CVGL-CUEBENCH: A BENCHMARK FOR VISUAL-CUE DEPENDENCE AND REPRESENTATION TRANSFER IN CROSS-VIEW GEO-LOCALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-view geo-localization (CVGL) models achieve high retrieval recall on benchmark RGB ground queries. However, standard RGB images intrinsically couple surface appearance with spatial geometric layouts. Evaluating RGB inputs alone obscures which visual cues actually drive alignment, as well as how sensitively models rely on each cue. To systematically diagnose visual cue reliance and model sensitivity, we introduce CVGL-CueBench, a multi-representation diagnostic benchmark comprising two sub-benchmarks, CVUSA-Cue and CVACT-Cue, derived from CVUSA and CVACT. While retaining ground–aerial spatial correspondences and the invariant RGB aerial gallery, CVGL-CueBench provides five parallel query representations: the original RGB image alongside four derived structural representations (class-agnostic region segmentation, estimated relative depth, edge maps, and sketches). We establish a systematic diagnostic protocol evaluating five representative CVGL methods across representation-matched training ($A \rightarrow A$) to measure localization capacity preserved within each visual cue, and cross-representation transfer ($A \rightarrow B$) to assess sensitivity to representation shifts, complemented by test-time spatial perturbations on an RGB baseline to examine reliance on high-frequency textures, spatial layout, and regional context. Evaluations demonstrate that derived structural representations retain sufficient geometric cues to support high retrieval recall when paired with capable backbones: on north-aligned CVUSA-Cue, Sample4Geo achieves 90.13%–98.68% Recall@1 across all query representations. However, cross-representation transfer ($RGB \rightarrow B$) reveals pronounced sensitivity, varying from 68.12% on sketches to a near-total collapse of 0.06% on relative depth queries. Furthermore, under a 70° field of view on CVACT-Cue, models trained on sketch queries surpass their RGB counterparts across all five evaluated methods by 0.96–15.48 percentage points, though this ranking advantage varies across methods on CVUSA-Cue. These findings demonstrate that high RGB recall masks acute sensitivity to representation shifts, while confirming that structural representations retain sufficient geometric cues to support cross-view alignment, particularly under restricted viewing constraints. CVGL-CueBench thus provides a controlled diagnostic foundation to decouple visual cues and evaluate structural robustness in cross-view geo-localization.

1 INTRODUCTION

Cross-view geo-localization (CVGL) provides an image-based alternative for estimating geographic location when GPS signals are unstable or unavailable. The task matches ground images with a geo-tagged aerial image database and returns the aerial reference that corresponds to the same place (Hu et al., 2022; Shi et al., 2019; Bansal et al., 2011). This capability can support applications such as autonomous driving (Hu et al., 2023), unmanned aerial vehicle navigation (Ryll et al., 2020), and search-and-rescue operations in GPS-denied disaster environments (Nasar et al., 2023).

Most CVGL methods address cross-view differences by learning a joint representation for ground and aerial images. Early works establish paired-image matching with convolutional networks, while subsequent studies introduce local feature aggregation, transformer architectures, and metric-learning objectives (Hu et al., 2018; Zhu et al., 2022). Other approaches leverage orientation encoding,

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.

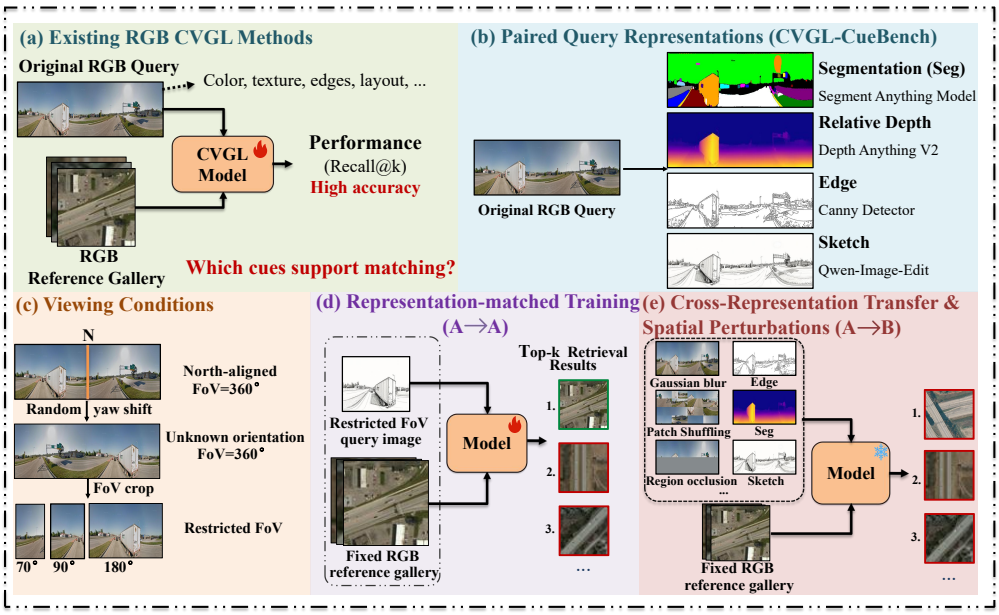


Figure 1: **Overview of the CVGL-CueBench evaluation framework. (a)** Existing CVGL pipelines evaluate models on standard RGB queries where visual cues are entangled. **(b)** CVGL-CueBench decouples ground observations into five parallel representations. **(c)** Realistic viewing configurations incorporating canonical north-aligned panoramas, unknown orientation, and restricted FoVs. **(d)** Representation-matched training evaluating the localization capacity preserved within each cue against invariant RGB aerial galleries. **(e)** Cross-representation transfer and test-time spatial perturbations on RGB queries probing sensitivity to representation shifts and spatial layout dependencies.

geometric layouts, semantic structures, or region selection to make correspondences more reliable (Liu & Li, 2019; Shi et al., 2019; Zhai et al., 2017; Shen et al., 2026; Rodrigues & Tani, 2021; Li et al., 2024), and recent studies have extended evaluations to variations in viewing direction, field of view (FoV), and image quality (Mi et al., 2024; Zhang & Zhu, 2024; Wu et al., 2026). These advances have driven CVGL models to achieve remarkable retrieval recall on benchmark RGB ground queries. However, as illustrated in Figure 1(a), standard RGB inputs intrinsically couple surface appearances with spatial geometric layouts. Evaluating RGB inputs alone therefore obscures which visual cues actually drive alignment, as well as how sensitively models rely on each cue.

To systematically examine cue reliance and model sensitivity, we introduce **CVGL-CueBench**, a multi-representation diagnostic benchmark constructed from CVUSA (Zhai et al., 2017) and CVACT (Liu & Li, 2019), comprising two sub-benchmarks: **CVUSA-Cue** and **CVACT-Cue**. For each ground scene, CVGL-CueBench provides five parallel query representations (Figure 1(b)): the original RGB image alongside four derived structural representations that reduce surface appearance details to emphasize distinct aspects of scene structure. Specifically, **edge** maps (Canny, 1986) record high-gradient intensity transitions as binary contours, retaining salient boundary contours while eliminating chromatic information and continuous shading gradients. **Estimated relative depth** maps (Yang et al., 2024) encode distance ordering relative to the camera through min–max normalized pseudo-color fields, reflecting relative surface ordering in the absence of absolute metric scale. **Class-agnostic region segmentations** (Kirillov et al., 2023) partition the ground scene into discrete region masks rendered via a cyclical color palette, delineating two-dimensional spatial layouts and boundary arrangements without assigning semantic class labels. Finally, generated **sketches** (Wu et al., 2025) depict the scene through line-based strokes that delineate prominent structural outlines while suppressing dense photographic textures. By altering the visual properties of ground queries while maintaining the invariant RGB aerial gallery and ground-truth spatial correspondences, CVGL-CueBench enables controlled evaluation of model behavior across distinct visual representations.

With these parallel representations established, two primary questions emerge: whether structural cues independently retain sufficient information to support cross-view matching, and whether models can generalize across representations without retraining. Motivated by diagnostic paradigms in shortcut learning and visual cue preference (Geirhos et al., 2018; 2020), we address both questions

by formulating an evaluation protocol that pairs **representation-matched training** ($A \to A$) with **cross-representation transfer** ($A \to B$, where $A \neq B$) against an invariant RGB aerial gallery. In the matched setting (Figure 1(d)), models are trained and evaluated on the identical query representation to measure the retrieval performance achievable with each representation under the evaluated training setup. In the transfer setting (Figure 1(e)), a model trained on representation A is evaluated on representation B without updating network parameters to probe sensitivity to representation shifts. Complementing evaluations across these derived representations, we investigate how model behavior changes when spatial structure is disrupted within the RGB queries. As a diagnostic case study on Sample4Geo (Deuser et al., 2023), the protocol introduces test-time spatial perturbations—Gaussian blur, patch shuffling, and region occlusion—to observe how cross-view matching responds when fine details are smoothed, spatial arrangements are scrambled, or view sectors are obstructed.

In accordance with canonical benchmark configurations in CVUSA (Zhai et al., 2017) and CVACT (Liu & Li, 2019), CVGL-CueBench initially provides north-aligned, 360° panoramic ground queries. In practical deployments, however, ground cameras frequently operate with arbitrary orientations and constrained angular spans. To separately examine how reliance on structural cues depends on directional alignment and visual context, the evaluation extends beyond canonical panoramas to encompass both unknown orientations and restricted fields of view ($70^\circ, 90^\circ, 180^\circ$; Figure 1(c)). Across these settings, we benchmark five representative methods across five query representations, five viewing conditions, and two benchmark datasets under representation-matched training ($A \to A$) and cross-representation transfer ($A \to B$). Complemented by the test-time spatial perturbations on an RGB baseline, this systematic evaluation spans 250 ($5 \times 5 \times 5 \times 2$) primary experimental configurations to identify which visual cues drive cross-view geo-localization and determine how this cue reliance shifts under directional and contextual constraints. Our contributions are as follows:

- Retrieval recall alone on standard RGB queries with coupled cues cannot reveal which visual cues support localization or how models rely on individual scene properties. To address this limitation, we introduce **CVGL-CueBench** (comprising CVUSA-Cue and CVACT-Cue), providing 444,160 ground queries across five parallel representations (original RGB alongside class-agnostic region segmentation, estimated relative depth, edge maps, and sketches) evaluated against invariant RGB aerial reference galleries.
- To decouple intrinsic representation capacity from model generalization, we formulate an evaluation protocol combining representation-matched training ($A \to A$) to measure localization capacity within each visual cue, with cross-representation transfer ($A \to B$) to evaluate the sensitivity of models to representation shifts without parameter updates. Furthermore, the protocol applies test-time spatial perturbations to an RGB baseline to isolate reliance on fine boundary details, global spatial layout, and localized feature visibility.
- Across five representative methods evaluated over 250 configurations (spanning all representations, viewing conditions, and datasets), we reveal that: (1) derived structural representations support high retrieval recall with the evaluated capable methods under north-aligned panoramic conditions; (2) RGB-trained models exhibit acute sensitivity under cross-representation transfer to structural cues; and (3) under restricted FoV (70° FoV) on CVACT-Cue, models trained on sketch queries surpass their RGB counterparts across all five evaluated methods, though this advantage varies across methods on CVUSA-Cue.

2 Related Work

Cross-view representation learning. Early approaches (Zhai et al., 2017) establish cross-view feature translation using Convolutional Neural Networks (Krizhevsky et al., 2012). Subsequent efforts focus on advanced feature aggregation and loss formulations: CVM-Net (Hu et al., 2018) introduces NetVLAD for viewpoint-invariant pooling, L2LTR (Yang et al., 2021) integrates CNN backbones with Transformers, TransGeo (Zhu et al., 2022) deploys a pure Vision Transformer architecture, and Sample4Geo (Deuser et al., 2023) designs geographically informed hard-negative mining within a contrastive learning objective.

