GEOCORE: RECOVERING MISSING CROSS-VIEW RELATIONS UNDER DISJOINT PAIRWISE SUPERVISION FOR TRI-VIEW GEO-LOCALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Tri-view geo-localization integrates UAV, satellite, and ground views, with fully supervised settings typically relying on direct paired supervision for the cross-view relations of interest. Such supervision is difficult to obtain when geospatial data are independently collected across platforms or datasets, leaving some relations unobserved during training. This work considers an incomplete tri-view setting by recovering the UAV–ground (u–g) relation from UAV–satellite (u–s) and satellite–ground (s–g) pairs, without direct u–g supervision or shared location instances between the two observed relations. Existing pairwise methods can learn both observed relations effectively yet still struggle to recover the missing u–g relation. Because the two observed relations are supervised on disjoint location sets, each pairwise objective constrains only its own relation, and the shared satellite view alone does not establish cross-edge comparability. This motivates organizing the two relations under shared coordinates and a common metric. GeoCoRe is introduced as a two-stage framework for recovering missing cross-view relations in tri-view geo-localization. Stage 1 induces shared Geo-primitives as a shared primitive reference, while Stage 2 uses this reference to guide coordinate assignment and align shared-coordinate representations across the observed edges. GeoCoRe improves the average missing-edge R@1 from 1.53% for the strongest baseline to 6.11%, demonstrating the value of cross-relation compositionality for recovering unobserved relations. Code is available at https://anonymous.4open.science/r/GeoCoRe.

1 INTRODUCTION

Cross-view Geo-Localization (CVGL) aims to localize images by matching observations captured from heterogeneous platforms, with applications in UAV navigation and remote sensing imagery interpretation (Wilson et al., 2024; Zheng et al., 2020). Existing methods commonly learn shared representations to bridge cross-view discrepancies for CVGL, achieving strong performance on individual cross-view relations, including UAV–satellite (u–s) and satellite–ground (s–g) (Hu et al., 2018; Shi et al., 2019; Dai et al., 2024; Zhu et al., 2022; Wang et al., 2022). These advances show that pairwise shared representation learning is effective for individual cross-view relations. Recent tri-view methods further extend this idea by jointly learning representations from same-location correspondences, enabling comparison across multiple cross-view relations (Liang et al., 2026).

However, such complete tri-view correspondences, remain comparatively limited due to the difficulty of coordinating data collection across heterogeneous platforms (Song et al., 2026). Accordingly, CVGL training data are more commonly available as separate pairwise correspondences (Li et al., 2024a), including u–s and s–g pairs (Wu et al., 2024). This motivates studying tri-view geo-localization where supervision is available only through separate u–s and s–g pairs drawn from disjoint location sets, as illustrated in Figure 1(a). Unlike complete tri-view supervision, no location jointly connects the three views in this setting, and the target UAV–ground (u–g) relation remains entirely unobserved during training. Satellite appears in both observed relations, but the two training sets share no location instances. The shared satellite view therefore provides only a view-level connection and does not directly constrain how the two independently supervised relations are or-

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.

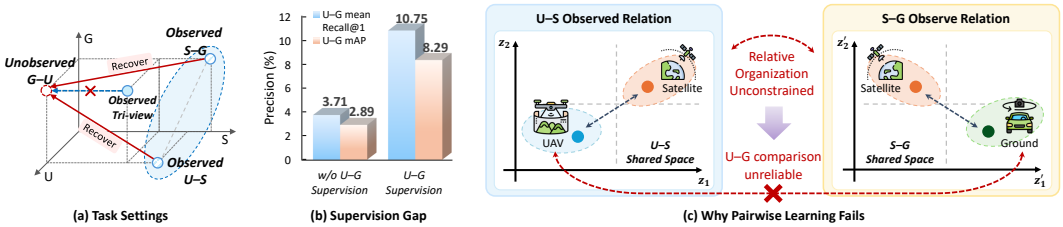


Figure 1: **Motivation.** (a) We study recovery of an unobserved u–g relation from disjoint u–s and s–g supervision. (b) Direct u–g supervision exposes a substantial missing-edge gap. (c) Pairwise learning aligns observed relations independently, leaving relative organization unconstrained.

ganized relative to each other. This setting raises the question of whether the two observed relations can be composed to recover the unobserved u–g relation.

To test whether such composition emerges under disjoint pairwise supervision, we construct disjoint u–s and s–g location splits in University-1652 dataset. As shown in Figure 1(b), Sample4Geo achieves only 3.71% average R@1 on the unobserved u–g relation when trained on the two observed relations, compared with 10.75% when direct u–g supervision is additionally provided. This gap indicates that learning the two observed relations alone does not naturally produce representations that are comparable on the unobserved relation. Pairwise supervision aligns each observed relation within its own learned space (Zhao et al., 2024; Chen et al., 2025; Wang et al., 2025). Without shared location instances, the relative organization of the u–s and s–g spaces remains unconstrained, so UAV and ground representations are not guaranteed to be directly comparable (Ji et al., 2025b), as illustrated in Figure 1(c). Therefore, the key challenge is to establish cross-edge comparability across independently supervised relations.

Addressing this challenge requires bringing independently supervised relations into a common frame of reference. Sharing the satellite view alone does not constrain the relative organization between the edge-specific embedding spaces learned for the two observed relations. Accordingly, a common reference can guide coordinate assignment so that view-specific representations are organized under shared coordinates (Wang et al., 2023; Shi et al., 2023). Supervision from both observed relations then aligns the shared-coordinate representations under a common metric (Wang et al., 2024a), making UAV and ground representations directly comparable without u–g supervision.

Building on this insight, we propose a two-stage framework, **GeoCoRe**, for tri-view geo-localization under disjoint pairwise supervision. GeoCoRe instantiates the common reference with shared Geo-primitives and uses it to organize independently supervised relations under shared coordinates and a common metric. Stage 1, **Observed-Edge Geo-Primitive Induction**, learns representations from the observed u–s and s–g relations and induces shared Geo-primitives with consistent identities across views, providing the reference for subsequent coordinate mapping. Stage 2, **Shared Coordinate Mapping for Missing-Edge Recovery**, uses this reference to guide coordinate assignment and organize view-specific representations under shared coordinates. Supervision from both observed relations further aligns these representations under a common metric, enabling direct comparison between UAV and ground views without u–g supervision. Experiments show substantial gains on the unobserved u–g relation, while ablations verify the contributions of the shared primitive reference, primitive-guided coordinate assignment, and common-metric alignment to cross-edge comparability. Our contributions are summarized as follows:

- We formulate tri-view geo-localization under disjoint pairwise supervision as a compositional generalization problem, where an entirely unobserved relation is recovered from independently supervised observed relations.
- We propose GeoCoRe, which establishes a shared primitive reference to guide coordinate assignment, organizes view-specific representations under shared coordinates, and aligns them under a common metric to establish cross-edge comparability.
- We demonstrate that, under a protocol using disjoint u–s and s–g supervision from DenseUAV and VIGOR, GeoCoRe improves missing-edge Recall@1 on University-1652 from 1.53% to 6.11% without direct u–g supervision.

2 Related Work

Cross-View Geo-Localization. Existing methods improve cross-view representation learning through local pattern modeling (Wang et al., 2022; Lv et al., 2024), Transformer-based global encoding (Zhu et al., 2022), geometric layout correspondence (Zhang et al., 2023; 2024) and recent multi-level or query-driven representation learning (Ge et al., 2024; Hu et al., 2025). Sample4Geo further improves retrieval discriminability through contrastive learning and hard-negative sampling (Deuser et al., 2023). Recent studies relax ideal pairing conditions by considering many-to-many matching under partial overlap (Ji et al., 2025a), spatial decentrality (Xia et al., 2025), and noisy correspondences caused by GPS offsets (Li et al., 2026). These advances improve matching within individual cross-view relations under increasingly relaxed pairing conditions, while the target relation itself remains directly observed during training. Beyond pairwise localization, recent studies have explored the tri-view geo-localization. University-1652 (Zheng et al., 2020) provides a unified benchmark containing UAV, satellite, and ground imagery. More recently, UniGeoRS (Liang et al., 2026) supports joint localization and evaluation across these three views, while GeoBridge (Song et al., 2026) supports multi-directional retrieval among UAV, street-view, and satellite imagery. These studies extend CVGL from isolated view-pair matching to unified modeling of multiple cross-view relations, with the target relations observed during training. In contrast, we study tri-view geo-localization under disjoint pairwise supervision, where u–s and s–g pairs are from disjoint sets and the u–g relation remains entirely unobserved during training.

Learning Cross-View Relations from Incomplete Pairings. Prior work shows that indirect alignment can emerge even when all data sources are not directly paired (Zhu et al., 2024; Yacobi et al., 2025; Zhou et al., 2022; Huang et al., 2024; Schnaus et al., 2025). ImageBind (Girdhar et al., 2023) for example, aligns multiple modalities through a common image modality, enabling transfer between modality pairs without direct supervision (Wang et al., 2024b; Chung et al., 2026). Recent studies further consider weakly paired data (Lee et al., 2022; Timilsina et al., 2024) and instance-unpaired data (Xi et al., 2024; Sturma et al., 2023; Gupta et al., 2026), showing that cross-source representations can still be aligned without complete pairing. However, such alignment does not by itself establish whether independently supervised relations can be composed to recover the unobserved relation (Li et al., 2024b). We study this compositional setting by establishing cross-edge comparability between the observed relations.

3 Method

3.1 Problem Definition and Overview

Problem Definition. We consider tri-view geo-localization across UAV, satellite, and ground views, denoted by $\mathcal{V} = \{u, s, g\}$. Training supervision is available only for two observed relations, u–s and s–g, with $\mathcal{D}_{us} = \{(x_i^u, x_i^s, y_i)\}_{i=1}^{N_{us}}$ and $\mathcal{D}_{sg} = \{(x_j^s, x_j^g, y_j)\}_{j=1}^{N_{sg}}$, where x^v denotes an image from view v and y is its location identity. The two observed relations are supervised on disjoint location sets, $\mathcal{Y}_{us} \cap \mathcal{Y}_{sg} = \varnothing$, and the u–g relation remains entirely unobserved during training, *i.e.*, $\mathcal{D}_{ug} = \varnothing$. Accordingly, the observed edge set is $\mathcal{E}_{\text{obs}} = \{(u, s), (s, g)\}$, while the missing edge set is $\mathcal{E}_{\text{miss}} = \{(u, g)\}$. Our goal is to recover the missing u–g relation using only $\mathcal{D}_{us}$ and $\mathcal{D}_{sg}$.

Overview. Because the two observed relations are supervised on disjoint location sets, each pairwise objective constrains only its own relation and does not establish cross-edge comparability. GeoCoRe establishes such comparability by building a shared primitive reference in Stage 1 and then organizing representations under shared coordinates and a common metric in Stage 2, as shown in Figure 2. Specifically, Stage 1, Observed-Edge Geo-Primitive Induction, extracts view-adapted local features T^v and induces shared Geo-primitives with consistent primitive indices across views, establishing a shared primitive reference. It retains the corresponding response maps A^v and produces the Stage 1 descriptor b^v. Stage 2, Shared Coordinate Mapping for Missing-Edge Recovery, uses A^v to guide content-adaptive assignment of T^v to shared coordinates and encodes the resulting assignments into a shared-coordinate descriptor s^v. Supervision from both observed relations further aligns these representations under a common metric, establishing cross-edge comparability. Finally, b^v and s^v are combined into the final descriptor d^v, enabling direct retrieval across all view pairs, including the unobserved u–g relation.

