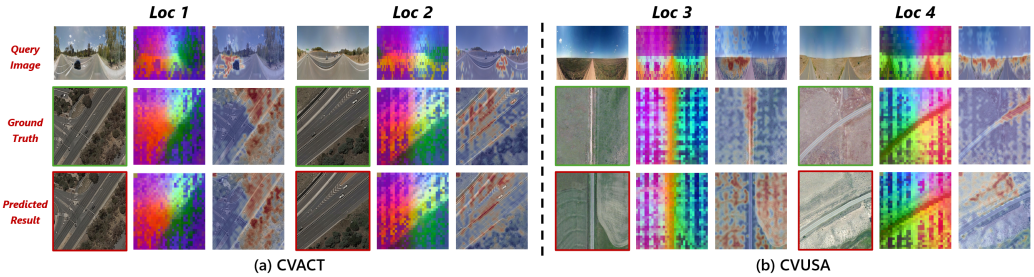


Figure 10: Visualization Analysis of Failed Case of Ground-view CVGL on CVACT and CVUSA.

candidates difficult to distinguish. As shown in Figure 9 (a), several predictions are attracted to locations with similar local structures despite differing from the ground truth in their overall spatial organization.

In contrast, queries captured at 300 m contain broader scene context and more complete geographic layouts, leading to stronger correspondence with satellite imagery, as illustrated in Figure 9 (b). The additional contextual information provides more reliable semantic and structural cues for distinguishing competing locations. These observations indicate that UAV altitude substantially affects localization difficulty: lower-altitude observations emphasize local details but suffer from limited spatial context, whereas higher-altitude views provide more global structural evidence that facilitates cross-view matching.

Failure Case Analysis on Ground-view CVGL Figure 10 shows representative failure cases on CVACT and CVUSA. The ground-truth and predicted locations often share highly similar scene semantics and spatial layouts, such as road geometry, vegetation, and surrounding structures. In several cases, the prediction is geographically close to the correct location, differing mainly by a small local displacement. This reveals an intrinsic ambiguity of ground-view CVGL, where neighboring locations may exhibit nearly identical image-level appearance and structure.

These cases suggest that image-level retrieval alone may be insufficient for precise localization in highly repetitive street scenes. When candidate locations differ only in subtle lane boundaries, object placement, or local topology, the decisive evidence may exist at a finer granularity. Future work may therefore explore object-level correspondence, pixel-level matching, or explicit local geometric reasoning to better distinguish geographically adjacent but visually similar locations.

Failure Case Analysis on Multi-weather Robustness Evaluation Figure 11 presents representative failure cases under Fog-Rain, Fog-Snow, and Rain-Snow conditions. These compound weather shifts introduce severe visibility degradation and partial occlusion, substantially suppressing both scene-level semantics and fine-grained structural details. As highlighted in the examples, distinctive buildings, roads, and local layouts can become blurred or completely obscured, causing the predicted location to share only partial geographic evidence with the ground truth.

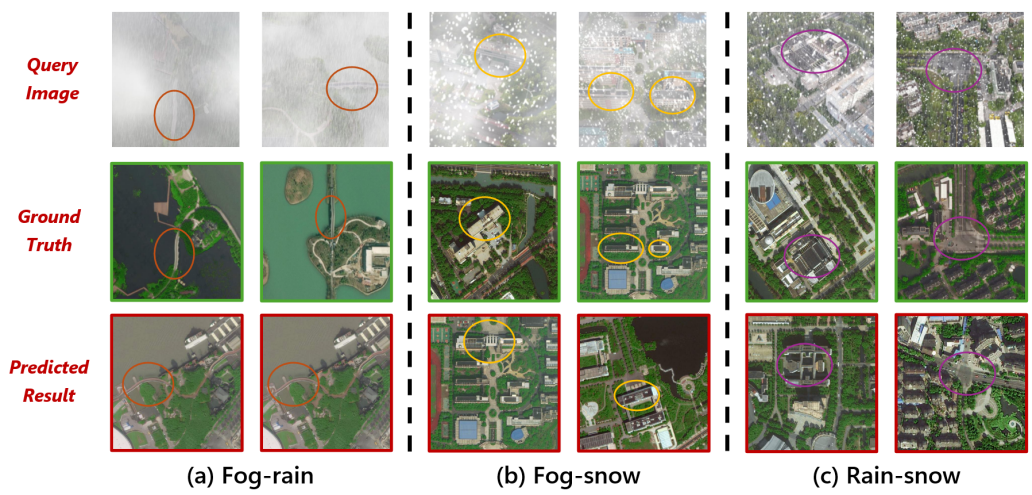


Figure 11: Failed Case Visualization Analysis of Multi-weather Robustness Evaluation on University-1652→SUES-200 at 150m. (Circles denote the ambiguous semantics)

Although GeoCore combines transferable semantic cues with reliability-aware spatial structures, such complementarity remains limited when both types of evidence are simultaneously corrupted or unavailable. In particular, extreme weather may remove the local structures required for fine-grained discrimination while also weakening global semantic recognition. These cases indicate that representation-level robustness alone cannot fully recover missing visual evidence. Future work may incorporate visibility-aware uncertainty modeling, multi-temporal observations, or additional sensing modalities to improve localization when severe environmental conditions fundamentally reduce the observable geographic information.