Spatial and semantic information. Large perspective disparities across views make scene layout and semantic structure highly valuable for cross-view correspondence. Orientation encoding (Liu & Li, 2019), polar transformation and spatial-aware feature aggregation (Shi et al., 2019) explicitly exploit directional information and spatial relationships. GeoDTR+ (Zhang et al., 2024) directly

models geometric layout to mitigate reliance on low-level appearance features. Complementary approaches predict ground-level semantic layouts from aerial imagery (Zhai et al., 2017) or introduce semantic supervision to align spatial features (Shen et al., 2026). Furthermore, semantic object masking (Rodrigues & Tani, 2021) addresses temporal appearance changes, while 4SCIG (Li et al., 2024) suppresses non-informative sky regions to reduce feature interference.

Robustness in CVGL. Ground queries frequently deviate from training distributions in viewing direction and visual quality. Existing works address viewing variations through orientation-aware matching (Shi et al., 2020), patch-level alignment (Li & Zhu, 2025), consistency learning (Mi et al., 2024), and progressive training (Chen et al., 2026). Regarding image quality, Zhang & Zhu (2024) establish a benchmark of common environmental corruptions on CVUSA (Zhai et al., 2017) and CVACT (Liu & Li, 2019) while holding aerial references fixed. Building on this, MRGeo (Wu et al., 2026) enhances corrupted queries through spatial and channel feature refinement.

Shortcut learning and visual-cue reliance. Recent findings in visual recognition reveal that high in-distribution benchmark accuracy often masks model vulnerabilities, as deep networks exploit shortcut features rather than invariant semantic structures (Geirhos et al., 2020). Specifically, Geirhos et al. (2018) demonstrate that standard ImageNet-trained CNNs exhibit a pronounced texture bias over shape, a vulnerability that can be revealed through controlled representation shifts. These insights highlight the necessity of examining visual-cue dependence in CVGL, without presupposing that cross-view models rely on the same invariant properties across extreme viewpoint shifts.

Distinction from prior work. Existing CVGL benchmarks evaluate models almost exclusively on standard RGB queries where surface appearances and spatial geometry are conflated, obscuring which cues actually support cross-view matching. While Zhang & Zhu (2024) examine pixel-level resilience against environmental noise, **CVGL-CueBench** investigates visual cue reliance and structural robustness. By evaluating five parallel query representations across representation-matched training, cross-representation transfer, and targeted baseline spatial perturbations, this work decouples intrinsic representation capacity from model generalization.

3 Benchmark Design and Construction

3.1 Problem Formulation

Let $\mathcal{Q} = \{q_i\}_{i=1}^{N_q}$ denote the set of ground queries and $\mathcal{R} = \{r_j\}_{j=1}^{N_r}$ the georeferenced RGB aerial gallery. For each ground query q_i, let $\pi(i) \in \{1, \dots, N_r\}$ denote the index of the corresponding ground-truth aerial reference. For each query representation A, let q_i^A denote the query image derived from q_i, with $q_i^{\mathrm{RGB}} = q_i$. Each derived query is paired with the identical aerial reference $r_{\pi(i)}$ as the original query. Let f_q^A and f_r^A denote the query and aerial encoders of a model trained using query representation A alongside RGB aerial images. When evaluating this model on query representation B, the retrieval score between q_i^B and r_j is formulated as:

$$S_{ij}^{A \to B} = s\big(f_q^A\big(q_i^B\big), f_r^A(r_j)\big), \tag{1}$$

where $s(\cdot, \cdot)$ denotes the cosine similarity function between feature embeddings. The aerial gallery remains RGB across all settings. When $A = B$, the model is trained and evaluated within the identical query representation, measuring localization capacity under **representation-matched training** (Figure 1(d)). When $A \neq B$, the model trained on representation A is evaluated directly on representation B without updating network parameters, measuring sensitivity to representation shifts under **cross-representation transfer** (Figure 1(e)). Evaluating against an invariant RGB aerial gallery reflects practical operational deployments, where existing satellite databases cannot be dynamically converted into derived representations; further rationale is detailed in Appendix Section D.

3.2 Query Representations

To evaluate cross-view retrieval across distinct descriptions of identical ground scenes, we establish five parallel query representations (Figure 1(b)):

$$\mathcal{V} = \{\mathrm{RGB}, \mathrm{Seg}, \mathrm{Depth}, \mathrm{Edge}, \mathrm{Sketch}\}. \tag{2}$$

RGB denotes the source panorama, which serves as the original-image query baseline and the input for generating the four derived query representations. **Edge** queries are generated by applying the

Canny detector (Canny, 1986) to grayscale ground panoramas with thresholds of 100 and 200; the binary outputs are inverted to display dark boundaries on a light background and saved as grayscale images. **Sketch** queries are generated using `Qwen-Image-Edit-2511` (Wu et al., 2025); the prompt requests a pencil sketch, prohibits adding or removing objects, and instructs the model not to depict content absent from the input image (Appendix Section E). **Depth** queries are generated using Depth Anything V2 (Yang et al., 2024) with a ViT-B backbone; each relative-depth prediction is independently min–max normalized to $[0, 255]$, and visualized using `plasma` colormap, without absolute metric calibration. **Seg** queries are produced using the Segment Anything Model (Kirillov et al., 2023) with a ViT-H backbone and its automatic mask generator; class-agnostic mask proposals are rendered into a color region map using a fixed cyclic palette indexed by mask return order.

3.3 Spatial Perturbations

Complementing representation-level comparisons, we introduce a diagnostic case study to observe how retrieval responds when spatial structure is corrupted within the RGB queries. Using an RGB-trained Sample4Geo baseline (Deuser et al., 2023) evaluated at test time without fine-tuning, we apply three targeted perturbation operations (Figure 1(e)): **Gaussian blur.** We apply 2D Gaussian filtering with square kernel dimensions of 3×3, 5×5, 9×9, and 15×15 (with 1×1 denoting the unperturbed baseline) to progressively smooth fine details and boundary transitions while retaining global scene layout. **Patch shuffling.** We divide each resized query image into regular non-overlapping grid partitions of 2×2, 4×4, and 8×8, and randomly permute the patch order while preserving pixel arrangements within each individual patch, disrupting global spatial arrangements while preserving local patch content. **Region occlusion.** We evaluate localized feature obstruction across six spatial regions: top, middle, bottom, left, center, and right. For vertical cuts (top, middle, bottom), masks span the full image width and a fraction ρ of the height; for horizontal cuts (left, center, right), masks span the full height and a fraction ρ of the width. Middle and center masks are centered symmetrically along the vertical and horizontal axes, respectively. We evaluate masking ratios $\rho \in \{0.10, 0.33, 0.50\}$, setting occluded pixels to the channel-wise mean values of the query image.

3.4 Viewing Conditions and Transformation Pipeline

To examine whether reliance on visual cues changes under directional alignment and visual context, the evaluation incorporates five distinct viewing conditions: north-aligned panoramas ($360°$), unknown-orientation panoramas ($360°$), and restricted FoVs of $180°$, $90°$, and $70°$ (Figure 1(c)). For unknown orientations, we uniformly sample an azimuth rotation angle $\theta \sim \mathcal{U}(0°, 359°)$ and apply the corresponding circular horizontal pixel shift to the ground panorama. For a target FoV $\phi \in \{70°, 90°, 180°\}$, we extract the leftmost $\lfloor W \cdot \phi/360 \rfloor$ columns, where W denotes the full panorama width. Across all FoV evaluations, training and test viewing spans remain strictly matched.

Standardized transformation pipeline. To standardize downstream geometric processing across all representations $\mathcal{V}$, all derived structural representations (Edge, Depth, Seg, Sketch) are generated offline from complete $360°$ panoramic ground images and stored as fixed dataset inputs. Subsequent transformations—including resolution resizing, orientation shifts, and FoV truncation—are applied dynamically using identical geometric parameters across all representations. The offline generation pipelines are independent of downstream localization training and receive no gradients from the localization networks. This protocol evaluates limited-FoV crops of precomputed panoramic representations, with the FoV restriction applied at the input to the downstream localization network.

4 Benchmark and Experimental Setup

4.1 Benchmark Composition and Splits

CVGL-CueBench comprises two sub-benchmarks, **CVUSA-Cue** and **CVACT-Cue**, derived respectively from CVUSA (Zhai et al., 2017) and CVACT (Liu & Li, 2019). Each sub-benchmark provides five parallel query representations: the original RGB image alongside class-agnostic region segmentation, estimated relative depth, edge maps, and sketches (Section 3.2). All queries strictly retain split assignments and ground-to-aerial spatial correspondences, evaluated against invariant RGB aerial reference galleries. CVUSA contains 35,532 training pairs and 8,884 test pairs. For CVACT,

Table 1: **Evaluation results under representation-matched training ($A \rightarrow A$) on CVUSA-Cue and CVACT-Cue.** All ground queries are evaluated in a north-aligned panoramic setting against invariant RGB aerial references. Subscripts denote percentage-point differences relative to the corresponding RGB baseline; red and green indicate performance drops and gains, respectively.