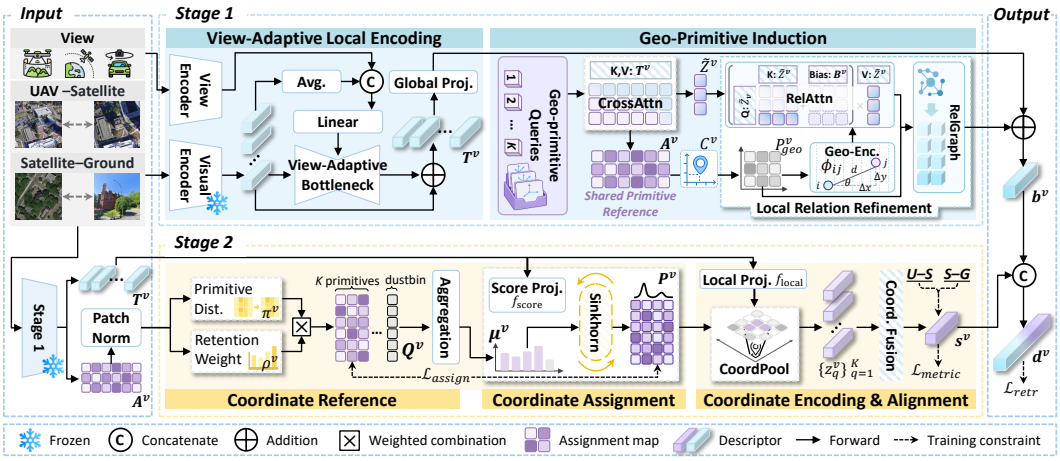


Figure 2: **Overview of GeoCoRe.** Stage 1 extracts view-adapted local features and induces shared Geo-primitives across views from the observed u–s and s–g relations. These primitives provide a shared reference for subsequent coordinate mapping through their responses A^v, while Stage 1 also produces the retrieval descriptor b^v. Stage 2 uses A^v to guide coordinate assignment and organizes local features under shared coordinates, producing the shared-coordinate descriptor s^v. Supervision from both observed relations further aligns these representations under a common metric, establishing cross-edge comparability. The final descriptor d^v combines b^v and s^v.

3.2 *Stage 1*: Observed-Edge Geo-Primitive Induction

View-Adaptive Local Encoding. To provide view-adapted local features for Geo-primitive induction across views, we apply a lightweight view-conditioned modulation to the frozen patch features from a frozen visual encoder. Given an image x^v from view $v \in \{u, s, g\}$, the encoder produces patch features $X^v = E_{\text{vis}}(x^v) \in \mathbb{R}^{N_v \times D}$. A lightweight bottleneck is conditioned on pooled image context and a learnable view embedding to produce an adaptation residual Δ^v_{view}. The resulting view-adapted local features are $T^v = X^v + \alpha\Delta^v_{\text{view}}$, where α controls the adaptation strength. We further summarize T^v with GeM pooling and a shared projection head to obtain the image-level descriptor $b^v_{\text{va}} = \text{Norm}(\text{MLP}_{\text{glob}}(\text{GeM}(T^v)))$.

Geo-Primitive Induction. Although T^v preserves fine-grained local information, it does not provide a common primitive reference across views. Therefore, we introduce a learnable query bank $Q_{\text{prim}} = \{q_k\}_{k=1}^K \in \mathbb{R}^{K \times D}$ shared across views, where each query defines a primitive index and induces a Geo-primitive from the local features. For each view $v \in \{u, s, g\}$, the shared queries attend to T^v to induce the corresponding Geo-primitives and their patch-level responses,

$$(\widetilde{Z}^v, A^v) = \text{CrossAttn}(Q_{\text{prim}}, T^v, T^v), \tag{1}$$

where $\widetilde{Z}^v$ contains the induced primitive representations, and A^v records their primitive-to-patch responses. Since each query q_k is shared across views, primitive index k maintains a consistent identity across views, establishing a shared primitive reference. The corresponding responses A^v are retained to guide coordinate assignment in Stage 2.

To enrich the induced primitives with local spatial relations, we obtain their response-weighted spatial supports as $P^v_{\text{geo}} = A^v C^v$, where $C^v \in \mathbb{R}^{N_v \times 2}$ denotes normalized patch coordinates. Then we encode pairwise displacement, distance, and orientation derived from P^v_{geo} into a geometric bias $B^v = f_{\text{geo}}(\phi^v)$ and refine the primitives as $Z^v = \text{RelAttn}(\widetilde{Z}^v; B^v)$. This incorporates local geometric context while preserving the query-defined primitive indices. We further refine the primitives with local neighborhood relations using $\widehat{Z}^v = \text{RelGraph}(Z^v, P^v_{\text{geo}})$, and summarize them into a graph-level descriptor b^v_{graph}. Combining it with the image-level descriptor b^v_{va} gives $b^v = \text{Norm}(b^v_{\text{va}} + \beta b^v_{\text{graph}})$, where β controls the contribution of the graph-level descriptor.

Stage 1 uses a single model shared across views and is alternately optimized on the two observed relations. For each observed relation $(a, b) \in \mathcal{E}_{\text{obs}}$, the Stage 1 objective is $\mathcal{L}^{a,b}_{\text{S1}} = \mathcal{L}^{a,b}_{\text{retr}} + \mathcal{L}^{a,b}_{\text{cons}} +$

$\mathcal{L}^{a,b}_{\text{div}}$, where the three terms respectively supervise retrieval, cross-view primitive consistency, and primitive diversity. Stage 1 outputs b^v for retrieval and retains the primitive responses A^v to guide coordinate assignment in Stage 2 under the shared primitive reference.

3.3 *Stage 2*: Shared Coordinate Mapping for Missing-Edge Recovery

With Stage 1 frozen, Stage 2 reuses its outputs (b^v, T^v, A^v). Using the primitive responses A^v to guide coordinate assignment, Stage 2 maps the local features T^v into shared-coordinate representations through primitive-guided coordinate assignment and shared-coordinate encoding. Supervision from both observed relations further aligns the resulting shared-coordinate representations under a common metric, enabling direct comparison on the unobserved u–g relation.

Primitive-Guided Coordinate Assignment. We use the primitive responses A^v from the frozen Stage 1 to guide the assignment of T^v to shared coordinates. Using these primitive responses, we derive a patch-level reference assignment $Q^v \in \mathbb{R}^{N_v \times (K+1)}$ over the K shared coordinate indices and an additional dustbin. After normalizing each primitive response over patches to obtain $\bar{A}^v$, we compute the relative primitive distribution π^v_{nq} for patch n as

$$\pi^v_{nq} = \frac{\bar{A}^v_{qn}}{\sum_{k=1}^{K} \bar{A}^v_{kn}}. \tag{2}$$

We further compute a retention weight $\rho^v_n \in [0,1]$ from the overall primitive-response strength of each patch, assigning the remaining mass to the dustbin. The patch-level reference Q^v and its image-level occupancy prior μ^v are then defined as

$$Q^v_{nq} = \rho^v_n \pi^v_{nq}, q \in \{1,\dots,K\}, \quad Q^v_{n,\varnothing} = 1 - \rho^v_n, \quad \mu^v_q = \frac{1}{N_v}\sum_{n=1}^{N_v} Q^v_{nq}, q \in \{1,\dots,K,\varnothing\}. \tag{3}$$

Thus, Q^v provides a primitive-guided reference assignment for individual patches, while μ^v summarizes the image-level occupancy prior over the shared coordinates.

To obtain content-adaptive coordinate assignments while preserving this primitive guidance, a shared scoring head predicts assignment scores from T^v. With an additional dustbin score, the content-score matrix L^v and final coordinate assignment P^v are defined as

$$L^v = [f_{\text{score}}(T^v); \ell_\varnothing] \in \mathbb{R}^{N \times (K+1)}, \quad P^v = \text{Sinkhorn}\left(L^v; \mu^v\right). \tag{4}$$

Sinkhorn normalization uses a uniform patch marginal and μ^v as the coordinate occupancy marginal, yielding the content-adaptive assignment P^v. To preserve consistency with the patch-level reference assignment, we align P^v with Q^v via

$$\mathcal{L}^v_{\text{assign}} = \frac{1}{N_v \log 2}\sum_{n=1}^{N_v} \text{JS}(P^v_n \| Q^v_n), \tag{5}$$

where $\text{JS}(\cdot\|\cdot)$ denotes the Jensen–Shannon divergence between the predicted coordinate assignment and the primitive-guided reference assignment. The resulting P^v determines how local features are assigned and aggregated under the shared coordinates.

Shared-Coordinate Encoding and Dual-Edge Metric Alignment. Given the coordinate assignment P^v, we aggregate the local features T^v at each shared coordinate to obtain coordinate-wise representations z^v_q, which are further fused to obtain the shared-coordinate descriptor s_v:

$$\begin{aligned} z^v_q &= \text{CoordPool}\left(\{f_{\text{local}}(T^v_n)\}_{n=1}^{N_v}, \{P^v_{nq}\}_{n=1}^{N_v}\right), \qquad q \in \{1,\dots,K\}, \\ s^v &= \text{Norm}\left(f_{\text{fuse}}\left([z^v_1;\dots;z^v_K]\right)\right), \end{aligned} \tag{6}$$

where f_{local} and f_{fuse} denote the shared local projection and coordinate fusion functions, respectively, and CoordPool performs assignment-weighted aggregation at each shared coordinate.

Shared coordinates provide a common organization for representations across views. Cross-edge comparability is further established by constraining the resulting descriptors under a common metric. We jointly supervise the shared-coordinate descriptors on both observed relations:

$$\mathcal{L}_{\text{metric}} = \mathbb{E}_{(u,s)\sim\mathcal{D}_{us}}\left[\mathcal{L}_{\text{MP}}(s^u, s^s)\right] + \mathbb{E}_{(s,g)\sim\mathcal{D}_{sg}}\left[\mathcal{L}_{\text{MP}}(s^g, s^s)\right], \tag{7}$$

Table 1: **Comparison on observed-edge and missing-edge geo-localization.** Baselines are grouped according to their original evaluation directions. U2S, S2U, G2S, S2G, U2G, and G2U denote $u \to s$, $s \to u$, $g \to s$, $s \to g$, $u \to g$, and $g \to u$ tasks, respectively.