A.4.2 Visualizing the Effect of RSSR

Figure 12 presents qualitative comparisons on University-1652→SUES-200 and CVACT→CVUSA using DINOv2-based variants. We compare anchor-weighted pooling, an ordered readout without rotation averaging, RSSR with shuffled regional layouts, and the complete RSSR. For the readout variants, we visualize the structural residual separately from the mean regional content to isolate its contribution to cross-view matching. Pooling maps are obtained by projecting regional matching contributions back to image patches through their assignment weights, whereas residual maps use signed gradient-based attribution.

In the illustrated examples, pooling produces relatively coarse responses, summarizing regional content without explicitly modeling inter-region relationships. The ordered and shuffled readouts yield different response patterns, highlighting the influence of regional arrangement on the structural residual. Full RSSR exhibits more spatially differentiated responses around paths, boundaries, and road–field layouts. In particular, the CVUSA examples show contrasts between linear scene structures and surrounding areas, while the SUES-200 examples reveal responses distributed over spatially extended layouts. These observations are consistent with the intended role of RSSR: retaining layout-dependent evidence while reducing sensitivity to discrete rotations of the regional grid. The visualizations provide qualitative interpretation of selected examples and complement the quantitative ablations.

A.5 Discussion

This work presents GeoCore for generalizable CVGL by distinguishing cross-view invariance from geographic discrimination. Rather than treating all invariant cues as equally informative, GeoCore combines transferable scene semantics with fine-grained regional structure. The semantic branch learns view-consistent scene concepts from VLM's textual supervision, while the structural branch organizes local features into anchored regions, models their spatial arrangements, and promotes regional discrimination across appearance variations. Our retrieval-margin analysis explains how

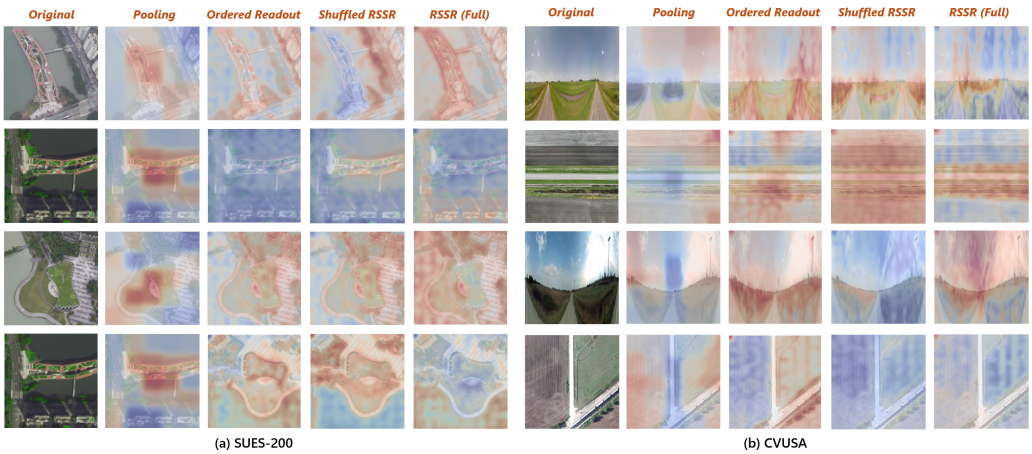


Figure 12: Qualitative visualization of anchor-weighted pooling and RSSR variants on (a) SUES-200 and (b) CVUSA. Each example shows a query image and its corresponding satellite image. Readout visualizations isolate the structural residual. Red and blue denote positive and negative relative contributions to cross-view matching.

structural evidence can resolve locations that remain ambiguous in the semantic space, while the reliability analysis relates CRR to a smooth approximation of the worst-environment regional margin. Experiments on five UAV and ground-view benchmarks demonstrate consistent improvements in cross-area transfer and adverse-weather robustness, with textual supervision and reliability prediction confined to training.

Several findings further clarify the behavior and scope of GeoCore. (1) The relative contribution of each core depends on the geographic characteristics of the benchmark. UAV images capture broad scene context, allowing the Semantic Core to provide effective coarse discrimination. In densely sampled ground-view settings, neighboring locations often share similar scene semantics, making fine-grained structural evidence more important. Their combination nevertheless performs best across both settings, supporting the complementary role of the two cores. (2) GeoCore achieves larger gains at lower UAV altitudes and under challenging weather shifts such as Fog-Snow. However, the failure cases show that limited spatial coverage or severe visibility degradation may simultaneously remove semantic and structural evidence. Representation learning cannot fully recover geographic information that is absent from the observations. (3) The RSSR ablations confirm that its benefit depends on meaningful regional ordering, while its symmetry applies specifically to quarter-turn permutations of the encoded anchor grid rather than arbitrary image rotations. In addition, CRR covers only the environmental variations represented during training. Future work may explore finer local correspondence and additional sensing modalities (e.g., infrared imagery), for localization under severe visual degradation.