Method	Backbone	Query	CVUSA-Cue				CVACT-Cue			
			R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
TransGeo (Zhu et al., 2022)	DeiT	RGB	94.08	98.36	99.04	99.77	84.95	94.14	95.78	98.37
		Sketch	88.65$_{-5.43}$	96.25$_{-2.11}$	97.65$_{-1.39}$	99.52$_{-0.25}$	79.51$_{-5.44}$	91.80$_{-2.34}$	94.15$_{-1.63}$	97.86$_{-0.51}$
		Edge	79.29$_{-14.79}$	93.19$_{-5.17}$	95.77$_{-3.27}$	99.33$_{-0.44}$	45.21$_{-39.74}$	70.15$_{-23.99}$	81.25$_{-14.53}$	92.56$_{-5.81}$
		Depth	81.35$_{-12.73}$	92.67$_{-5.69}$	95.11$_{-3.93}$	98.97$_{-0.80}$	70.74$_{-14.21}$	84.84$_{-9.30}$	88.52$_{-7.26}$	93.56$_{-4.81}$
		Seg	80.58$_{-13.50}$	93.76$_{-4.60}$	95.89$_{-3.15}$	99.41$_{-0.36}$	77.01$_{-7.94}$	90.24$_{-3.90}$	92.91$_{-2.87}$	97.62$_{-0.75}$
GeoDTR (Zhang et al., 2023)	ResNet-34	RGB	93.76	98.47	99.22	99.85	85.43	94.81	96.11	98.26
		Sketch	89.36$_{-4.40}$	97.26$_{-1.21}$	98.42$_{-0.80}$	99.81$_{-0.04}$	77.74$_{-7.69}$	91.24$_{-3.57}$	93.61$_{-2.50}$	98.15$_{-0.11}$
		Edge	82.23$_{-11.53}$	94.96$_{-3.51}$	96.59$_{-2.63}$	99.57$_{-0.28}$	76.86$_{-8.57}$	90.63$_{-4.18}$	93.49$_{-2.62}$	97.94$_{-0.32}$
		Depth	80.32$_{-13.44}$	93.28$_{-5.19}$	95.93$_{-3.29}$	98.14$_{-1.71}$	71.03$_{-14.40}$	87.80$_{-7.01}$	91.14$_{-4.97}$	97.29$_{-0.97}$
		Seg	80.68$_{-13.08}$	93.82$_{-4.65}$	96.18$_{-3.04}$	99.48$_{-0.37}$	77.55$_{-7.88}$	91.20$_{-3.61}$	93.83$_{-2.28}$	97.89$_{-0.37}$
Sample4Geo (Deuser et al., 2023)	ConvNeXt-B	RGB	98.68	99.68	99.78	99.87	90.81	96.74	97.48	98.77
		Sketch	95.95$_{-2.73}$	99.13$_{-0.55}$	99.54$_{-0.24}$	99.83$_{-0.04}$	88.57$_{-2.24}$	94.93$_{-1.81}$	96.10$_{-1.38}$	98.23$_{-0.54}$
		Edge	92.57$_{-6.11}$	98.25$_{-1.43}$	98.73$_{-1.05}$	99.59$_{-0.28}$	85.78$_{-5.03}$	94.00$_{-2.74}$	95.35$_{-2.13}$	97.71$_{-1.06}$
		Depth	90.13$_{-8.55}$	97.10$_{-2.58}$	98.13$_{-1.65}$	99.50$_{-0.37}$	85.62$_{-5.19}$	93.62$_{-3.12}$	94.78$_{-2.70}$	97.61$_{-1.16}$
		Seg	91.72$_{-6.96}$	97.61$_{-2.07}$	98.68$_{-1.10}$	99.63$_{-0.24}$	86.47$_{-4.34}$	94.41$_{-2.33}$	95.67$_{-1.81}$	97.90$_{-0.87}$
ConGeo (Mi et al., 2024)	ConvNeXt-B	RGB	98.27	99.59	99.70	99.86	90.12	95.69	96.56	98.24
		Sketch	96.40$_{-1.87}$	98.86$_{-0.73}$	99.30$_{-0.40}$	99.70$_{-0.16}$	85.19$_{-4.93}$	92.62$_{-3.07}$	94.30$_{-2.26}$	97.37$_{-0.87}$
		Edge	94.00$_{-4.27}$	97.91$_{-1.68}$	98.60$_{-1.10}$	99.47$_{-0.39}$	81.70$_{-8.42}$	91.51$_{-4.18}$	93.64$_{-2.92}$	97.25$_{-0.99}$
		Depth	91.29$_{-6.98}$	96.55$_{-3.04}$	97.57$_{-2.13}$	99.13$_{-0.73}$	82.27$_{-7.85}$	91.88$_{-3.81}$	93.69$_{-2.87}$	97.28$_{-0.96}$
		Seg	91.36$_{-6.91}$	96.80$_{-2.79}$	97.73$_{-1.97}$	99.23$_{-0.63}$	82.74$_{-7.38}$	91.77$_{-3.92}$	93.69$_{-2.87}$	97.18$_{-1.06}$
SinGeo (Chen et al., 2026)	ConvNeXt-B	RGB	96.90	99.04	99.26	99.71	85.13	92.73	94.34	97.27
		Sketch	95.02$_{-1.88}$	98.40$_{-0.64}$	98.90$_{-0.36}$	99.59$_{-0.12}$	84.96$_{-0.17}$	92.41$_{-0.32}$	94.05$_{-0.29}$	97.14$_{-0.13}$
		Edge	89.99$_{-6.91}$	96.27$_{-2.77}$	97.59$_{-1.67}$	99.07$_{-0.64}$	77.56$_{-7.57}$	89.05$_{-3.68}$	91.74$_{-2.60}$	96.37$_{-0.90}$
		Depth	89.09$_{-7.81}$	95.22$_{-3.82}$	96.51$_{-2.75}$	98.54$_{-1.17}$	81.37$_{-3.76}$	90.33$_{-2.40}$	92.58$_{-1.76}$	96.42$_{-0.85}$
		Seg	86.71$_{-10.19}$	94.82$_{-4.22}$	96.37$_{-2.89}$	98.61$_{-1.10}$	82.37$_{-2.76}$	91.25$_{-1.48}$	93.12$_{-1.22}$	96.57$_{-0.70}$

following established benchmark conventions (Liu & Li, 2019), we adopt the 35,532 training pairs and the 8,884 validation pairs (CVACT_val) as the standardized test split. Spanning five parallel representations across both splits, CVGL-CueBench encompasses 444,160 ground queries in total. Qualitative query visualizations across all representations are provided in Appendix Section C.

4.2 Evaluation Protocol and Implementation Details

Evaluation metrics. Following the standard CVGL evaluation setting, we report top-K retrieval recall (R@K), which measures the percentage of ground queries for which the ground-truth aerial reference is retrieved within the top K candidates. We report R@1, R@5, R@10, and R@1%, where R@1% denotes retrieval within the top 1% of the entire reference gallery.

Implementation details. For all methods, we adopt their official open-source implementations, native input resolutions, and original training hyperparameters. Models are trained independently on each query representation for representation-matched evaluations ($A \rightarrow A$), while cross-representation transfer ($A \rightarrow B$) and test-time spatial perturbations are evaluated directly without fine-tuning. All experiments are conducted on NVIDIA A100-SXM4-40GB GPUs using PyTorch.

5 Benchmarking Results

5.1 Evaluation on North-Aligned Panoramas

Table 1 reports cross-view retrieval performance across the five evaluated methods under representation-matched training ($A \rightarrow A$) on north-aligned 360° panoramas.

Information sufficiency of structural representations. While RGB queries achieve the highest recall in Table 1, structural representations preserve sufficient geometric cues to support accurate retrieval even when natural chromaticity and surface textures are attenuated. With capable architectures such as Sample4Geo, structural representations attain 90.13%–95.95% R@1 on CVUSA-Cue and 85.62%–88.57% on CVACT-Cue. Across all five methods, Sketch achieves the highest R@1 among the non-RGB representations in Table 1. The remaining structural cues (Edge, Depth, Seg) sustain strong retrieval performance, surpassing 90% on CVUSA-Cue and 85% on CVACT-Cue with Sample4Geo. These results indicate that spatial layout, distance ordering, and boundary contours independently provide sufficient discriminatory cues for cross-view alignment, demonstrating that photographic textures are not indispensable for localization. This geometric sufficiency highlights potential for privacy-preserving localization and bandwidth-efficient transmission, where uploading or storing raw photographic imagery is restricted or telemetrically constrained.

Table 2: **Evaluation results under representation-matched training ($A \rightarrow A$) with unknown orientation and $70°$ FoV.** Ground queries are evaluated under matched unknown orientation and restricted $70°$ FoV on CVUSA-Cue and CVACT-Cue against invariant RGB aerial galleries.

Method	Backbone	Query	CVUSA-Cue				CVACT-Cue			
			R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
TransGeo (Zhu et al., 2022)	DeiT	RGB	24.08	46.63	56.84	84.32	7.20	20.05	28.16	62.06
		Sketch	19.01$_{-5.07}$	41.76$_{-4.87}$	52.93$_{-3.91}$	86.24$_{+1.92}$	10.56$_{+3.36}$	29.17$_{+9.12}$	40.33$_{+12.17}$	77.70$_{+15.64}$
		Edge	5.72$_{-18.36}$	16.02$_{-30.61}$	23.23$_{-33.61}$	56.07$_{-28.25}$	1.21$_{-5.99}$	4.96$_{-15.09}$	8.01$_{-20.15}$	30.69$_{-31.37}$
		Depth	5.60$_{-18.48}$	15.41$_{-31.22}$	21.96$_{-34.88}$	54.32$_{-30.00}$	1.37$_{-5.83}$	5.44$_{-14.61}$	8.71$_{-19.45}$	32.63$_{-29.43}$
		Seg	4.82$_{-19.26}$	13.45$_{-33.18}$	19.90$_{-36.94}$	51.03$_{-33.29}$	2.79$_{-4.41}$	8.72$_{-11.33}$	13.14$_{-15.02}$	42.70$_{-19.36}$
GeoDTR (Zhang et al., 2023)	ResNet-34	RGB	9.78	26.15	37.15	77.19	3.86	11.89	18.51	53.91
		Sketch	11.58$_{+1.80}$	29.85$_{+3.70}$	40.39$_{+3.24}$	77.91$_{+0.72}$	4.82$_{+0.96}$	13.69$_{+1.80}$	21.45$_{+2.94}$	56.68$_{+2.77}$
		Edge	4.08$_{-5.70}$	12.91$_{-13.24}$	20.00$_{-17.15}$	55.78$_{-21.41}$	2.40$_{-1.46}$	7.81$_{-4.08}$	13.05$_{-5.46}$	43.37$_{-10.54}$
		Depth	2.37$_{-7.41}$	8.44$_{-17.71}$	13.31$_{-23.84}$	46.73$_{-30.46}$	2.31$_{-1.55}$	7.76$_{-4.13}$	12.65$_{-5.86}$	42.57$_{-11.34}$
		Seg	2.49$_{-7.29}$	8.74$_{-17.41}$	14.34$_{-22.81}$	46.52$_{-30.67}$	2.84$_{-1.02}$	9.80$_{-2.09}$	15.58$_{-2.93}$	48.13$_{-5.78}$
Sample4Geo (Deuser et al., 2023)	ConvNeXt-B	RGB	36.77	62.20	71.92	92.06	18.37	38.88	48.98	78.55
		Sketch	22.86$_{-13.91}$	47.61$_{-14.59}$	59.36$_{-12.56}$	89.01$_{-3.05}$	22.43$_{+4.06}$	49.27$_{+10.39}$	61.51$_{+12.53}$	89.80$_{+11.25}$
		Edge	10.61$_{-26.16}$	26.48$_{-35.72}$	35.64$_{-36.28}$	71.33$_{-20.73}$	4.14$_{-14.23}$	11.93$_{-26.95}$	18.20$_{-30.78}$	49.81$_{-28.74}$
		Depth	7.65$_{-29.12}$	20.62$_{-41.58}$	28.41$_{-43.51}$	64.81$_{-27.25}$	3.45$_{-14.92}$	10.73$_{-28.15}$	16.91$_{-32.07}$	47.94$_{-30.61}$
		Seg	7.32$_{-29.45}$	19.60$_{-42.60}$	27.27$_{-44.65}$	62.71$_{-29.35}$	6.67$_{-11.70}$	19.39$_{-19.49}$	27.36$_{-21.62}$	61.30$_{-17.25}$
ConGeo (Mi et al., 2024)	ConvNeXt-B	RGB	49.10	70.80	78.00	93.10	24.60	45.30	54.30	80.60
		Sketch	43.70$_{-5.40}$	69.11$_{-1.69}$	77.57$_{-0.43}$	94.23$_{+1.13}$	33.56$_{+8.96}$	60.13$_{+14.83}$	70.80$_{+16.50}$	90.38$_{+9.78}$
		Edge	17.18$_{-31.92}$	33.17$_{-37.63}$	41.40$_{-36.60}$	71.40$_{-21.70}$	4.23$_{-20.37}$	11.65$_{-33.65}$	16.67$_{-37.63}$	43.93$_{-36.67}$
		Depth	15.51$_{-33.59}$	32.86$_{-37.94}$	41.64$_{-36.36}$	72.71$_{-20.39}$	5.27$_{-19.33}$	14.74$_{-30.56}$	21.02$_{-33.28}$	51.69$_{-28.91}$
		Seg	14.62$_{-34.48}$	29.11$_{-41.69}$	37.16$_{-40.84}$	66.63$_{-26.47}$	6.44$_{-18.16}$	16.67$_{-28.63}$	23.25$_{-31.05}$	53.17$_{-27.43}$
SinGeo (Chen et al., 2026)	ConvNeXt-B	RGB	44.88	66.63	73.31	89.41	22.61	41.78	50.10	74.41
		Sketch	48.46$_{+3.58}$	71.48$_{+4.85}$	79.44$_{+6.13}$	93.37$_{+3.96}$	38.09$_{+15.48}$	63.55$_{+21.77}$	71.55$_{+21.45}$	89.52$_{+15.11}$
		Edge	18.30$_{-26.58}$	34.48$_{-32.15}$	42.96$_{-30.35}$	68.67$_{-20.74}$	5.11$_{-17.50}$	13.82$_{-27.96}$	19.59$_{-30.51}$	45.73$_{-28.68}$
		Depth	18.13$_{-26.75}$	34.52$_{-32.11}$	42.71$_{-30.60}$	70.00$_{-19.41}$	13.57$_{-9.04}$	29.87$_{-11.91}$	38.68$_{-11.42}$	67.51$_{-6.90}$
		Seg	15.29$_{-29.59}$	28.35$_{-38.28}$	35.97$_{-37.34}$	61.90$_{-27.51}$	10.62$_{-11.99}$	24.51$_{-17.27}$	31.73$_{-18.37}$	59.92$_{-14.49}$