Method	Backbone	Observed Edges																Missing Edge								
		U2S				S2U				G2S				S2G				R_{obs}	U2G				G2U				R_{miss}
		R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP		R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	
Original Eval. Dir.: G2S																											
GeoDTR (Zhang et al., 2023)	ResNet-34	1.47	5.45	8.78	2.95	1.43	3.71	6.13	0.98	0.70	1.90	3.72	1.42	1.00	2.28	4.71	0.87	1.15	0.18	0.91	1.96	0.40	0.12	0.74	1.09	0.22	0.15
GeoDTR+ (Zhang et al., 2024)	ConvNeXt-T	0.20	0.72	1.40	0.59	0.43	1.28	2.00	0.28	0.35	1.43	2.33	0.93	0.71	2.28	3.42	0.64	0.42	0.06	0.39	0.90	0.25	0.04	0.16	0.19	0.18	0.05
Original Eval. Dir.: U2S																											
DMNIL (Chen et al., 2026b)	ConvNeXt-T	30.27	49.39	58.25	35.02	63.62	77.32	81.03	35.46	0.16	1.51	2.75	0.81	0.43	1.85	3.28	0.59	23.62	0.60	2.45	4.20	0.73	0.85	2.33	3.49	0.46	0.73
Game4Loc (Ji et al., 2025a)	RoPE ViT-B/16	0.80	4.07	7.20	2.07	2.28	6.28	10.13	1.04	1.05	2.71	4.42	1.86	1.43	4.71	6.56	1.38	1.39	0.26	0.95	1.76	0.39	0.12	0.58	1.16	0.22	0.19
Original Eval. Dir.: U2S + S2U																											
University-1652 (Zheng et al., 2020)	ResNet-50	4.86	11.80	17.23	7.11	7.85	14.55	18.12	3.97	0.71	2.14	3.71	0.94	0.50	1.32	2.75	1.12	3.48	0.63	2.04	3.50	0.67	0.35	0.97	1.67	0.37	0.49
LKGL (Fan et al., 2026)	DINOv2 ViT-B/14	11.88	27.14	35.98	15.97	32.95	49.93	56.06	12.18	0.27	1.94	4.03	1.04	1.00	3.57	6.56	1.06	11.53	0.54	2.68	4.68	0.79	0.66	1.43	2.48	0.48	0.60
UniABG (Chen et al., 2026a)	ConvNeXt-B	17.85	32.13	39.41	21.62	56.30	71.33	78.32	17.31	0.35	1.09	2.02	0.80	0.29	1.14	2.85	0.59	18.70	0.57	2.60	4.20	0.74	0.27	1.51	2.21	0.30	0.42
Original Eval. Dir.: U2S + S2U + G2S																											
LPN (Wang et al., 2022)	ResNet-50	9.51	20.41	27.22	12.67	20.54	29.53	34.95	8.93	0.23	0.70	1.47	0.66	0.57	1.85	2.85	0.59	7.71	0.29	1.12	2.52	0.51	0.00	0.12	0.39	0.25	0.14
TransGeo (Zhu et al., 2022)	DeiT-S/16	3.26	11.06	17.54	5.80	3.57	8.99	12.70	2.05	0.93	3.99	6.51	2.12	1.28	6.42	8.70	1.42	2.26	0.41	1.73	3.40	0.66	0.27	0.62	1.09	0.30	0.34
Sample4Geo (Deuser et al., 2023)	ConvNeXt-B	14.90	30.66	39.31	19.04	43.94	63.77	71.90	14.70	1.09	3.72	6.28	2.16	1.57	5.71	8.70	1.50	15.38	1.55	5.50	9.22	1.52	1.51	4.11	6.32	0.97	1.53
GeoCoRe with Different Backbones																											
GeoCoRe	ConvNeXt-T	22.08	41.63	51.87	27.03	48.07	61.20	67.90	25.22	1.09	4.30	7.02	2.33	1.71	7.28	10.84	2.09	18.24	1.84	5.79	9.31	1.90	1.43	3.18	4.85	1.14	1.64
	ConvNeXt-B	23.22	44.58	54.89	28.50	50.36	65.91	72.04	31.00	1.24	5.16	8.69	2.74	2.85	7.99	12.27	2.69	19.42	1.97	6.41	10.40	2.08	2.02	4.38	6.01	1.52	2.00
	DeiT-S Distilled	17.96	36.82	46.45	22.74	47.65	63.34	70.61	23.40	0.43	2.40	4.65	1.36	0.86	3.71	6.85	1.16	16.73	1.36	4.57	7.58	1.37	0.78	2.29	3.68	0.80	1.07
	DINOv2 ViT-B/14	41.38	65.82	75.59	47.13	66.19	79.60	83.17	44.32	4.42	12.64	18.88	7.03	7.85	17.26	25.25	5.88	29.96	5.93	15.04	21.84	5.14	5.23	10.04	13.26	3.34	5.58
	DINOv3 ViT-B/16	**43.88**	**69.05**	**77.94**	**49.66**	**69.19**	**83.45**	**88.16**	**45.72**	**4.46**	**11.36**	**18.11**	**6.71**	**6.13**	**16.12**	**22.11**	**5.27**	**30.92**	**5.98**	**15.08**	**21.64**	**4.93**	**6.24**	**11.44**	**14.97**	**4.01**	**6.11**

where $\mathcal{L}_{\text{MP}}$ is the matching loss. Supervision from both observed relations jointly aligns the shared-coordinate representations under a common metric, establishing cross-edge comparability for the unobserved u–g relation. Finally, we form the direct-sum descriptor by combining the Stage 1 descriptor b^v with the shared-coordinate descriptor s^v, $d^v = \frac{1}{\sqrt{2}}[b^v; s^v]$, where $\|b^v\|_2 = \|s^v\|_2 = 1$.

Training and Inference. During Stage 2 training, Stage 1 remains frozen and Stage 2 is optimized using supervision only from the two observed relations. The objective is $\mathcal{L}_{\text{S2}} = \mathcal{L}_{\text{retr}} + \mathcal{L}_{\text{metric}} + \mathcal{L}_{\text{assign}}$, where $\mathcal{L}_{\text{retr}}$ applies the matching loss to d^v. At inference, each image is independently encoded into d^v, and retrieval on the unobserved u–g relation uses $\text{score}(u, g) = (d^u)^\top d^g$.

4 Experimental

4.1 Experimental Setup

Datasets. DenseUAV (Dai et al., 2024) and VIGOR (Zhu et al., 2021) provide the observed u–s and s–g relations for training, respectively, with disjoint location sets across the two relations. DenseUAV provides u–s pairs over 2,256 locations, comprising 6,768 UAV images and 13,536 satellite images, while VIGOR provides s–g correspondences with 52,609 ground panoramas and 42,199 satellite images. University-1652 (Zheng et al., 2020) contains UAV, satellite, and ground imagery and is used exclusively for final evaluation over six directed cross-view retrieval tasks. Specifically, the u–g relation remains entirely unobserved during training.

Evaluation Metrics. We evaluate six directed geo-localization tasks, $u \to s$, $s \to u$, $s \to g$, $g \to s$, $u \to g$, and $g \to u$, using standard CVGL metrics of Recall@k ($k \in \{1, 5, 10\}$) and mean Average Precision (mAP). To summarize performance at the relation level, we further define $R_{\text{obs}} = \frac{1}{4}(R@1^{u \to s} + R@1^{s \to u} + R@1^{s \to g} + R@1^{g \to s})$ and $R_{\text{miss}} = \frac{1}{2}(R@1^{u \to g} + R@1^{g \to u})$. R_{obs} averages Recall@1 over the observed directions, while R_{miss} summarizes recovery of the unobserved u–g relation.

Implementation. The visual encoder in stage 1 is a frozen DINOv3 ViT-B/16 (Siméoni et al., 2025). All images are resized to 384×384. Stage 1 applies standard geometric augmentation, including random rotation for satellite images and random perspective sampling for ground panoramas, while Stage 2 uses four fixed horizontal views for ground panoramas. Both stages are trained for 30 epochs with AdamW and a batch size of 96, using learning rates of 3×10^{-4} and 1×10^{-4} for Stages 1 and 2, respectively. To balance supervision from the two observed relations, we sample DenseUAV and VIGOR with fixed probabilities during training. Stage 2 adopts location-balanced sampling to

provide multiple positives within each mini-batch. Model selection uses validation performance on the two observed relations only. Additional implementation details are provided in Appendix C.

4.2 Can GeoCoRe Recover the Unobserved Relation?

We compare GeoCoRe with representative CVGL baselines spanning pair-wise and recent tri-view retrieval methods. As shown in Table 1, the baselines generally achieve low retrieval performance on the unobserved u–g relation. A higher R_{obs} does not necessarily correspond to better missing-edge recovery. UniABG achieves a higher R_{obs} than Sample4Geo, 18.70% versus 15.38%, while its R_{miss} is lower, 0.42% versus 1.53%. This discrepancy illustrates that effective matching on the observed relations alone does not necessarily establish cross-edge comparability for the missing relation. Among the baselines, Sample4Geo achieves the highest R_{miss}. GeoCoRe increases R_{miss} from 1.53% to 6.11%, with R@1 of 5.98% and 6.24% on U2G and G2U, respectively. This improvement over Sample4Geo is also consistent across R@5, R@10, and mAP in both missing directions. The backbone comparison further shows consistent missing-edge recovery across ConvNeXt and ViT families. These variants also provide more direct reference points for baselines using different backbone families, with DINOv3 ViT-B/16 achieving the highest R_{obs} and R_{miss}.

4.3 What Is Required to Establish Shared Coordinates?

Table 2: **Ablation of shared primitive reference and primitive-guided coordinate assignment.** Base, Shared-Coord, and Direct-Sum denote the Stage 1 descriptor b^v, shared-coordinate descriptor s^v, and their direct-sum representation d^v, respectively.

Variant	Primitive Reference	Primitive Guidance	Rep.	R_{obs}	U2G R@1	G2U R@1	R_{miss}
Only Stage 1	Shared Q_{prim}	–	Base	22.14	3.99	3.88	3.94
Edge-Specific Primitives	$Q^{us}_{\text{prim}}, Q^{sg}_{\text{prim}}$	✓	Shared-Coord	26.49	3.90	5.00	4.45
			Direct-Sum	29.47	5.15	5.39	5.27
***w/o* Primitive Guidance**	Shared Q_{prim}	×	Shared-Coord	27.46	4.40	4.15	4.28
			Direct-Sum	30.27	5.32	5.51	5.42
GeoCoRe	Shared Q_{prim}	✓	Shared-Coord	28.78	4.79	5.16	4.98
			Direct-Sum	**30.92**	**5.98**	**6.24**	**6.11**

Is a Cross-Edge Shared Primitive Reference Necessary? To isolate the role of the shared primitive reference, we replace the shared query bank Q_{prim} with two independently learned banks, Q^{us}_{prim} and Q^{sg}_{prim}, while keeping the subsequent primitive-guided coordinate assignment unchanged. As shown in Table 2, Edge-Specific Primitives reduce the shared-coordinate R_{miss} from 4.98% to 4.45%, and the final Direct-Sum R_{miss} from 6.11% to 5.27%. These results demonstrate the importance of a shared primitive reference for maintaining consistent primitive identities across the observed relations and supporting missing-edge recovery.