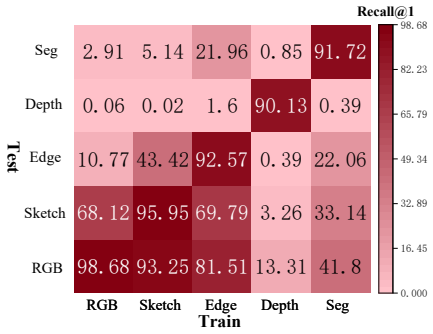
Method-dependent sensitivity to structural representations. Under north-aligned, representation-matched evaluation on CVACT-Cue, TransGeo and GeoDTR achieve similar RGB performance (84.95% and 85.43% R@1, respectively), yet their performance diverges markedly on Edge queries (45.21% versus 76.86%). This gap of 31.65 percentage points contrasts with the much smaller differences on Depth and Seg queries, at 0.29 and 0.54 percentage points, respectively. These results reveal a pronounced interaction between the retrieval method and query representation: comparable RGB recall does not imply a comparable ability to exploit boundary-based structural cues. A plausible explanation is that differences in spatial aggregation and architectural inductive biases affect how the two pipelines integrate fragmented contours into discriminative representations. The comparatively close performance on Depth and Seg further highlights the representation-specific nature of this divergence. Edge queries thus expose a weakness of TransGeo under the evaluated setting that is not apparent from its RGB baseline.

5.2 Evaluation under Unknown Orientation and Restricted FoV

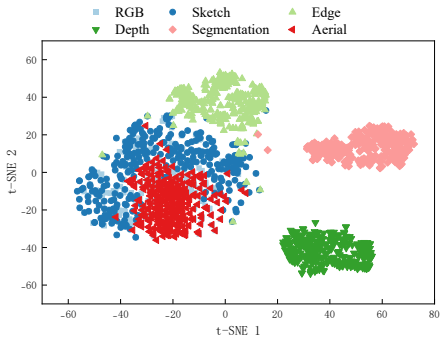
Table 2 reports retrieval performance under representation-matched training ($A \rightarrow A$) where ground queries are constrained by unknown orientation and a restricted $70°$ FoV during both training and evaluation. Visual context truncation causes retrieval accuracy to degrade noticeably across all evaluated architectures and representations compared to the north-aligned panoramic setting. Results for unknown-orientation queries with other FoVs are provided in the Appendix Section A.

Context-dependent ranking reversal between Sketch and RGB. On the urban CVACT-Cue benchmark, Sketch queries outperform RGB baselines across all evaluated methods and metrics, yielding R@1 improvements ranging from 0.96 to 15.48 percentage points. For instance, SinGeo attains 38.09% R@1 with Sketch compared to 22.61% with RGB, while ConGeo advances from 24.60% to 33.56%. On CVUSA-Cue, Sketch improves R@1 for GeoDTR and SinGeo by 1.80 and 3.58 percentage points, respectively. This divergence between the two benchmarks reflects differences in their dominant visual characteristics under visual truncation. In CVACT-Cue (urban scenes), narrow ground perspectives are dominated by transient photometric factors (such as cast shadows and variable facade textures) that correlate poorly with overhead imagery; sketch queries filter these variations to highlight stable architectural silhouettes that align reliably with aerial footprints. In CVUSA-Cue (rural and roadway scenes), distinctive structural contours are sparse, allowing broader color distributions and horizon profiles to retain greater utility for distinguishing locations.

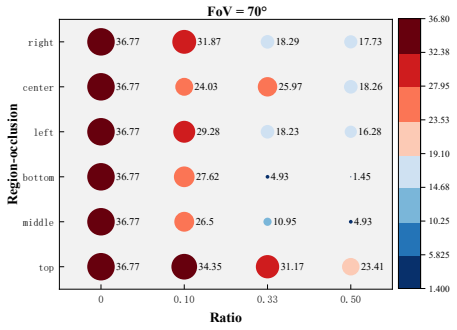
Contextual vulnerability of localized structural representations. In sharp contrast to Sketch, the remaining structural representations (Edge, Depth, and Seg) experience severe degradation under restricted FoVs across all architectures, trailing the RGB baseline by substantial margins. For instance, Sample4Geo achieves R@1 scores of only 7.32%–10.61% on CVUSA-Cue and 3.45%–6.67% on CVACT-Cue. This sharp collapse relative to their robust panoramic performance reveals a



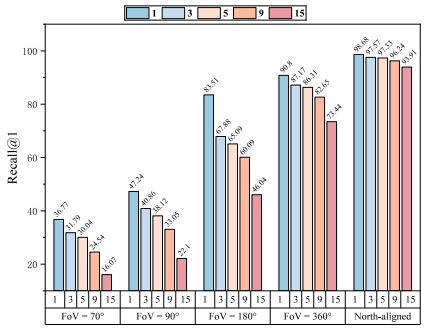
(a) Cross-representation transfer matrix ($A \rightarrow B$)



(b) Latent feature manifold visualization (t-SNE)



(c) Spatial sensitivity under region occlusion



(d) Sensitivity to high-frequency detail across FoVs

Figure 2: **Representation transfer and spatial perturbation evaluations of Sample4Geo on CVUSA-Cue.** (a) Cross-representation transfer matrix for north-aligned panoramas; the horizontal and vertical axes denote the training (A) and test (B) representations, respectively. (b) t-SNE visualization of ground and aerial embeddings across query representations, using the frozen RGB-trained model across 300 randomly sampled pairs. (c) Test-time region occlusion under a 70° FoV across six regions with different masking ratios. (d) Test-time Gaussian blur across varying FoVs and north-aligned panoramas ($k = 1$ indicates the unblurred baseline, and k denotes kernel size).

fundamental reliance on global spatial continuity. While 360° panoramas supply complete horizon transitions, smooth depth gradients, and interconnected region arrangements to resolve perceptual aliasing, a narrow 70° aperture isolates fragmented edge segments and unanchored depth patches that frequently correlate with multiple aerial candidates. Consequently, localized edge boundaries and depth profiles become uninformative in isolation, demonstrating that the utility of these structural cues is strictly conditioned on panoramic visual context.

5.3 Cross-Representation Transfer and Perturbation Analysis

We evaluate cross-representation transfer and test-time spatial perturbations using Sample4Geo on CVUSA-Cue. Figure 2 analyzes cross-representation transfer alongside model sensitivity to region occlusion and Gaussian blur low-pass filtering. Evaluations on patch shuffling, comprehensive metrics, and corresponding CVACT-Cue results are provided in Appendix Section B.

Cross-representation transfer matrix. Figure 2a demonstrates that representation-matched configurations ($A = B$) along the diagonal achieve the highest retrieval accuracy. Conversely, evaluating the RGB-trained model directly on derived structural queries incurs performance drops, yielding 68.12% on Sketch, 10.77% on Edge, 2.91% on Seg, and 0.06% on Depth. Because matched target baselines remain consistently high, this degradation is primarily driven by an unadapted input distribution shift rather than an intrinsic deficiency of localization cues. Furthermore, transfer between RGB and structural representations exhibits directional asymmetry: while RGB $\rightarrow$ Sketch drops to 68.12%, Sketch $\rightarrow$ RGB preserves 93.25% R@1. A similar asymmetry emerges for Edge, where Edge $\rightarrow$ RGB achieves 81.51% compared to 10.77% for RGB $\rightarrow$ Edge. This asymmetry indicates that training on structural representations compels networks to encode foundational geometric layouts that naturally

persist within RGB imagery, whereas training exclusively on RGB imagery allows networks to rely on superficial photometric shortcuts that fail to generalize to derived structural inputs.

Latent feature manifold analysis. Figure 2b presents a t-SNE visualization of query embeddings from the frozen RGB-trained Sample4Geo checkpoint across 300 sampled pairs on CVUSA-Cue. Within the latent space, Sketch and ground RGB queries exhibit extensive overlap proximate to aerial embeddings, demonstrating that the encoder primarily relies on shared structural contours rather than surface textures to achieve alignment (preserving 68.12% R@1 under cross-representation transfer). In contrast, Edge, Seg, and Depth diverge into peripheral clusters, confirming that distinct geometric query representations project into disjoint latent subspaces under an unadapted RGB encoder and causing severe transfer degradation (Edge at 10.77%, Seg at 2.91%, and Depth at 0.06%).

Spatial sensitivity under region occlusion. To diagnose spatial sensitivity under 70° FoV, Figure 2c evaluates directional masking across varying ratios. The results reveal a steep vertical performance gradient: masking the bottom or middle bands incurs severe retrieval degradation, with R@1 dropping to 1.45% and 4.93% at $\rho = 0.50$, respectively, whereas top occlusion retains a substantially higher recall of 23.41%. This disparity indicates that cross-view correspondence relies predominantly on lower and middle visual structures—such as ground-plane road trajectories, roadside terrain boundaries, and horizon profiles—while top-band masking is less damaging under this evaluation setting. In contrast, horizontal occlusions (left, center, right) exhibit uniform degradation (16.28%–18.26% R@1 at $\rho = 0.50$), reflecting the azimuth ambiguity inherent to unaligned orientation queries.

Sensitivity to high-frequency detail across FoVs. Figure 2d evaluates model sensitivity to high-frequency attenuation by applying Gaussian low-pass filtering across varying FoVs. Under north-aligned panoramas, retrieval performance exhibits minimal degradation, with R@1 decreasing by 4.77 percentage points ($98.68\% \rightarrow 93.91\%$) under the largest kernel ($k = 15$). This resilience indicates that when query context is complete, low-frequency spatial arrangements provide sufficient localization evidence despite smoothed surface details. However, as the visual field narrows to 70°, blur sensitivity escalates, with R@1 collapsing from 36.77% to 16.07% ($k = 15$). Truncating spatial context eliminates global geometric anchoring, compelling the method to rely on high-frequency contours to disambiguate locations. This contextual dependency clarifies why region-level structural representations (Depth and Seg), which inherently lack high-frequency edge transitions, suffer degradation under restricted FoVs despite strong performance under panoramas.

6 Conclusion and Discussion

This paper systematically investigates visual cue dependencies in cross-view geo-localization, addressing a problem previously obscured by reliance on standard RGB imagery. By establishing CVGL-CueBench across CVUSA-Cue and CVACT-Cue, we evaluate five complementary query representations under varying FoVs and orientations. Benchmarking demonstrates that derived structural representations preserve sufficient geometric information to support competitive retrieval, suggesting that high recall can be achieved with these queries in the evaluated panoramic settings. Under a restricted 70° FoV, sketch queries outperform RGB baselines in dense urban environments by preserving invariant architectural silhouettes while filtering out transient surface textures. Furthermore, cross-representation transfer evaluations reveal directional asymmetry, while spatial perturbations demonstrate a critical reliance on ground-plane and horizon geometry as well as heightened vulnerability to high-frequency detail loss when visual context is truncated.

Limitations and future scope. CVGL-CueBench assesses downstream localization performance using specific representation-generation pipelines, including off-the-shelf monocular depth estimation, class-agnostic segmentation models, contour extraction algorithms, and multimodal foundation models. Consequently, the empirical findings reflect the properties of these representations coupled with their designated extraction models, prompt formulations, algorithmic thresholds, and color rendering choices, rather than generator-independent properties of the underlying spatial and structural cues. Exploring alternative generation paradigms or incorporating physical sensor measurements constitutes important avenues for future research. Overall, CVGL-CueBench provides empirical insights and diagnostic protocols to facilitate the design of representation-resilient architectures and the deployment of robust, privacy-preserving localization systems in practical cross-view applications.

Author AI Use Statement

Generative AI tools were employed exclusively for language polishing and grammatical refinement to improve the readability of the manuscript. All generated suggestions were critically reviewed, edited, and approved by the authors, who take full responsibility for the entire contents and technical claims of this paper. All conceptual ideation, mathematical formulation, experimental design, empirical validation, and analytical interpretations were conducted entirely by the authors. The automated vision models mentioned in the benchmark construction (such as the sketch generation pipeline detailed in Section 3.2 and Appendix E) serve strictly as offline experimental methodologies rather than authorial assistance.

Ethics Statement

This research utilizes publicly available, standard academic cross-view geo-localization benchmarks (CVUSA and CVACT) strictly in accordance with their respective licensing agreements and academic research protocols. The derived multi-representation benchmark (CVGL-CueBench) operates exclusively on these existing datasets without collecting new real-world imagery or personal identifying information (PII). By evaluating derived structural ground query representations—including depth profiles, semantic masks, edge maps, and structural sketches—this work actively contributes to privacy-preserving visual localization frameworks, demonstrating that reliable geographical matching can be achieved while eliminating the transmission of sensitive photographic surface textures and transient personal details.

Reproducibility Statement

To ensure the full reproducibility of the experimental results and benchmark construction reported in this work: (1) Detailed descriptions of the problem formulation, coordinate transformations, and restricted FoV cropping protocols are provided in Section 3.1 and Section 3.4; (2) Exact generation hyperparameters, prompt configurations, inference steps, algorithmic thresholds, and deterministic random seeds for the derived query representations (Canny edge detection, SAM, Depth Anything V2, and Qwen-Image-Edit-2511) are documented in Section 3.2 and Appendix E; (3) Complete training regimes and spatial perturbation protocols (covering directional region occlusion, Gaussian low-pass filtering, and patch shuffling) are delineated in Section 4.2, Section 3.3, and Appendix B; (4) Code, pre-trained model weights, and the CVGL-CueBench dataset generation scripts will be made publicly available upon publication.

References

Mayank Bansal, Harpreet S Sawhney, Hui Cheng, and Kostas Daniilidis. Geo-localization of street views with aerial image databases. In *Proceedings of the 19th ACM international conference on Multimedia*, pp. 1125–1128, 2011.

John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, 1986. doi: 10.1109/TPAMI.1986.4767851.

Yang Chen, Xieyuanli Chen, Junxiang Li, Jie Tang, and Tao Wu. Singeo: Unlock single model's potential for robust cross-view geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19403–19412, June 2026.

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.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

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

Wenmiao Hu, Yichen Zhang, Yuxuan Liang, Yifang Yin, Andrei Georgescu, An Tran, Hannes Kruppa, See-Kiong Ng, and Roger Zimmermann. Beyond geo-localization: Fine-grained orientation of street-view images by cross-view matching with satellite imagery. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, pp. 6155–6164, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392037. doi: 10.1145/3503161.3548102. URL https://doi.org/10.1145/3503161.3548102.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17853–17862. IEEE, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Jiangshan Li, Chunfang Yang, Baojun Qi, Ma Zhu, and Nan Wu. 4scig: A four-branch framework to reduce the interference of sky area in cross-view image geo-localization. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–18, 2024. doi: 10.1109/TGRS.2024.3379376.

Yiru Li and Yingying Zhu. Plgeo: A patch-level framework to overcome orientation discrepancies in cross-view geo-localization. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 6057–6065, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3755181. URL https://doi.org/10.1145/3746027.3755181.

Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *CVPR*, pp. 5624–5633, 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.

Wajeeha Nasar, Ricardo Da Silva Torres, Odd Erik Gundersen, and Anniken T Karlsen. The use of decision support in search and rescue: A systematic literature review. *ISPRS International Journal of Geo-Information*, 12(5):182, 2023.

Royston Rodrigues and Masahiro Tani. Are these from the same place? seeing the unseen in cross-view image geo-localization. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 3752–3760. IEEE, 2021.

Markus Ryll, John Ware, John Carter, and Nick Roy. Semantic trajectory planning for long-distant unmanned aerial vehicle navigation in urban environments. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1551–1558. IEEE, 2020.

Ji Shen, Lixing Chen, Yang Bai, Zhongqi Miao, Zhe Qu, Pan Zhou, and Jianhua Li. Augmenting cross-view geo-localization with spatial semantics from vision foundation models. In *Proceedings of the ACM Web Conference 2026*, WWW '26, pp. 5177–5188, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400723070. doi: 10.1145/3774904.3792233. URL https://doi.org/10.1145/3774904.3792233.

Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/ba2f0015122a5955f8b3a50240fb91b2-Paper.pdf.

Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. Where am i looking at? joint location and orientation estimation by cross-view matching. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4063–4071. IEEE, 2020.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Le Wu, Lv Bo, Songsong Ouyang, and Yingying Zhu. Mrgeo: Robust cross-view geo-localization of corrupted images via spatial and channel feature enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10691–10699, 2026.

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.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in neural information processing systems*, 37:21875–21911, 2024.

Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. Predicting ground-level scene layout from aerial imagery. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4132–4140. IEEE, 2017.

Qingwang Zhang and Yingying Zhu. Benchmarking the robustness of cross-view geo-localization models. In *European Conference on Computer Vision*, pp. 36–53. Springer, 2024.

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.

Sijie Zhu, Mubarak Shah, and Chen Chen. Transgeo: Transformer is all you need for cross-view image geo-localization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1152–1161. IEEE, 2022.

OVERVIEW

In this appendix, we provide the following supplementary materials and detailed analyses to complement the main paper:

- **Section A: Supplementary Results under Varying Fields of View.**
- **Section B: Supplementary Results on Cross-Representation Transfer and Spatial Perturbations.**
- **Section C: Qualitative Examples of CVGL-CueBench.**
- **Section D: Rationale for Fixed RGB Aerial Imagery.**
- **Section E: Derived Representation Generation Pipelines and Analysis.**

A SUPPLEMENTARY RESULTS UNDER VARYING FIELDS OF VIEW

To complement the north-aligned panorama and restricted 70° FoV evaluations presented in the main paper, Tables 3, 4, and 5 document retrieval performance under representation-matched training ($A \rightarrow A$) across intermediate and extended fields of view ($90^\circ, 180^\circ$, and 360°) under unknown orientation. Across evaluated architectures and representations, retrieval recall (R@1) scales monotonically as the visual field widens, reflecting the progressive accumulation of spatial evidence. Crucially, these multi-view benchmarks reveal how the relative utility of derived structural representations evolves along the continuum from localized perspective crops to complete panoramic coverage.

Sketch query resilience across extended viewing angles. On the urban CVACT-Cue benchmark, Sketch queries systematically outperform standard RGB baselines across all five evaluated methods at a 90° FoV (Table 3), yielding R@1 improvements ranging from 1.51 percentage points with GeoDTR to 11.89 percentage points with SinGeo. As the field of view expands to 180° (Table 4) and full 360° panoramas under unknown orientation (Table 5), this performance advantage remains consistent across all evaluated architectures, securing R@1 gains of 0.22–12.31 percentage points at 180° and 0.87–6.71 percentage points at 360°. These sustained improvements confirm that the efficacy of sketch representations is not an artifact confined to narrow viewing crops (70°). In dense urban scenes characterized by repetitive facade geometries, volatile cast shadows, and transient foreground objects, continuous structural outlines extracted by multimodal foundation models provide geometric constraints that align with overhead footprints across diverse viewing angles. Conversely, on the rural CVUSA-Cue benchmark—where scenes are dominated by open roadway corridors and natural horizon profiles—Sketch performance relative to RGB varies substantially across methods and viewing angles.

Field of view dependency of Edge, Depth, and Segmentation. Unlike sketch queries that perform reliably across different view angles, the remaining derived representations (Edge, Depth, and Seg) depend heavily on the available FoV. Under a restricted 90° FoV, these representations fall substantially behind the RGB baseline, with R@1 decreasing by more than 20–30 percentage points across multiple models. This drop occurs because narrow perspective crops provide limited spatial context: without broader surroundings, localized depth gradients or flat segmentation regions often look similar across different locations, creating visual ambiguity. Expanding the field of view to 180° noticeably improves retrieval capability, with ConGeo reaching 57.06% R@1 on CVUSA-Cue and 54.07% on CVACT-Cue using depth queries. Under full 360° panoramas with unknown orientation, the performance gap between these derived representations and the RGB baseline narrows substantially. Specifically, on CVACT-Cue, Seg achieves 81.70% R@1 with ConGeo and 83.21% with SinGeo, trailing RGB by only 1.30 and 0.20 percentage points, respectively. These results indicate that the poor performance of edge, depth, and segmentation queries under narrow views (70°–90°) comes from visual ambiguity in small views rather than a failure of the representations themselves. When sufficient surrounding layout information is available, structural representations provide reliable evidence for cross-view localization.

Impact of unknown orientation on full panoramas. Comparing the 360° results under unknown orientation (Table 5) with the north-aligned setting in Table 1 shows how orientation alignment affects retrieval. When camera heading is unknown, the model must search across all horizontal directions rather than matching along a fixed heading, which generally reduces R@1, with several

Table 3: **Evaluation results under representation-matched training ($A \rightarrow A$) with unknown orientation and 90° FoV.** In this benchmark, ground-level query images are restricted to an unknown azimuth orientation and a limited 90° FoV during both training and evaluation phases, while the corresponding aerial satellite references maintain standard full-view RGB imagery. Performance is measured by Recall@k (R@1, R@5, R@10) and R@1%. The subscript values indicate the performance difference relative to the standard RGB baseline of the respective method, where green denotes positive performance gains and red denotes performance degradation.