Does Coordinate Assignment Require Shared Primitive Guidance? Having established the importance of a shared primitive reference, we next examine whether it should explicitly guide coordinate assignment. We construct a *w/o* Primitive Guidance variant by removing Q^v and $\mathcal{L}_{\text{assign}}$, replacing the image-specific μ^v with a fixed uniform marginal over the K coordinates and the dustbin, while keeping Sinkhorn normalization and the remaining Stage 2 architecture unchanged. As shown in Table 2, removing Primitive Guidance reduces the shared-coordinate R_{miss} from 4.98% to 4.28% and the Direct-Sum R_{miss} from 6.11% to 5.42%, while Direct-Sum R_{obs} remains at 30.27% compared with 30.92% for GeoCoRe. These results indicate that parameter sharing in Stage 2 alone is insufficient. Explicit guidance from the shared primitive reference is important for organizing the two observed relations under shared coordinates, thereby improving missing-edge recovery.

4.4 Is Dual-Edge Metric Alignment Necessary for Cross-Edge Comparability?

Building on above findings, we next examine whether dual-edge metric alignment is necessary to establish cross-edge comparability.

Removing Dual-Edge Metric Alignment. To isolate dual-edge metric alignment, we keep f_{score} shared to preserve the same primitive-guided coordinate assignment, while making f_{local} and f_{fuse}

Table 3: **Ablation of dual-edge metric alignment.** S and P denote shared and edge-private parameters, respectively. Base, Shared-Coord, and Direct-Sum correspond to b^v, s^v, and d^v.

Variant	Cross-edge Sharing			Rep.	Observed	Missing Edge		
	f_{score}	f_{local}	f_{fuse}		R_{obs}	U2G	G2U	R_{miss}
Only Stage 1	–	–	–	Base	22.14	3.99	3.88	3.94
***w/o* Dual-Edge Metric Alignment**	S	P	P	Shared-Coord	27.01	1.00	0.58	0.79
				Direct-Sum	32.83	4.12	3.76	3.94
Edge-Specific Coordinate Assignment	P	S	S	Shared-Coord	32.64	2.52	1.94	2.23
				Direct-Sum	34.73	4.55	5.31	4.93
Fully Edge-Specific Stage 2	P	P	P	Shared-Coord	31.09	0.85	0.62	0.74
				Direct-Sum	35.79	4.01	3.64	3.83
GeoCoRe	S	S	S	Shared-Coord	28.78	4.79	5.16	4.98
				Direct-Sum	30.92	**5.98**	**6.24**	**6.11**

edge-specific. The two observed relations therefore retain the same shared-coordinate organization, but their shared-coordinate descriptors are optimized through edge-specific mappings. As shown in Table 3, this change reduces the shared-coordinate R_{miss} from 4.98% to 0.79%, despite increasing the final Direct-Sum R_{obs} from 30.92% to 32.83%. After direct-sum fusion with the Stage 1 descriptor, R_{miss} increases to 3.94%, which remains below the 6.11% achieved by GeoCoRe. The large reduction in R_{miss} shows that shared-coordinate organization alone is not sufficient for cross-edge comparability. The resulting representations must also be aligned under a common metric.

Alternative Cross-Edge Sharing Patterns. The complementary variants exhibit a similar pattern under alternative cross-edge sharing settings. Edge-Specific Coordinate Assignment, which makes only f_{score} edge-specific, increases Direct-Sum R_{obs} to 34.73% but reduces R_{miss} to 4.93%. Fully Edge-Specific Stage 2 makes all three functions edge-specific, further increases R_{obs} to 35.79%, while the shared-coordinate and Direct-Sum R_{miss} fall to 0.74% and 3.83%, respectively. These results show that higher observed-edge retrieval performance does not imply cross-edge comparability. Aligning shared-coordinate representations from both observed relations under a common metric is therefore important for missing-edge recovery.

4.5 Sensitivity to the Number of Geo-primitives K

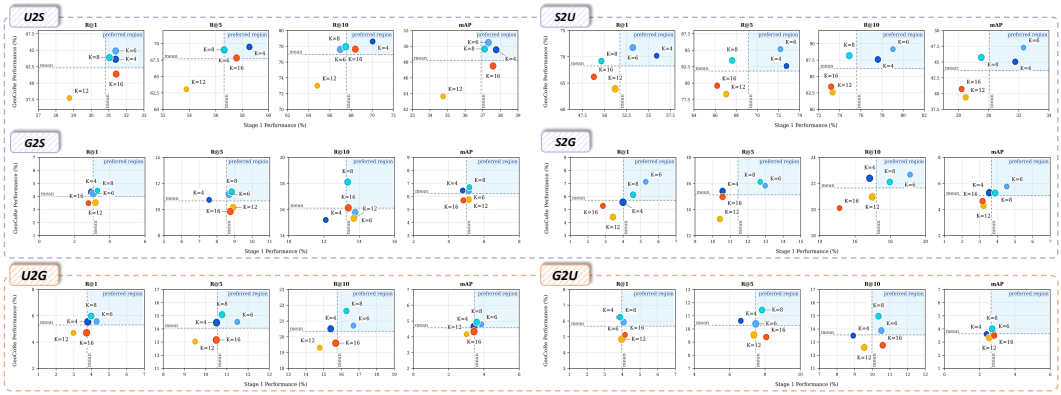


Figure 3: **Sensitivity to the number of Geo-primitives K.** Results are reported for all six retrieval directions, with R@1, R@5, R@10, and mAP for each direction. The x- and y-axes denote the Stage 1 and final GeoCoRe performance, respectively. Dashed lines indicate the mean performance across K, and the shaded upper-right region denotes configurations above the mean in both stages.

We evaluate the sensitivity to the number of Geo-primitives by varying $K \in \{4, 6, 8, 12, 16\}$ and reporting performance for both the Stage 1 and Direct-Sum descriptors. Since K determines the number of shared Geo-primitives in Stage 1 and shared coordinates in Stage 2, this comparison reflects its influence on both stages. As shown in Figure 3, $K = 8$ provides the most balanced

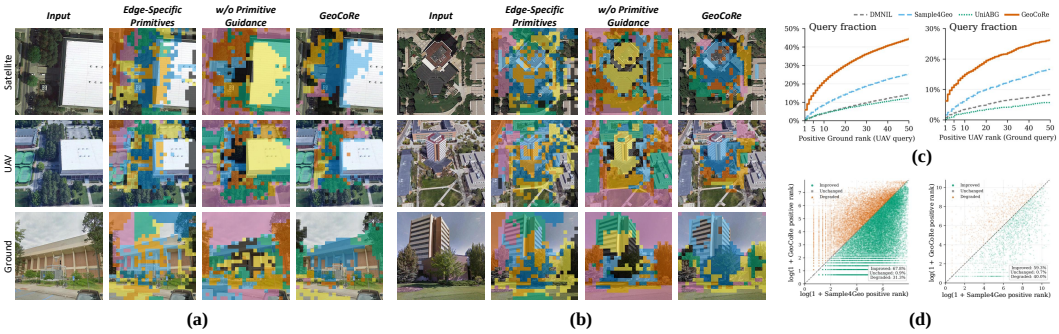


Figure 4: **Qualitative comparison and missing-edge rank analysis.** (a, b) Shared-coordinate assignments under Edge-Specific Primitives, *w/o* Primitive Guidance, and GeoCoRe. Colors denote coordinate indices, consistent across views within each variant but not across variants. (c) CMC curves based on the first positive rank for the u–g relation. (d) Query-wise first-positive-rank comparison between Sample4Geo and GeoCoRe, where points below the diagonal indicate higher ranks under GeoCoRe.

performance across observed and missing retrieval directions and across all four metrics. Smaller or larger values lead to less stable performance across directions, so we use $K = 8$ in GeoCoRe.

4.6 Visualization

Shared-Coordinate Assignment Analysis. We visualize the shared-coordinate assignments P^v on University-1652 test locations. Edge-Specific Primitives removes the cross-edge shared primitive reference through independently learned primitive banks, while *w/o* Primitive Guidance removes primitive guidance from coordinate assignment. As shown in Figure 4(a), *w/o* Primitive Guidance assigns the white-roof region in UAV and satellite to the same shared coordinate, but the corresponding assignment extends to a broader building region in ground. In GeoCoRe, the shared-coordinate assignments are more spatially localized, with the coordinate covering the main building region in satellite remaining concentrated on the upper building region and facade in ground. Figure 4(b) shows a similar observation. GeoCoRe keeps the assignments associated with the central high-rise concentrated on corresponding building regions across the three views, whereas the two ablations exhibit broader or shifted assignments.

Missing-Edge Rank Analysis To characterize ranking on the unobserved u–g relation, Figure 4(c) reports Cumulative Matching Characteristic (CMC) curves of the first positive rank for U2G and G2U. GeoCoRe yields consistently higher CMC curves than other methods in both directions, indicating that the improvement extends beyond the reported R@1, R@5, and R@10 thresholds to a broader range of retrieval ranks. Figure 4(d) further compares query-wise first positive ranks between GeoCoRe and Sample4Geo. The query-wise comparison further shows that GeoCoRe ranks the first positive match higher for a majority of queries in both U2G and G2U.

5 Conclusion

We study tri-view geo-localization under disjoint pairwise supervision and show that effective learning of the observed relations does not guarantee reliable recovery of the unobserved relation. This setting highlights that the shared satellite view alone does not establish cross-edge comparability between independently supervised relations. Therefore, recovering the missing relation requires organizing their representations under shared coordinates and a common metric. Building on this motivation, GeoCoRe induces shared Geo-primitives as a common reference, uses them to guide shared-coordinate mapping, and aligns the resulting representations under a common metric to establish cross-edge comparability. Without any u–g supervision, GeoCoRe improves the average missing-edge R@1 from 1.53% for the strongest baseline to 6.11%, with systematic ablations further supporting the key design choices. Overall, we cast missing-relation recovery under incomplete pairwise supervision as a problem of cross-relation compositionality, providing a broader perspective for recovering unobserved relations from independently supervised observed relations.

LLM USAGE

Large Language Models (LLMs) were used solely to assist with language editing and polishing, including grammar correction, sentence rephrasing, and readability improvement. They were not involved in research ideation, methodology, experimental design, data analysis, or scientific interpretation. All AI-assisted edits were reviewed by the authors, who take full responsibility for the content of the manuscript.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of GeoCoRe, including model configurations, training protocols, data splits, hyperparameters, and evaluation settings, in the main paper and appendix. Code, configuration files, and instructions for reproducing the experiments are provided in an anonymous repository. All datasets used in our experiments, including DenseUAV, VIGOR, and University-1652, are publicly available.

REFERENCES

Cuiqun Chen, Qi Chen, Bin Yang, and Xingyi Zhang. Uniabg: Unified adversarial view bridging and graph correspondence for unsupervised cross-view geo-localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2823–2831, 2026a. doi: 10.1609/aaai.v40i4.37272.

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

Zhongwei Chen, Zhao-Xu Yang, Hai-Jun Rong, and Guoqi Li. Without paired labeled data: End-to-end self-supervised learning for drone-view geo-localization. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2026b. doi: 10.1109/TNNLS.2026.3696684.