Method	Backbone	Query	CVUSA-Cue				CVACT-Cue			
			R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
TransGeo Zhu et al. (2022)	DeiT	RGB	30.56	55.06	65.44	89.52	9.50	24.35	33.12	68.16
		Sketch	27.30$_{-3.26}$	53.62$_{-1.44}$	64.03$_{-1.41}$	92.08$_{+2.56}$	12.55$_{+3.05}$	32.49$_{+8.14}$	44.22$_{+11.10}$	80.30$_{+12.14}$
		Edge	7.85$_{-22.71}$	20.99$_{-34.07}$	28.93$_{-36.51}$	63.40$_{-26.12}$	2.04$_{-7.46}$	7.86$_{-16.49}$	11.41$_{-21.71}$	40.20$_{-27.96}$
		Depth	7.50$_{-23.06}$	20.47$_{-34.59}$	28.31$_{-37.13}$	62.85$_{-26.67}$	2.43$_{-7.07}$	8.20$_{-16.15}$	12.94$_{-20.18}$	41.94$_{-26.22}$
		Seg	7.32$_{-23.24}$	19.61$_{-35.45}$	27.46$_{-37.98}$	61.61$_{-27.91}$	4.73$_{-4.77}$	14.71$_{-9.64}$	21.03$_{-12.09}$	54.02$_{-14.14}$
GeoDTR Zhang et al. (2023)	ResNet-34	RGB	13.15	32.40	43.74	81.23	5.41	15.49	22.68	59.84
		Sketch	14.86$_{+1.71}$	35.50$_{+3.10}$	47.27$_{+3.53}$	83.14$_{+1.91}$	6.92$_{+1.51}$	19.63$_{+4.14}$	28.36$_{+5.68}$	65.78$_{+5.94}$
		Edge	5.90$_{-7.25}$	17.57$_{-14.83}$	25.61$_{-18.13}$	63.64$_{-17.59}$	3.59$_{-1.82}$	11.13$_{-4.36}$	17.30$_{-5.38}$	50.43$_{-9.41}$
		Depth	3.37$_{-9.78}$	11.27$_{-21.13}$	17.76$_{-25.98}$	53.50$_{-27.73}$	2.82$_{-2.59}$	9.51$_{-5.98}$	15.54$_{-7.14}$	47.34$_{-12.50}$
		Seg	3.74$_{-9.41}$	12.10$_{-20.30}$	18.65$_{-25.09}$	53.84$_{-27.39}$	3.66$_{-1.75}$	12.53$_{-2.96}$	19.14$_{-3.54}$	53.46$_{-6.38}$
Sample4Geo Deuser et al. (2023)	ConvNeXt-B	RGB	47.24	71.24	79.14	93.96	26.70	50.69	60.96	85.05
		Sketch	29.87$_{-17.37}$	56.01$_{-15.23}$	66.91$_{-12.23}$	92.13$_{-1.83}$	31.46$_{+4.76}$	60.98$_{+10.29}$	72.07$_{+11.11}$	92.85$_{+7.80}$
		Edge	16.62$_{-30.62}$	36.77$_{-34.47}$	47.51$_{-31.63}$	80.27$_{-13.69}$	6.78$_{-19.92}$	19.30$_{-31.39}$	27.30$_{-33.66}$	61.40$_{-23.65}$
		Depth	10.94$_{-36.30}$	27.40$_{-43.84}$	37.55$_{-41.59}$	73.12$_{-20.84}$	6.95$_{-19.75}$	19.36$_{-31.33}$	27.48$_{-33.48}$	62.29$_{-22.76}$
		Seg	12.66$_{-34.58}$	30.11$_{-41.13}$	39.38$_{-39.76}$	74.23$_{-19.73}$	11.04$_{-15.66}$	27.81$_{-22.88}$	37.42$_{-23.54}$	70.38$_{-14.67}$
ConGeo Mi et al. (2024)	ConvNeXt-B	RGB	55.50	75.40	81.50	93.90	40.60	62.60	69.80	86.60
		Sketch	57.59$_{+2.09}$	79.35$_{+3.95}$	88.95$_{+7.45}$	96.46$_{+2.56}$	43.38$_{+2.78}$	69.21$_{+6.61}$	76.71$_{+6.91}$	92.19$_{+5.59}$
		Edge	25.87$_{-29.63}$	45.44$_{-29.96}$	54.53$_{-26.97}$	79.04$_{-14.86}$	7.77$_{-32.83}$	18.92$_{-43.68}$	25.84$_{-43.96}$	53.87$_{-32.73}$
		Depth	23.58$_{-31.92}$	42.82$_{-32.58}$	52.07$_{-29.43}$	79.23$_{-14.67}$	13.09$_{-27.51}$	29.80$_{-32.80}$	38.64$_{-31.16}$	68.70$_{-17.90}$
		Seg	22.87$_{-32.63}$	40.60$_{-34.80}$	48.85$_{-32.65}$	74.35$_{-19.55}$	12.07$_{-28.53}$	26.63$_{-35.97}$	34.43$_{-35.37}$	62.78$_{-23.82}$
SinGeo Chen et al. (2026)	ConvNeXt-B	RGB	54.33	75.01	81.20	93.10	38.58	58.93	66.24	84.44
		Sketch	60.11$_{+5.78}$	80.11$_{+5.10}$	86.22$_{+5.02}$	95.85$_{+2.75}$	50.47$_{+11.89}$	72.69$_{+13.76}$	78.95$_{+12.71}$	91.51$_{+7.07}$
		Edge	25.09$_{-29.24}$	43.25$_{-31.76}$	51.94$_{-29.26}$	75.34$_{-17.76}$	7.76$_{-30.82}$	17.86$_{-41.07}$	24.57$_{-41.67}$	50.58$_{-33.86}$
		Depth	24.39$_{-29.94}$	43.41$_{-31.60}$	52.17$_{-29.03}$	77.09$_{-16.01}$	27.98$_{-10.60}$	49.00$_{-9.93}$	57.41$_{-8.83}$	80.22$_{-4.22}$
		Seg	22.48$_{-31.85}$	38.55$_{-36.46}$	45.94$_{-35.26}$	69.65$_{-23.45}$	20.32$_{-18.26}$	38.72$_{-20.21}$	46.98$_{-19.26}$	70.98$_{-13.46}$

Table 4: **Evaluation results under representation-matched training ($A \rightarrow A$) with unknown orientation and 180° FoV.** Ground queries are evaluated under matched unknown orientation and 180° FoV on CVUSA-Cue and CVACT-Cue against invariant RGB aerial galleries.

Method	Backbone	Query	CVUSA-Cue				CVACT-Cue			
			R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
TransGeo Zhu et al. (2022)	DeiT	RGB	58.62	81.86	88.19	98.13	36.01	62.13	71.42	91.52
		Sketch	61.14$_{+2.52}$	85.65$_{+3.79}$	91.34$_{+3.15}$	99.07$_{+0.94}$	48.32$_{+12.31}$	75.82$_{+13.69}$	83.83$_{+12.41}$	96.38$_{+4.86}$
		Edge	24.52$_{-34.10}$	47.54$_{-34.32}$	58.19$_{-30.00}$	88.05$_{-10.08}$	14.90$_{-21.11}$	34.36$_{-27.77}$	44.65$_{-26.77}$	77.79$_{-13.73}$
		Depth	35.87$_{-22.75}$	60.24$_{-21.62}$	70.21$_{-17.98}$	93.69$_{-4.44}$	17.15$_{-18.86}$	38.52$_{-23.61}$	49.11$_{-22.31}$	80.63$_{-10.89}$
		Seg	34.23$_{-24.39}$	59.04$_{-22.82}$	69.00$_{-19.19}$	92.23$_{-5.90}$	22.02$_{-13.99}$	45.03$_{-17.10}$	55.72$_{-15.70}$	84.26$_{-7.26}$
GeoDTR Zhang et al. (2023)	ResNet-34	RGB	30.29	58.16	70.40	94.65	17.99	39.81	51.64	83.80
		Sketch	37.92$_{+7.63}$	67.27$_{+9.11}$	77.93$_{+7.53}$	96.74$_{+2.09}$	21.12$_{+3.13}$	45.74$_{+5.93}$	57.37$_{+5.73}$	87.98$_{+4.18}$
		Edge	15.67$_{-14.62}$	37.94$_{-20.22}$	50.01$_{-20.39}$	85.95$_{-8.70}$	11.26$_{-6.73}$	27.58$_{-12.23}$	37.39$_{-14.25}$	73.45$_{-10.35}$
		Depth	11.56$_{-18.73}$	29.68$_{-28.48}$	41.77$_{-28.63}$	80.45$_{-14.20}$	10.61$_{-7.38}$	26.50$_{-13.31}$	36.21$_{-15.43}$	73.09$_{-10.71}$
		Seg	15.37$_{-14.92}$	37.02$_{-21.14}$	48.61$_{-21.79}$	84.56$_{-10.09}$	11.38$_{-6.61}$	29.05$_{-10.76}$	39.47$_{-12.17}$	76.08$_{-7.72}$
Sample4Geo Deuser et al. (2023)	ConvNeXt-B	RGB	83.51	94.54	96.23	98.83	58.36	77.25	83.01	93.73
		Sketch	60.02$_{-23.49}$	84.57$_{-9.97}$	90.20$_{-6.03}$	98.79$_{-0.04}$	61.14$_{+2.78}$	84.36$_{+7.11}$	89.29$_{+6.28}$	96.65$_{+2.92}$
		Edge	42.26$_{-41.25}$	68.33$_{-26.21}$	77.26$_{-18.97}$	94.77$_{-4.06}$	30.63$_{-27.73}$	54.87$_{-22.38}$	63.59$_{-19.42}$	87.02$_{-6.71}$
		Depth	35.27$_{-48.24}$	61.00$_{-33.54}$	70.73$_{-25.50}$	92.88$_{-5.95}$	39.53$_{-18.83}$	64.65$_{-12.60}$	73.40$_{-9.61}$	91.43$_{-2.30}$
		Seg	37.96$_{-45.55}$	63.89$_{-30.65}$	73.66$_{-22.57}$	93.66$_{-5.17}$	37.51$_{-20.85}$	62.10$_{-15.15}$	70.36$_{-12.65}$	89.64$_{-4.09}$
ConGeo Mi et al. (2024)	ConvNeXt-B	RGB	92.30	97.90	98.70	99.70	70.30	85.20	88.60	95.10
		Sketch	87.16$_{-5.14}$	96.01$_{-1.89}$	97.42$_{-1.28}$	99.35$_{-0.35}$	70.52$_{+0.22}$	85.43$_{+0.23}$	88.82$_{+0.22}$	96.01$_{+0.91}$
		Edge	59.23$_{-33.07}$	78.36$_{-19.54}$	84.06$_{-14.64}$	94.37$_{-5.33}$	31.62$_{-38.68}$	51.78$_{-33.42}$	59.98$_{-28.62}$	81.82$_{-13.28}$
		Depth	57.06$_{-35.24}$	76.91$_{-20.99}$	82.71$_{-15.99}$	94.40$_{-5.30}$	54.07$_{-16.23}$	73.26$_{-11.94}$	79.27$_{-9.33}$	91.90$_{-3.20}$
		Seg	55.65$_{-36.65}$	75.90$_{-22.00}$	81.97$_{-16.73}$	94.39$_{-5.31}$	44.90$_{-25.40}$	65.91$_{-19.29}$	72.95$_{-15.65}$	88.54$_{-6.56}$
SinGeo Chen et al. (2026)	ConvNeXt-B	RGB	89.82	97.11	98.17	99.48	68.16	82.04	85.82	93.39
		Sketch	86.56$_{-3.26}$	95.70$_{-1.41}$	97.16$_{-1.01}$	99.20$_{-0.28}$	74.12$_{+5.96}$	87.15$_{+5.11}$	90.02$_{+4.20}$	95.61$_{+2.22}$
		Edge	54.31$_{-35.51}$	74.59$_{-22.52}$	81.32$_{-16.85}$	92.59$_{-6.89}$	31.88$_{-36.28}$	50.29$_{-31.75}$	58.31$_{-27.51}$	78.42$_{-14.97}$
		Depth	62.10$_{-27.72}$	79.77$_{-17.34}$	84.37$_{-13.80}$	93.85$_{-5.63}$	64.38$_{-3.78}$	80.20$_{-1.84}$	84.96$_{-0.86}$	93.51$_{+0.12}$
		Seg	55.72$_{-34.10}$	75.55$_{-21.56}$	81.39$_{-16.78}$	93.06$_{-6.42}$	60.34$_{-7.82}$	76.27$_{-5.77}$	81.13$_{-4.69}$	91.64$_{-1.75}$

exceptions in the SinGeo results. For example, on CVUSA-Cue, the R@1 of Sample4Geo on RGB drops from 98.68% to 90.80%, and derived representations show similar decreases. Nevertheless, models such as ConGeo and SinGeo still maintain high retrieval accuracy ($> 94\%$ R@1 on both RGB and Sketch on CVUSA-Cue). This result demonstrates that structural representations remain effective for geo-localization even without orientation calibration.