Dahyun Chung, Donghyun Shin, Yujin Sung, Seunggi Moon, Jinwoo Jeon, and Byung-Jun Lee. uclip: Parameter-efficient multilingual extension of vision-language models with unpaired data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3399–3406, 2026.

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

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

Xuyi Fan, Hongguang Li, Yangzhu Wang, Huaxin Qiu, Miaomiao Liu, Wei Li, and Li Shen. Enhancing cross-view geo-localization via global–local feature matching and keypoint guidance. *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2026. doi: 10.1109/TCSVT.2026.3704558.

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

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind one embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190, 2023.

Sharut Gupta, Shobhita Sundaram, Chenyu Wang, Stefanie Jegelka, and Phillip Isola. Better together: Leveraging unpaired multimodal data for stronger unimodal models. In *The Fourteenth International Conference on Learning Representations, ICLR*, 2026. URL https://openreview.net/forum?id=5OIgg5YkC3.

Shuyu Hu, Zelin Shi, Tong Jin, and Yunpeng Liu. Query-driven feature learning for cross-view geo-localization. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–15, 2025.

Sixing Hu, Mengdan Feng, Rang MH Nguyen, and Gim Hee Lee. Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7258–7267, 2018.

Yan Huang, Yuming Wang, Yunan Zeng, Junshi Huang, Zhenhua Chai, and Liang Wang. Unpaired image-text matching via multimodal aligned conceptual knowledge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5160–5176, 2024.

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, 2025a.

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 (ICCV)*, pp. 25165–25175, 2025b.

Sangmin Lee, Hyung-Il Kim, and Yong Man Ro. Weakly paired associative learning for sound and image representations via bimodal associative memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10524–10533, 2022.

Guopeng Li, Ming Qian, and Gui-Song Xia. Unleashing unlabeled data: A paradigm for cross-view geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16719–16729, 2024a.

Tianhong Li, Sangnie Bhardwaj, Yonglong Tian, Han Zhang, Jarred Barber, Dina Katabi, Guillaume Lajoie, Huiwen Chang, and Dilip Krishnan. Leveraging unpaired data for vision-language generative models via cycle consistency. In *The Twelfth International Conference on Learning Representations, ICLR*, volume 2024, pp. 34254–34274, 2024b.

Zheng Li, Xueyi Zhang, Yanming Guo, Yuxiang Xie, Zhaoyun Ding, Siqi Cai, Haizhou Li, and Mingrui Lao. Paul: Uncertainty-guided partition and augmentation for robust cross-view geo-localization under noisy correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5389–5398, June 2026.

Xiao Liang, Huaizhi Tang, Feiyang Zhang, Shiji Yuan, Chun Hu, Dezhi Zheng, and Kang Ma. Unigeors: A unified benchmark for tri-view geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5399–5408, 2026.

Hongxiang Lv, Hai Zhu, Runzhe Zhu, Fei Wu, Chunyuan Wang, Meiyu Cai, and Kaiyu Zhang. Direction-guided multiscale feature fusion network for geo-localization. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024.

Dominik Schnaus, Nikita Araslanov, and Daniel Cremers. It'sa (blind) match! towards vision-language correspondence without parallel data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24983–24992, 2025.

Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. In *Advances in Neural Information Processing Systems*, pp. 10090–10100, 2019.

Yujiao Shi, Fei Wu, Akhil Perincherry, Ankit Vora, and Hongdong Li. Boosting 3-dof ground-to-satellite camera localization accuracy via geometry-guided cross-view transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21459–21469, 2023.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

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 (CVPR)*, pp. 27793–27803, 2026.

Nils Sturma, Chandler Squires, Mathias Drton, and Caroline Uhler. Unpaired multi-domain causal representation learning. In *Advances in Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=zW1uVN6Mbv.

Subash Timilsina, Sagar Shrestha, and Xiao Fu. Identifiable shared component analysis of unpaired multimodal mixtures. In *Advances in Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=ivCX2cjwcT.

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

Tingyu Wang, Zhedong Zheng, Zunjie Zhu, Yaoqi Sun, Chenggang Yan, and Yi Yang. Learning cross-view geo-localization embeddings via dynamic weighted decorrelation regularization. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–12, 2024a.

Xiaolong Wang, Runsen Xu, Zhuofan Cui, Zeyu Wan, and Yu Zhang. Fine-grained cross-view geo-localization using a correlation-aware homography estimator. In *Advances in Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=cgiP4cMBP9.

Xueyi Wang, Lele Zhang, Zheng Fan, Yang Liu, Chen Chen, and Fang Deng. From coarse to fine: A matching and alignment framework for unsupervised cross-view geo-localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8024–8032, 2025.

Yabing Wang, Fan Wang, Jianfeng Dong, and Hao Luo. Cl2cm: Improving cross-lingual cross-modal retrieval via cross-lingual knowledge transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5651–5659, 2024b.

Daniel Wilson, Xiaohan Zhang, Waqas Sultani, and Safwan Wshah. Image and object geo-localization. *International Journal of Computer Vision*, 132(4):1350–1392, 2024.

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.

Johnny Xi, Jana Osea, Zuheng Xu, and Jason S. Hartford. Propensity score alignment of unpaired multimodal data. In *Advances in Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=hT4y7D2o2T.

Panwang Xia, Lei Yu, Yi Wan, Qiong Wu, Peiqi Chen, Liheng Zhong, Yongxiang Yao, Dong Wei, Xinyi Liu, Lixiang Ru, Yingying Zhang, Jiangwei Lao, Jingdong Chen, Ming Yang, and Yongjun Zhang. Cross-view geo-localization with panoramic street-view and vhr satellite imagery in decentrality settings. *ISPRS Journal of Photogrammetry and Remote Sensing*, 227:1–11, 2025. ISSN 0924-2716. doi: https://doi.org/10.1016/j.isprsjprs.2025.04.033.

Amitai Yacobi, Nir Ben-Ari, Ronen Talmon, and Uri Shaham. Learning shared representations from unpaired data. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=idxHcrwBYP.

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.

Hu Zhao, Keyan Ren, Tianyi Yue, Chun Zhang, and Shuai Yuan. Transfg: A cross-view geo-localization of satellite and uavs imagery pipeline using transformer-based feature aggregation and gradient guidance. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–12, 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.

Mingyang Zhou, Licheng Yu, Amanpreet Singh, Mengjiao Wang, Zhou Yu, and Ning Zhang. Unsupervised vision-and-language pretraining via retrieval-based multi-granular alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16464–16473, 2022.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, Hongfa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Caiwan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *The Twelfth International Conference on Learning Representations, ICLR*, 2024. URL https://openreview.net/forum?id=QmZKc7UZCy.

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 (CVPR)*, 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 (CVPR)*, pp. 1152–1161, 2022.

Appendix

A Cross-Relation Comparability with Disjoint Supervision

The two observed relations in our setting are supervised on disjoint location sets. The u–s objective directly constrains similarities within $\mathcal{D}_{us}$, while the s–g objective constrains similarities within $\mathcal{D}_{sg}$. The central question is whether satisfying these two sets of pairwise constraints is sufficient to determine the similarity structure of the unobserved u–g relation.

A direct way to examine this question is to ask whether the observed-edge similarities can remain unchanged under an alternative representation organization while the cross-relation similarity changes. Consider the two disjoint representation blocks induced by $\mathcal{D}_{us}$ and $\mathcal{D}_{sg}$. Let z_i^u and $z_i^{s,us}$ denote normalized UAV and Satellite representations from the first block, and let $z_j^{s,sg}$ and z_j^g denote normalized Satellite and Ground representations from the second block. Here, the two Satellite superscripts distinguish instances drawn from the two disjoint location sets.

For arbitrary orthogonal transformations Q_{us} and Q_{sg}, applying a common transformation within each observed block preserves all inner products within that block:

$$\begin{aligned}(Q_{us}z_i^u)^\top(Q_{us}z_k^{s,us}) &= (z_i^u)^\top z_k^{s,us},\\ (Q_{sg}z_j^{s,sg})^\top(Q_{sg}z_l^g) &= (z_j^{s,sg})^\top z_l^g.\end{aligned} \tag{8}$$

Thus, the two observed objectives can remain unchanged even when the two supervision blocks are expressed under different equivalent coordinate choices. We next examine whether the same coordinate changes also preserve the similarity across the two blocks.

Consider a UAV representation from the first block and a ground representation from the second block. Their u–g similarity becomes

$$(Q_{us}z_u)^\top(Q_{sg}z_g) = z_u^\top Q_{us}^\top Q_{sg}z_g. \tag{9}$$

Unlike the two observed-edge similarities, this cross-block similarity depends on the relative coordinate choice between the two supervision blocks. Thus, the u–s and s–g similarities can both remain unchanged under their respective common coordinate transformations, while the u–g similarity can

vary because its two endpoints are transformed by different coordinate choices. This shows that preserving the two observed relations is insufficient to uniquely determine the similarity of the missing u–g relation.

This ambiguity is directly relevant to our setting, where the two observed relations are supervised on disjoint location sets. Sharing the Satellite view type and encoder provides parameter-level coupling, but does not introduce shared location instances that explicitly constrain the relative coordinate organization of the two supervision blocks. In an actual shared-encoder model, parameter sharing, pretrained representations, and common regularities in the data further restrict the feasible representation space. The analysis above isolates what is determined by the pairwise similarity constraints, which do not by themselves specify the relative organization of the two disjoint supervision signals. This observation motivates the cross-relation coupling introduced by GeoCoRe. Stage 1 induces shared Geo-primitives from the observed relations to provide a common primitive reference. Stage 2 uses this reference to guide coordinate assignment and organize view-specific representations under shared coordinates. Joint supervision from the two observed relations further aligns these shared-coordinate representations under a common metric, establishing cross-edge comparability for recovery of the unobserved u–g relation.

B Complete Method Details

B.1 View-Conditioned Local Encoding

We use a frozen DINOv3 ViT-B/16 visual encoder with LVD-1689M pretrained weights. At an input resolution of 384×384, the encoder produces a 24×24 grid of patch features with $N_v = 576$ and feature dimension $D = 768$. The view-conditioned modulation uses a learnable 16-dimensional embedding for each view. Given the patch features X^v, the image context is obtained by average pooling, $\bar{x}^v = \frac{1}{N_v}\sum_{n=1}^{N_v} X_n^v$, and the view identity is represented by a learnable embedding $e^v = E(v)$. Their concatenation is passed through a two-layer gating MLP, $g^v = \mathrm{GateMLP}\left([\bar{x}^v; e^v]\right)$. The lightweight bottleneck is modulated by the image-level gate:

$$\Delta^v_{\mathrm{view}} = W^{\mathrm{view}}_{\mathrm{up}}\left[\phi\left(W^{\mathrm{view}}_{\mathrm{down}} X^v\right) \odot g^v\right], \tag{10}$$

where $\phi(\cdot)$ denotes GELU. The image-level gate is broadcast across all patch tokens of the same image. We set $\alpha = 0.05$, use no dropout, and zero-initialize the weights and biases of $W^{\mathrm{view}}_{\mathrm{up}}$.