Table 5: **Evaluation results under representation-matched training ($A \to A$) with unknown orientation and 360° FoV.** Ground queries are evaluated under matched unknown orientation and 360° FoV on CVUSA-Cue and CVACT-Cue against invariant RGB aerial galleries.

Method	Backbone	Query	CVUSA-Cue				CVACT-Cue			
			R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
TransGeo Zhu et al. (2022)	DeiT	RGB	78.75	92.36	95.42	99.30	67.04	84.62	88.89	96.45
		Sketch	86.22$_{+7.47}$	95.86$_{+3.50}$	97.54$_{+2.12}$	99.68$_{+0.38}$	73.75$_{+6.71}$	89.18$_{+4.56}$	92.37$_{+3.48}$	97.78$_{+1.33}$
		Edge	47.32$_{-31.43}$	71.46$_{-20.90}$	79.18$_{-16.24}$	95.62$_{-3.68}$	36.63$_{-30.41}$	60.92$_{-23.70}$	70.26$_{-18.63}$	90.15$_{-6.30}$
		Depth	75.24$_{-3.51}$	90.99$_{-1.37}$	94.61$_{-0.81}$	99.01$_{-0.29}$	49.71$_{-17.33}$	72.92$_{-11.70}$	80.05$_{-8.84}$	93.95$_{-2.50}$
		Seg	73.22$_{-5.53}$	90.26$_{-2.10}$	93.84$_{-1.58}$	98.91$_{-0.39}$	55.98$_{-11.06}$	78.96$_{-5.66}$	84.21$_{-4.68}$	95.03$_{-1.42}$
GeoDTR Zhang et al. (2023)	ResNet-34	RGB	49.66	75.97	84.20	97.74	43.22	68.51	77.12	93.80
		Sketch	46.57$_{-3.09}$	73.81$_{-2.16}$	82.44$_{-1.76}$	97.30$_{-0.44}$	46.18$_{+2.96}$	73.03$_{+4.52}$	80.54$_{+3.42}$	95.46$_{+1.66}$
		Edge	29.80$_{-19.86}$	55.80$_{-20.17}$	67.12$_{-17.08}$	93.48$_{-4.26}$	29.40$_{-13.82}$	54.25$_{-14.26}$	64.13$_{-12.99}$	89.89$_{-3.91}$
		Depth	42.99$_{-6.67}$	70.19$_{-5.78}$	80.63$_{-3.57}$	97.28$_{-0.46}$	34.53$_{-8.69}$	60.41$_{-8.10}$	70.45$_{-6.67}$	92.13$_{-1.67}$
		Seg	51.51$_{+1.85}$	79.15$_{+3.18}$	87.56$_{+3.36}$	98.28$_{+0.54}$	42.26$_{-0.96}$	68.86$_{+0.35}$	77.17$_{+0.05}$	93.85$_{+0.05}$
Sample4Geo Deuser et al. (2023)	ConvNeXt-B	RGB	90.80	97.14	98.05	99.10	82.40	90.60	92.30	95.40
		Sketch	82.75$_{-8.05}$	94.00$_{-3.14}$	96.04$_{-2.01}$	98.45$_{-0.65}$	83.27$_{+0.87}$	92.01$_{+1.41}$	93.59$_{+1.29}$	96.93$_{+1.53}$
		Edge	66.82$_{-23.98}$	84.06$_{-13.08}$	88.54$_{-9.51}$	94.74$_{-4.36}$	66.98$_{-15.42}$	80.65$_{-9.95}$	84.07$_{-8.23}$	91.98$_{-3.42}$
		Depth	68.89$_{-21.91}$	85.65$_{-11.49}$	89.67$_{-8.38}$	95.95$_{-3.15}$	78.56$_{-3.84}$	88.36$_{-2.24}$	90.54$_{-1.76}$	95.23$_{-0.17}$
		Seg	73.63$_{-17.17}$	89.93$_{-7.21}$	92.97$_{-5.08}$	97.12$_{-1.98}$	75.54$_{-6.86}$	86.81$_{-3.79}$	89.23$_{-3.07}$	94.20$_{-1.20}$
ConGeo Mi et al. (2024)	ConvNeXt-B	RGB	96.60	98.90	99.20	99.70	83.00	90.60	92.40	96.30
		Sketch	94.85$_{-1.75}$	98.51$_{-0.39}$	99.00$_{-0.20}$	99.66$_{-0.04}$	84.95$_{+1.95}$	92.10$_{+1.50}$	93.73$_{+1.33}$	97.00$_{+0.70}$
		Edge	79.44$_{-17.16}$	91.19$_{-7.71}$	93.55$_{-5.65}$	97.70$_{-2.00}$	63.23$_{-19.77}$	77.81$_{-12.79}$	81.85$_{-10.55}$	91.34$_{-4.96}$
		Depth	89.65$_{-6.95}$	95.38$_{-3.52}$	96.56$_{-2.64}$	98.82$_{-0.88}$	80.99$_{-2.01}$	89.89$_{-0.71}$	91.86$_{-0.54}$	96.62$_{+0.32}$
		Seg	88.33$_{-8.27}$	95.58$_{-3.32}$	96.89$_{-2.31}$	98.93$_{-0.77}$	81.70$_{-1.30}$	90.54$_{-0.06}$	92.39$_{-0.01}$	96.81$_{+0.51}$
SinGeo Chen et al. (2026)	ConvNeXt-B	RGB	97.03	99.18	99.44	99.72	83.41	91.06	92.01	95.07
		Sketch	94.72$_{-2.31}$	98.31$_{-0.87}$	98.78$_{-0.66}$	99.56$_{-0.16}$	84.89$_{+1.48}$	92.42$_{+1.36}$	93.98$_{+1.97}$	96.98$_{+1.91}$
		Edge	74.88$_{-22.15}$	88.06$_{-11.12}$	91.19$_{-8.25}$	96.39$_{-3.33}$	66.31$_{-17.10}$	79.57$_{-11.49}$	83.36$_{-8.65}$	91.60$_{-3.47}$
		Depth	89.00$_{-8.03}$	95.04$_{-4.14}$	96.30$_{-3.14}$	98.40$_{-1.32}$	81.85$_{-1.56}$	90.35$_{-0.71}$	92.32$_{+0.31}$	96.36$_{+1.29}$
		Seg	86.80$_{-10.23}$	95.01$_{-4.17}$	96.58$_{-2.86}$	98.64$_{-1.08}$	83.21$_{-0.20}$	90.87$_{-0.19}$	92.57$_{+0.56}$	96.42$_{+1.35}$

B Supplementary Results on Cross-Representation Transfer and Spatial Perturbations

To complement the analysis in Section 5.3, this section provides extended evaluations of cross-representation transfer ($A \to B$) and diagnostic spatial perturbations using Sample4Geo. Specifically, we report: (1) multi-metric transfer evaluations (R@5, R@10, and R@1%) on CVUSA-Cue alongside complete transfer matrices on CVACT-Cue; (2) spatial sensitivity under directional region occlusion across extended viewing geometries (90°, 180°, and 360° under unknown orientation); (3) multi-metric evaluations of high-frequency attenuation via Gaussian low-pass filtering across varying fields of view; and (4) sensitivity analyses under spatial patch shuffling across multiple grid partitions. All experimental settings follow the protocols defined in Section 5.3.

Extended cross-representation transfer across metrics and datasets. Figure 3 presents cross-representation transfer matrices ($A \to B$) evaluated under broader candidate ranks across CVUSA-Cue and CVACT-Cue. When models are trained and tested on the identical representation ($A = B$), retrieval accuracy remains high, reaching near-saturation ($\geqslant 99.50\%$ at Recall@1% on CVUSA-Cue and $\geqslant 97.61\%$ on CVACT-Cue) across all five query representations. When evaluating off-diagonal transfer from the unadapted RGB checkpoint, expanding the evaluation window from Recall@1 to Recall@1% shows notable performance recovery. On CVACT-Cue, evaluating the RGB checkpoint on Sketch queries yields 31.97% at Recall@1, but recovers to 52.29% at Recall@5, 59.63% at Recall@10, and 80.84% at Recall@1%. A similar recovery trend appears for Edge queries, where retrieval increases from 4.25% at Recall@1 to 31.08% at Recall@1%. This trend indicates that while an unadapted RGB model struggles to identify the exact top-1 match without natural surface colors, its learned geometric representations still successfully place the correct reference within the top candidate pool. Furthermore, the directional transfer asymmetry observed in the main paper persists across both datasets and across all recall metrics. Models trained on structural line drawings transfer effectively to RGB imagery, achieving 77.19% Recall@1 and 96.58% Recall@1% for Sketch $\to$ RGB on CVACT-Cue, compared to 31.97% Recall@1 and 80.84% Recall@1% for RGB $\to$ Sketch. A similar pattern appears between Edge and RGB. This directional difference occurs because training on structural contours encourages the network to focus on visual outlines that also exist in real photographs. In contrast, standard RGB training allows models to rely on surface colors and textures, which are largely removed in derived line queries. In contrast, depth queries transferred from the RGB checkpoint remain near zero across all recall thresholds (0.01%–1.13% on CVACT-Cue and 0.06%–2.94% on CVUSA-Cue), confirming that smooth distance gradient maps differ fundamentally from the feature space learned from RGB imagery.

Sensitivity to directional region occlusion across viewing angles. Figures 4(a)–(c) evaluate the retrieval accuracy of Sample4Geo under directional region masking across $90°, 180°$, and $360°$ fields of view under unknown orientation on CVUSA-Cue. Across all tested fields of view, the results confirm a consistent vertical performance difference. Masking the bottom region produces the largest drop in retrieval accuracy, with Recall@1 decreasing from unmasked baselines ($47.24\%, 83.51\%$, and 90.80%) to 1.55% at $90°$, 1.07% at $180°$, and 1.38% at $360°$ under half-image occlusion ($\rho = 0.50$). Masking the middle region shows a similar downward trend, dropping to 6.96% at $90°$, 12.98% at $180°$, and 11.01% at $360°$. In the open rural environments of CVUSA-Cue, the bottom band contains road trajectories and nearby ground boundaries, while the middle band contains the horizon line. These visual elements provide the main reference points for matching against overhead satellite views. In contrast, top occlusion causes much smaller performance drops, retaining 30.49% Recall@1 at $90°$, 60.02% at $180°$, and 73.78% at $360°$ at $\rho = 0.50$. This stability confirms that upper sky regions contain little useful information for localization. Across horizontal partitions (left, center, and right), retrieval performance degrades at similar rates, reflecting the fact that informative visual landmarks are equally likely to appear at any horizontal position under unknown camera orientations.

Sensitivity to high-frequency blur across fields of view and metrics. Figures 4(d)–(f) illustrate retrieval accuracy under Gaussian low-pass filtering across multiple recall metrics (Recall@5, Recall@10, and Recall@1%) and five viewing setups on CVUSA-Cue. The effect of blur depends directly on the available field of view. Under north-aligned panoramas, retrieval performance remains stable even under the largest blur kernel ($k = 15$), with Recall@5 and Recall@10 staying above 98% and 99%, respectively. When complete panoramic context is available, the broad layout of the scene provides sufficient information to identify the correct geographic location despite smoothed surface details. However, as the visual field narrows to $90°$ and $70°$, sensitivity to blur increases substantially across all recall metrics. Under a $70°$ field of view, increasing the filter kernel from $k = 1$ to $k = 15$ reduces Recall@5 from 62.20% to 35.15%, Recall@10 from 71.92% to 45.70%, and Recall@1% from 92.06% to 78.41%. When broad visual context is missing, the model must rely on fine visual details and local boundaries to distinguish candidate locations, making narrow views highly vulnerable to high-frequency detail loss.