For GeM pooling, we set $\epsilon = 10^{-6}$. The exponent is parameterized as $p_v = \mathrm{softplus}(\widetilde{p}_v) + \epsilon$, initialized to approximately 3 for each view, and kept fixed during Stage 1. The shared projection head consists of LayerNorm followed by a two-layer MLP with GELU activation.

B.2 Geo-Primitive Induction and Local Relation Modeling

We use $K = 8$ shared Geo-primitive queries. For the 24×24 patch grid, the center of patch (r, c) is linearly normalized to $[-1, 1]^2$ as $c^v_{rc} = \left(2\frac{c+1/2}{24} - 1,\ 2\frac{r+1/2}{24} - 1\right)$, forming the coordinate matrix C^v. The spatial support of the k-th Geo-primitive is computed as the response-weighted patch coordinate,

$$p^v_k = \sum_{n=1}^{N_v} a^v_{kn} c^v_n, \qquad P^v_{\mathrm{geo}} = [p^v_1; \ldots; p^v_K] = A^v C^v. \tag{11}$$

For primitives i and j, their relative geometry is represented by

$$\phi^v_{ij} = \left[\Delta x^v_{ij},\ \Delta y^v_{ij},\ d^v_{ij},\ \cos\theta^v_{ij},\ \sin\theta^v_{ij}\right], \tag{12}$$

where $\Delta x^v_{ij} = p^v_{j,x} - p^v_{i,x}$, $\Delta y^v_{ij} = p^v_{j,y} - p^v_{i,y}$, $d^v_{ij} = \|p^v_j - p^v_i\|_2$, and $\theta^v_{ij} = \mathrm{atan2}(\Delta y^v_{ij}, \Delta x^v_{ij})$. The corresponding geometric bias is $B^v_{ij} = f_{\mathrm{geo}}(\phi^v_{ij})$.

RelGraph further models local neighborhood relations among the induced Geo-primitives. For each primitive pair, we construct a rule-based geometric relation distribution over six relation types: *null, proximity, separation, local grouping, directional alignment*, and *symmetry*. The relation type is determined from the relative primitive positions. In parallel, a learned relation distribution is predicted

from the primitive pair and its relative geometry. The two distributions are embedded independently and combined as $R_{ij}^{v} = \sum_{t=0}^{5} r_{ij,t}^{\text{rule}} E_t^{\text{rule}} + \sum_{t=0}^{5} r_{ij,t}^{\text{learn}} E_t^{\text{res}}$. The sparse relation graph is constructed from the non-null mass of the two distributions. Taking relation type 0 as the null relation, the edge weight is $w_{ij}^{v} = \frac{1}{2}\left[1 - r_{ij,0}^{\text{rule}} + 1 - r_{ij,0}^{\text{learn}}\right], i \neq j$. For each node, only the strongest local connections are retained, and the resulting adjacency is symmetrized.

Before message aggregation, each relation R_{ij}^{v} is projected through a shared relation dictionary that combines predefined geometric relation types and learnable relation prototypes, yielding $\overline{R}_{ij}^{v}$. We further associate each primitive with a role embedding derived from its local graph topology and spatial layout. The node update jointly uses the primitive feature, aggregated relation message, and role embedding:

$$\widehat{z}_i^{v} = z_i^{v} + \text{MLP}_{\text{node}}\left(\left[z_i^{v}; \sum_j M_{ij}^{v} \overline{R}_{ij}^{v}; e_i^{\text{role}}\right]\right). \tag{13}$$

The updated Geo-primitives are mean-pooled and mapped by the graph readout head:

$$b_{\text{graph}}^{v} = h_{\text{graph}}\left(\frac{1}{K}\sum_{i=1}^{K} \widehat{z}_i^{v}\right). \tag{14}$$

We set the graph-descriptor fusion weight to $\beta = 0.05$. The relation modeling modules and graph readout are shared for all views and observed relations.

B.3 Stage 1 Training Objectives

Cross-View Primitive Correspondence. For each observed relation, we establish soft correspondences between the RelGraph-updated Geo-primitives of the paired views. Each primitive is projected by a shared MLP followed by node-wise ℓ_2 normalization, yielding features F^a and F^b. We compute the correspondence score by combining feature similarity with the discrepancy between primitive spatial supports:

$$C_{ij}^{ab} = \frac{(F_i^a)^\top F_j^b}{0.07} - 0.3\frac{\|p_i^a - p_j^b\|_2}{\overline{d}^{ab}}, \qquad \overline{d}^{ab} = \max\left(\text{sg}\left[\frac{1}{K^2}\sum_{i,j}\|p_i^a - p_j^b\|_2\right], 10^{-4}\right), \tag{15}$$

where sg denotes stop-gradient. After clipping the scores to $[-50, 50]$, softmax along the target primitive dimension gives the two directional correspondence matrices $\Pi_{ab} = \text{softmax}_j(C^{ab})$, $\Pi_{ba} = \text{softmax}_i((C^{ab})^\top)$. They are used to compare the instance-specific primitive representations and relational organization across the paired views in the Stage 1 consistency objectives.

Cross-View Primitive Consistency. Given the correspondence matrices above, we constrain the updated Geo-primitives at the node and local-relation levels. At the node level, each primitive representation is reconstructed from its soft correspondences in the paired view:

$$\mathcal{L}_{\text{node}}^{a,b} = \text{MSE}\left(\widehat{Z}^a, \Pi_{ab}\widehat{Z}^b\right) + \text{MSE}\left(\widehat{Z}^b, \Pi_{ba}\widehat{Z}^a\right). \tag{16}$$

We further constrain the local relational organization induced by RelGraph. For $x \in \{u, g\}$, the weighted adjacency matrix and relation embeddings are transported to the satellite primitive index space before comparison:

$$\mathcal{L}_{\text{graph}}^{x,s} = \text{MSE}\left(M^s, \Pi_{xs}^\top M^x \Pi_{xs}\right), \tag{17}$$

and

$$\mathcal{L}_{\text{rel}}^{x,s} = \text{MSE}\left(R^s, \mathcal{T}(R^x, \Pi_{xs})\right), \qquad \mathcal{T}(R, \Pi)_{ijc} = \sum_{k,l} \Pi_{ki} R_{klc} \Pi_{lj}. \tag{18}$$

The correspondence matrix transports both endpoints of each relation, allowing the graph topology and relation embeddings to be compared in the satellite primitive index space. We additionally impose cycle consistency on the two directional correspondences:

$$\mathcal{L}_{\text{cyc}}^{us} = \text{MSE}\left(\Pi_{us}\Pi_{su}, I_K\right), \qquad \mathcal{L}_{\text{cyc}}^{sg} = \text{MSE}\left(\Pi_{sg}\Pi_{gs}, I_K\right). \tag{19}$$

The consistency objective is $\mathcal{L}_{\text{cons}}^{e} = 0.05\mathcal{L}_{\text{cyc}}^{e} + 0.10\mathcal{L}_{\text{graph}}^{e} + 0.05\mathcal{L}_{\text{node}}^{e} + 0.05\mathcal{L}_{\text{rel}}^{e}$.

Primitive Diversity Regularization. To maintain diversity among the induced Geo-primitives and their response patterns, we penalize pairwise similarity among the RelAttn outputs Z^v and primitive-to-patch responses A^v. For $U \in \mathbb{R}^{B\times K\times d}$, we define

$$\mathcal{D}(U) = \frac{1}{BK(K-1)} \sum_{b=1}^{B} \sum_{i\neq j} \left(\mathrm{Norm}(U_{bi})^\top \mathrm{Norm}(U_{bj})\right)^2. \tag{20}$$

For an observed relation $e = (a, b)$, the diversity term is

$$\mathcal{L}^e_{\mathrm{div}} = \frac{0.05}{2} \left[\mathcal{D}(Z^a) + \mathcal{D}(Z^b)\right] + \frac{0.02}{2} \left[\mathcal{D}(A^a) + \mathcal{D}(A^b)\right]. \tag{21}$$

B.4 Primitive-Guided Coordinate Assignment

We detach A^v when constructing the reference assignment. Each primitive response is normalized over the patch dimension as

$$\bar{A}^v_{qn} = \frac{A^v_{qn}}{\sum_{m=1}^{N_v} A^v_{qm}}. \tag{22}$$

The relative primitive distribution π^v_n is then computed from $\bar{A}^v$ as defined in Primitive-Guided Coordinate Assignment.

The retention weight is computed from the overall primitive-response strength of each patch relative to the image average:

$$\rho^v_n = \mathrm{clip}\left(\frac{\sum_{q=1}^{K} \bar{A}^v_{qn}}{\frac{1}{N_v}\sum_{m=1}^{N_v}\sum_{q=1}^{K} \bar{A}^v_{qm}}, 0, 1\right). \tag{23}$$

The shared scoring head f_{score} is implemented as a two-layer MLP following LayerNorm. The dustbin score is learnable, and Sinkhorn normalization uses $\varepsilon_{\mathrm{OT}} = 1.0$ for three iterations.

B.5 Shared-Coordinate Encoding

The shared local projection f_{local} maps each patch feature to a 128-dimensional representation, followed by ℓ_2 normalization. For each shared coordinate $q \in \{1, \ldots, K\}$, assignment-weighted aggregation is implemented as

$$z^v_q = \frac{\sum_{n=1}^{N_v} P^v_{nq}\, f_{\mathrm{local}}(T^v_n)}{\sum_{n=1}^{N_v} P^v_{nq}}. \tag{24}$$

The dustbin assignment is excluded from coordinate aggregation. The K coordinate representations are concatenated in shared coordinate index order and mapped by the coordinate fusion function f_{fuse} to obtain the shared-coordinate descriptor s^v. With $K = 8$, f_{fuse} is implemented as a bias-free linear projection from 1024 to 768 dimensions.

B.6 Stage 2 Objective Details

We use PK sampling with three images per location, allowing multiple positives from the paired view within each mini-batch. For an observed relation $e = (a, b)$, let $\mathcal{P}_e(i) = \{j \mid y_j = y_i\}$ denote the positive set for query i. The directional multi-positive contrastive loss is

$$\ell_{a\to s}(U, V) = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\sum_{j\in\mathcal{P}_e(i)} \exp(U_i^\top V_j/\tau)}{\sum_{j=1}^{B} \exp(U_i^\top V_j/\tau)}, \tag{25}$$

with the symmetric form $\mathcal{L}^e_{\mathrm{MP}}(U, V) = \frac{1}{2}\left[\ell_{a\to s}(U, V) + \ell_{s\to a}(V, U)\right], \tau = 0.07$. This loss is applied to d^v for $\mathcal{L}_{\mathrm{retr}}$ and to s^v for $\mathcal{L}_{\mathrm{metric}}$. Mini-batches from the two observed relations are sampled alternately, with each iteration optimizing the corresponding relation. For $\mathcal{L}_{\mathrm{assign}}$, the per-view assignment losses are averaged over the samples and the two views in each paired mini-batch. Stage 2 optimizes f_{local}, f_{score}, f_{fuse}, and the learnable dustbin score.

C Experimental Protocol and Reproducibility Details

C.1 Disjoint-Pair Evaluation Protocol

University-1652 is reserved exclusively for final evaluation and is excluded from training, validation, hyperparameter selection, and checkpoint selection. Checkpoints are selected solely by the mean validation performance over the four observed directions, $u \to s$, $s \to u$, $s \to g$, and $g \to s$. The unobserved u–g relation is evaluated only after model selection is complete.

C.2 Data Preprocessing and Augmentation

Training. UAV and satellite images use standard geometric augmentation, with additional random rotation for satellite images. For ground images, Stage 1 generates one perspective view online from each VIGOR panorama by sampling yaw $\theta \sim \mathcal{U}(0^\circ, 360^\circ)$, horizontal field of view $f \sim \mathcal{U}(90^\circ, 150^\circ)$, and pitch $\varphi \sim \mathcal{U}(-10^\circ, 10^\circ)$. In Stage 2, each panorama is uniformly divided along its width into four fixed, non-overlapping crops, which are treated as independent training samples. In both stages, DenseUAV and VIGOR are sampled with probabilities 0.42 and 0.58, respectively, to balance supervision from the two observed relations.

Validation and Testing. VIGOR validation uses the complete panorama with deterministic resizing and normalization for both stages. Thus, the four-crop scheme is used only during Stage 2 training. University-1652 test images use the same deterministic preprocessing, with the provided ground perspective images encoded directly. For test-time horizontal-flip ensembling, the Stage 1 descriptor b^v and shared-coordinate descriptor s^v are extracted from the original and horizontally flipped images and concatenated within each branch. The resulting branch descriptors are then combined to form the final direct-sum descriptor.

C.3 Baseline Adaptation Protocol

We adapt all baselines to the same experimental setting as GeoCoRe while preserving their core representation and learning mechanisms whenever possible. For methods with view-specific branches, UAV and ground retain separate branches, while the satellite branch is shared across the two observed relations. The identity-classification objectives of the University-1652 baseline and LPN require a common set of location identities. Under the disjoint-pair setting, we replace these objectives with bidirectional contrastive learning while retaining their global and local-region representations, respectively. TransGeo, GeoDTR, and GeoDTR+ retain their original representation mechanisms and metric-learning objectives, using a shared satellite branch across the two observed relations. TransGeo uses its first-stage retrieval model, while the GeoDTR variants retain their geometric-layout modeling and augmentation strategies. Sample4Geo, LKGL, and Game4Loc likewise retain their primary representation designs, with supervision restricted to the u–s and s–g relations. DMNIL and UniABG are included as unsupervised references. For DMNIL, cross-view pseudo-label interaction is restricted to the two observed relations. UniABG follows its original learning mechanism by jointly clustering UAV and ground samples. Pseudo-correspondences are then mined from a shared satellite candidate pool, allowing broader cross-view interaction without u–g supervision.

D Additional Quantitative Results

D.1 Full Ablation Results

Primitive Reference and Coordinate Assignment. Table 4 extends the corresponding ablation in the main paper with Recall@1/5/10 and mAP for all six retrieval directions.

Dual-Edge Metric Alignment. Table 5 reports the complete Recall@1/5/10 and mAP results for all six retrieval directions under the dual-edge metric alignment ablations.

D.2 Controlled In-Domain Disjoint-Pair Evaluation

The main experiments use cross-dataset supervision, where the observed u–s and s–g relations are provided by different datasets. To separate cross-relation compositionality from cross-dataset do-

Table 4: **Full metrics for the ablation of shared primitive reference and primitive-guided coordinate assignment.** We report Recall@1/5/10 and mAP for all six retrieval directions.

Variant	Rep.	R_{obs}	R_{miss}	U2S				S2U				G2S				S2G				U2G				G2U			
				R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
Only Stage 1	Base	22.14	3.94	31.04	56.64	67.51	37.10	49.64	67.62	74.89	29.52	3.30	8.84	13.34	5.17	4.56	12.70	17.97	3.87	3.99	10.77	16.29	3.59	3.88	7.83	10.35	2.71
Edge-Specific Primitives	Shared-Coord	26.49	4.45	34.60	58.28	68.34	40.30	64.19	80.46	85.88	39.99	2.87	8.26	13.18	4.69	4.28	11.13	16.55	3.53	3.90	11.46	17.34	3.50	5.00	8.53	11.36	2.77
	Direct-Sum	29.47	5.27	41.06	65.86	75.32	46.86	66.48	81.88	87.02	44.25	4.07	11.01	15.86	6.27	6.28	14.69	20.97	4.94	5.15	13.49	19.48	4.44	5.39	10.16	13.22	3.51
w/o Primitive Guidance	Shared-Coord	27.46	4.28	37.79	61.96	71.20	43.48	64.76	80.17	84.45	38.25	3.02	7.25	11.05	4.66	4.28	12.41	16.83	3.66	4.40	11.84	17.33	3.75	4.15	8.26	11.32	2.56
	Direct-Sum	30.27	5.42	44.15	69.34	78.17	49.96	66.19	81.46	85.31	42.77	4.03	10.82	15.86	6.22	6.70	15.41	21.40	5.38	5.32	13.58	19.73	4.59	5.51	10.16	13.61	3.36
GeoCoRe	Shared-Coord	28.78	4.98	37.64	61.69	71.06	43.33	68.33	82.31	86.02	41.95	3.14	8.84	12.91	4.96	5.99	13.12	18.54	4.41	4.79	13.14	18.91	4.03	5.16	9.77	12.49	3.25
	Direct-Sum	**30.92**	**6.11**	**43.88**	**69.05**	**77.94**	**49.66**	**69.19**	**83.45**	**88.16**	**45.72**	**4.46**	**11.36**	**18.11**	**6.71**	**6.13**	**16.12**	**22.11**	**5.27**	**5.98**	**15.08**	**21.64**	**4.93**	**6.24**	**11.44**	**14.97**	**4.01**

Table 5: **Full metrics for the ablation of dual-edge metric alignment.** S and P denote shared and edge-private parameters, respectively. Base, Shared-Coord, and Direct-Sum denote the Stage 1 base descriptor b^v, the Stage-2 shared-coordinate descriptor s^v, and the final descriptor d^v, respectively.

Variant	Rep.	Cross-edge Sharing			Summary		U2S				S2U				U2G				G2U				G2S				S2G			
		f_{score}	f_{local}	f_{fuse}	R_{obs}	R_{miss}	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
Only Stage 1	Base	–	–	–	22.14	3.94	31.04	56.65	67.52	37.10	49.64	67.62	74.89	29.52	3.99	10.77	16.29	3.59	3.88	7.83	10.35	2.71	3.30	8.84	13.34	5.17	4.56	12.70	17.97	3.87
w/o Dual-Edge Metric Alignment	Shared-Coord	S	P	P	27.01	0.79	35.12	57.08	65.80	40.35	66.62	80.88	85.02	33.71	1.00	3.57	5.87	1.20	0.58	1.86	2.68	0.75	3.30	8.96	13.77	5.19	0.75	3.30	5.87	1.20
	Direct-Sum				32.83	3.94	46.94	73.07	78.64	52.40	74.32	85.73	87.93	44.66	4.12	11.10	16.58	3.77	3.76	7.33	9.50	2.69	4.07	9.58	13.77	5.93	4.01	12.14	17.64	6.53
Edge-Specific Coordinate Assignment	Shared-Coord	P	S	S	32.64	2.23	47.78	71.40	79.88	53.22	74.89	86.02	89.44	45.21	2.52	8.10	12.57	2.69	1.94	4.46	6.71	1.73	2.91	8.38	12.91	4.77	2.52	8.10	12.57	2.69
	Direct-Sum				34.73	4.93	53.56	77.34	84.77	58.92	76.18	88.45	91.73	50.65	4.55	12.74	18.62	4.31	5.31	9.77	12.64	3.56	4.03	11.32	16.63	6.37	4.55	12.74	18.62	4.31
Fully Edge-Specific Stage 2	Shared-Coord	P	P	P	31.09	0.74	44.60	67.98	76.48	50.03	73.75	85.16	89.30	41.80	0.85	3.18	5.10	1.12	0.62	1.36	1.74	0.53	2.99	8.72	13.34	4.91	0.85	3.18	5.10	1.12
	Direct-Sum				35.79	3.83	53.98	77.28	84.40	59.27	78.32	88.87	92.01	50.68	4.01	10.97	16.53	3.68	3.64	7.29	9.77	2.72	4.01	10.97	16.53	3.68	4.58	11.67	16.87	6.82
GeoCoRe	Shared-Coord	S	S	S	28.78	4.98	37.64	61.69	71.06	43.33	68.33	82.31	86.02	41.95	4.79	13.14	18.91	4.03	5.16	9.77	12.49	3.25	4.46	11.36	18.11	6.71	3.14	8.84	12.91	4.96
	Direct-Sum				**30.92**	**6.11**	43.88	69.05	77.94	49.66	69.19	83.45	88.16	45.72	**5.98**	15.08	21.64	4.93	**6.24**	11.44	14.97	4.01	4.46	11.36	18.11	6.71	6.13	16.12	22.11	5.27

main differences, we further construct a controlled in-domain disjoint-pair setting within University-1652. The location set is split into two disjoint subsets for the two observed relations, satisfying $Y_{us} \cap Y_{sg} = \varnothing$. The u–s and s–g pairs are sampled from the corresponding subsets, while all u–g correspondences are excluded from training. As shown in Table 6, we compare *Pairwise*, *Stage 1*, and *GeoCoRe*. *Pairwise* learns the two observed relations using separate pairwise supervision without cross-edge coupling. GeoCoRe increases R_{miss} from 7.20% for *Pairwise* to 10.83% under the controlled in-domain setting. The improvement therefore persists after removing cross-dataset domain differences, providing additional evidence for missing-relation recovery under disjoint pairwise supervision. Compared with *Stage 1*, GeoCoRe further increases R_{miss} from 7.44% to 10.83% and mAP_{miss} from 5.64% to 7.85%. This further improvement supports the contribution of shared-coordinate mapping and common metric alignment to missing-relation recovery.

Table 6: **Full retrieval results under the controlled in-domain disjoint-pair setting.** R_{obs} denotes the average R@1 over the four observed directions, and R_{miss} denotes the average R@1 over the two missing directions.

Method	R_{obs}	R_{miss}	U2S				S2U				G2S				S2G				U2G				G2U			
			R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
Pairwise	34.56	7.20	55.98	81.96	89.05	61.80	73.75	86.59	92.15	53.97	3.37	10.70	16.09	5.64	5.14	14.27	19.83	4.63	7.22	18.49	25.79	5.99	7.17	12.91	17.10	5.13
Stage 1	34.38	7.44	56.79	80.94	87.80	62.21	73.32	84.17	89.87	53.86	3.14	10.43	15.35	5.44	4.28	13.27	18.97	3.89	7.10	18.92	26.70	6.07	7.79	14.42	18.88	5.21
GeoCoRe Full	**42.58**	**10.83**	**73.58**	**91.51**	**94.86**	**77.56**	**86.45**	**93.72**	**95.44**	**71.92**	**4.46**	**13.14**	**20.01**	**7.22**	**5.85**	**17.12**	**24.96**	**5.63**	**10.30**	**23.16**	**31.35**	**8.24**	**11.36**	**17.56**	**21.75**	**7.46**

E Analysis of Shared Primitive Reference and Coordinate Assignment

The shared Geo-primitives establish a shared primitive reference that subsequently guides coordinate assignment. Figure 5 visualizes the progression from primitive responses to the reference assignment Q^v and final coordinate assignment P^v across different locations.

The primitive-response overlays exhibit view-dependent activation patterns, with the same primitive indices appearing at different spatial positions across satellite, UAV, and ground views. The reference assignment Q^v transfers these responses to the shared coordinate indices, but the resulting assignments remain relatively fragmented in several regions. After refinement to P^v, the coordinate assignments generally form larger and more spatially coherent regions, while recurring coordinate indices are retained across views despite changes in their spatial extent and position. This pattern is consistent across the six examples, indicating that the shared primitive reference provides a common

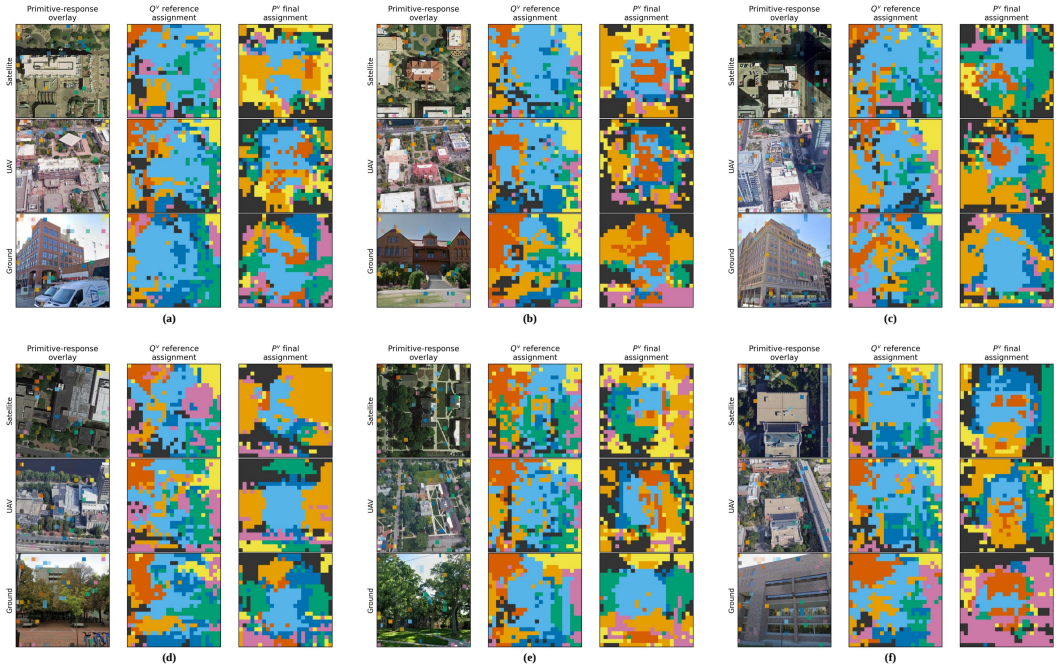


Figure 5: **Visualization of shared primitive reference and coordinate assignment.** For each location, the three rows correspond to satellite, UAV, and ground views. The first column shows the primitive-response overlay, where colors indicate shared Geo-primitive indices. The second column presents the primitive-guided reference assignment Q^v derived from the primitive responses, and the third column shows the final content-adaptive coordinate assignment P^v. The examples illustrate how the shared primitive reference guides coordinate assignment across views under substantial viewpoint variation.

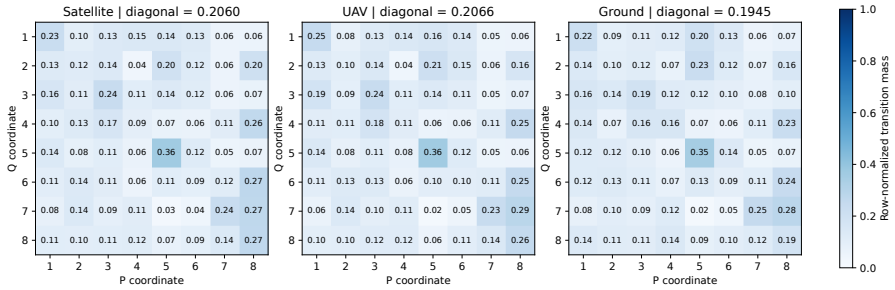


Figure 6: **Transition matrices from reference assignment Q^v to final coordinate assignment P^v.** Rows and columns denote shared coordinate indices in Q^v and P^v, respectively. The consistent diagonal bias across views indicates that the final assignment remains coupled to the primitive-guided reference while allowing redistribution across coordinates.

coordinate basis, while the subsequent assignment step reorganizes the view-dependent responses into more coherent shared-coordinate partitions.

The transition statistics in Figure 6 further quantify the refinement from Q^v to P^v. The row-normalized transition matrices measure how assignment mass is redistributed from the reference assignment Q^v to the final coordinate assignment P^v. Across satellite, UAV, and ground views, the diagonal transition mass remains consistently above the uniform level, reaching 0.2060, 0.2066, and 0.1945, respectively. At the same time, substantial off-diagonal mass remains, indicating that P^v retains a preference for the corresponding coordinate indices while allowing content-adaptive redistribution across coordinates. The qualitative examples in Figure 5 show a similar pattern. Co-

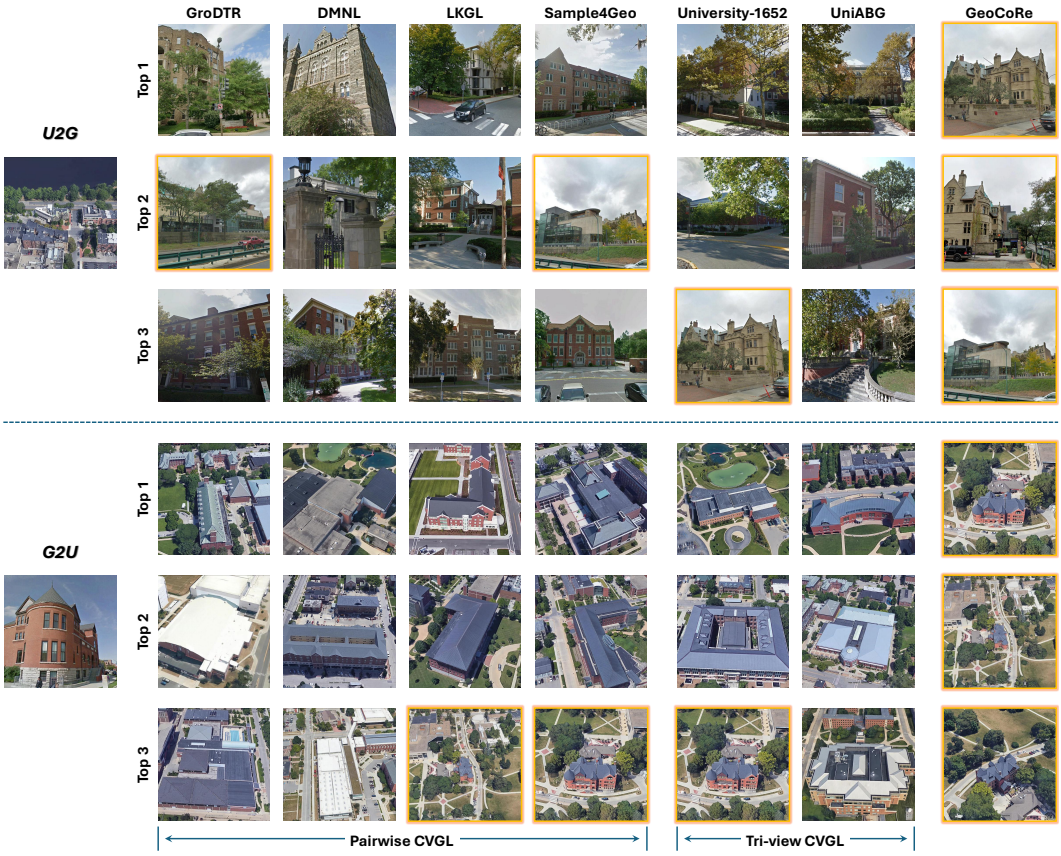


Figure 7: **Visualization of geo-localization results on the unobserved u–g relation.**

ordinate indices that cover salient scene regions in Q^v often remain associated with related regions in P^v, although their spatial extent and location vary across views. The refinement from Q^v to P^v also reorganizes fragmented assignments into more spatially coherent regions in several examples.

Together, these observations show that the shared primitive reference provides a persistent reference for coordinate assignment, while the subsequent assignment step adapts this reference to view-dependent observations. This behavior is consistent with the role of Stage 2 in organizing view-specific representations under shared coordinates.

F Additional Qualitative Results and Failure Cases

F.1 Visualization of Missing-Edge Geo-localization Results

Figure 7 shows representative geo-localization results on the unobserved u–g relation. Across both U2G and G2U, several pairwise and tri-view baselines retrieve candidates with similar scene or architectural appearance, while ground-truth positives often appear at lower ranks. In the shown examples, GeoCoRe ranks the ground-truth positives higher in both directions. These qualitative results are consistent with the quantitative improvements on the unobserved u–g relation.

F.2 Failure Cases and Analysis

Figure 8 shows representative failure cases for both UAV-to-ground and ground-to-UAV retrieval. Cases 1 and 2 exhibit large viewpoint differences with limited cross-view overlap. In Case 1, the UAV query mainly captures surrounding layout and roof-level geometry, whereas the ground image contains only a local facade. Case 2 shows the reverse situation, where the ground query is dominated by a largely textureless facade while the corresponding UAV image contains roof geometry



Figure 8: **Representative failure cases on the unobserved u–g relation.** Each row presents the query, the ground-truth positive, and the top three retrieved false positives.

and site layout. This limited overlap increases ambiguity among visually similar candidates. Case 3 illustrates ambiguity caused by repeated architectural patterns. The query and retrieved candidates share similar red-brick buildings, roof structures, and surrounding vegetation, providing limited location-specific cues for distinguishing the correct match. Case 4 contains limited location-specific information in the ground view, where vegetation occupies much of the image and only a small portion of the target building is visible. The corresponding UAV image contains a broader spatial layout, while the retrieved false positives share similar vegetation and campus context. These cases illustrate recurring difficulties associated with limited cross-view overlap, repeated architectural patterns, and weak location-specific cues in recovery of the unobserved u–g relation.

G Scope and Limitations

The current study focuses on the three-view chain $u{-}s{-}g$, where two observed relations are used to recover a single missing relation. This setting directly captures the cross-relation compositionality studied in this work, while more general N-view relation graphs remain unexplored. With additional views or multiple unobserved relations, the observed-edge pattern and the availability of multiple compositional paths may introduce additional challenges for establishing cross-edge comparability. Extending the current formulation to these more general settings is left for future work.