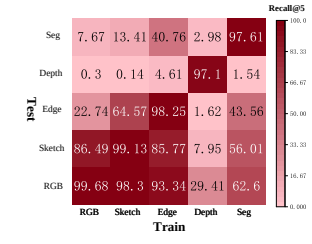
Impact of spatial patch shuffling on global layout. Figures 4(g)–(j) analyze model sensitivity to spatial layout disruption by testing patch shuffling across four grid partitions ($1\times1, 2\times2, 4\times4, 8\times8$) and five viewing conditions on CVUSA-Cue. Across all evaluated criteria, breaking the spatial structure leads to a continuous drop in retrieval performance as the grid becomes finer. While a coarse 2×2 partition retains partial retrieval capability under wide viewing angles, dividing images into 4×4 and 8×8 grids causes Recall@1 to fall close to zero across all viewing geometries. Although evaluating with broader windows like Recall@1% expands the recall boundaries in the radar plots, the performance ranking ($1 \times 1 > 2 \times 2 > 4 \times 4 > 8 \times 8$) remains identical across all viewing conditions. Because patch shuffling preserves local textures within each patch while scrambling their spatial positions, this steep drop confirms that cross-view geo-localization models rely on coherent two-dimensional spatial layouts rather than simply matching isolated visual textures.

C Qualitative Examples of CVGL-CueBench

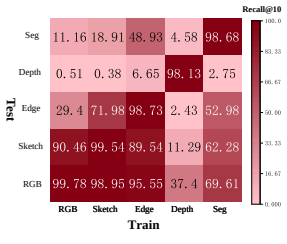
Figure 5 illustrates the visual query representations and perturbation protocols using representative ground scenes from CVUSA-Cue and CVACT-Cue. Within each dataset column, all panels correspond to the identical physical scene. The visual queries comprise the five benchmark representations: RGB, Sketch, Edge, Depth, and Seg. Spatial perturbations include Gaussian low-pass filtering across kernel sizes $k \in \{3, 5, 9, 15\}$ (corresponding to filter sizes 3×3, 5×5, 9×9, and 15×15) alongside spatial patch shuffling across 2×2, 4×4, and 8×8 regular grids. Directional region occlusions cover six spatial bands (top, middle, bottom, left, center, and right) under masking ratios $\rho \in \{0.10, 0.33, 0.50\}$, defined as the fraction of the query area occluded. Masked pixels are filled with the channel-wise means computed from the visible image prior to occlusion.

D Rationale for Fixed RGB Aerial Imagery

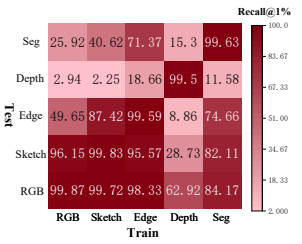
Throughout CVGL-CueBench, the aerial reference gallery remains RGB imagery across all training and evaluation configurations. This asymmetric protocol is motivated by practical operational



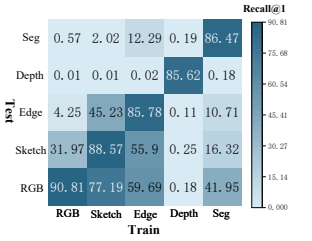
(a) CVUSA-Cue: Recall@5



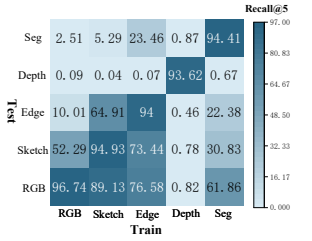
(b) CVUSA-Cue: Recall@10



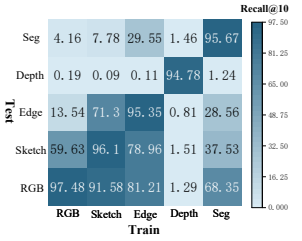
(c) CVUSA-Cue: Recall@1%



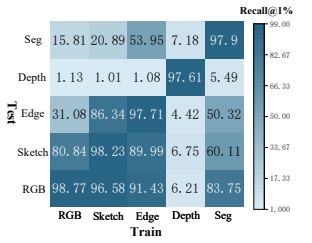
(d) CVACT-Cue: Recall@1



(e) CVACT-Cue: Recall@5



(f) CVACT-Cue: Recall@10



(g) CVACT-Cue: Recall@1%

Figure 3: **Cross-representation transfer matrices ($A \rightarrow B$) of Sample4Geo on CVUSA-Cue and CVACT-Cue.** Evaluations are conducted on north-aligned full panoramas without test-time fine-tuning. The horizontal and vertical axes denote the training (A) and test (B) query representations, respectively. (a)–(c) Extended evaluations on CVUSA-Cue reporting Recall@5, Recall@10, and Recall@1%. (d)–(g) Comprehensive evaluations on CVACT-Cue across all standard retrieval criteria: Recall@1, Recall@5, Recall@10, and Recall@1%.

constraints in large-scale visual geo-localization systems. Standard aerial galleries, such as those maintained by regional GIS repositories or satellite mapping providers, consist of extensive collections of high-resolution orthophotos. In practical deployment, transforming or storing these repository-scale satellite databases into multiple derived structural representations—such as synthetic depth or segmentation mosaics—would introduce unsustainable computational and storage overhead.

Furthermore, operational constraints in cross-view matching typically emerge on the ground platform rather than the overhead satellite repository. Ground acquisition systems (such as autonomous vehicle sensors or mobile edge cameras) frequently encounter variable illumination, transmission bandwidth bottlenecks, or privacy regulations. Converting ground observations into compact structural representations (such as edge maps, depth estimates, or sketches) allows client devices to minimize bandwidth or suppress sensitive photographic details before querying the server. Maintaining the aerial gallery in its native photographic form provides a rigorous testbed to evaluate whether derived structural ground queries can directly interface with pre-existing satellite repositories without requiring server-side data restructuring.

E Derived Representation Generation Pipelines and Analysis

Implementation Details of Derived Representations. All four derived query representations are generated offline from source RGB panoramas using fixed processing settings for each pipeline. The generation procedures are described below.

- **Sketch Generation Pipeline:** Sketch queries are generated using the open-source LightX2V image-to-image framework with the Qwen-Image-Edit-2511 foundation model. Generation uses the 4-step Lightning LoRA adapter (`Qwen-Image-Edit-2511-Lightning-4steps-V1.0`) with a fixed LoRA scale of 1.0. The inference configuration is held constant across images: the number of sampling steps is set to 4, the guidance scale is set to 1.0, and the generation seed is fixed to 42. The pipeline uses scaled dot-product attention and an adaptive resizing option. Before generation, ground panoramas are resampled to an intermediate resolution of 1664×928 pixels (width $\times$ height) using Lanczos filtering. Generated sketches are subsequently resampled back to the original image dimensions using the same filter. The negative prompt is left empty, and the following conditioning prompt is applied to all images:

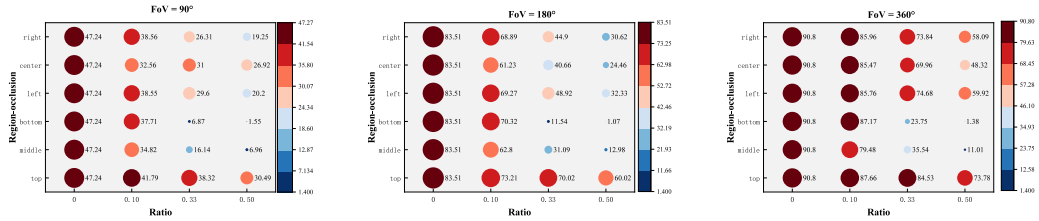
```
Transform this panoramic daytime street view into a hand-drawn
artistic sketch.  STRICT REQUIREMENTS:
1.  ARTISTIC SKETCH: Create a pencil-drawing style sketch, NOT
an edge detection map.
2.  NO OBJECT ALTERATION: Do NOT add/remove any objects.  Only
change the rendering style.
3.  NO HALLUCINATION: Since the output is a sketch of the
input, do not draw anything that is not visible in the input
image.
```

- **Edge Map Extraction:** Edge queries are obtained by converting source panoramas to grayscale and applying the Canny edge detector. The lower and upper hysteresis thresholds are set to 100 and 200, respectively. The OpenCV implementation uses a 3×3 Sobel aperture and the L_1 norm for gradient magnitude, without an additional Gaussian prefilter. The resulting binary maps are inverted so that boundaries appear dark against a light background and are saved as single-channel grayscale images.
- **Relative Depth Estimation:** Depth queries are generated using the pretrained Depth Anything V2 model with a ViT-B backbone. The standard inference pipeline uses a nominal input size of 518 pixels, applies aspect-ratio-aware resizing, and interpolates the predicted maps back to the source image dimensions. Each relative-depth prediction is independently min–max normalized to $[0, 255]$ and applies the *plasma* colormap. These representations encode estimated relative depth without an absolute metric scale.
- **Class-Agnostic Region Masks:** Seg queries are generated using the Segment Anything Model with a ViT-H backbone and its automatic mask generator. The generator uses its standard configuration, including a regular sampling grid with 32 points per side, a predicted intersection-over-union threshold of 0.88, and a mask stability threshold of 0.95. The returned class-agnostic mask proposals are rendered sequentially onto an initially black RGB canvas in their returned order, without additional sorting. Colors are assigned from a fixed 27-entry palette containing repeated colors; assignment restarts from the first entry for each image and cycles through the palette as additional masks are rendered. Later masks overwrite earlier ones in overlapping areas, while uncovered pixels remain black. The resulting maps depict the spatial arrangement of the rendered regions; color assignments do not encode semantic categories.

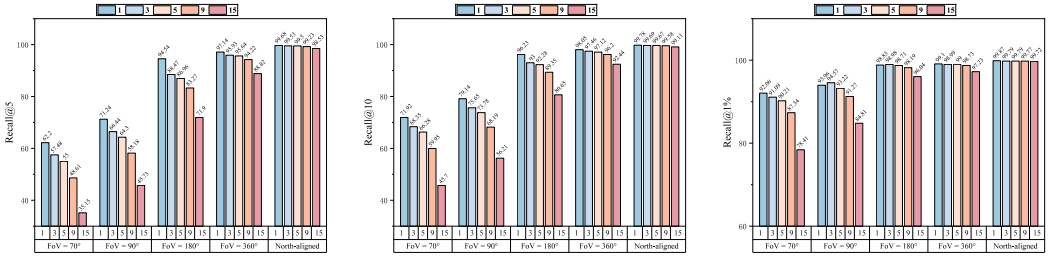
Structural Characteristics and Offline Protocol. Traditional edge detectors like Canny detect local pixel intensity changes, which can produce broken lines and visual noise from tree leaves or road surfaces. In contrast, sketch queries capture clean, continuous outlines of building boundaries, road margins, and horizons while suppressing surface textures and illumination variations. Under a restricted 70° FoV in urban scenes (such as CVACT-Cue), ground images are heavily influenced by perspective-dependent shadows, reflections, and lighting shifts that do not appear in aerial satellite

images. By filtering out these misleading surface appearances, clean structural outlines provide reliable geometric shapes that match the physical layout visible in overhead satellite views.

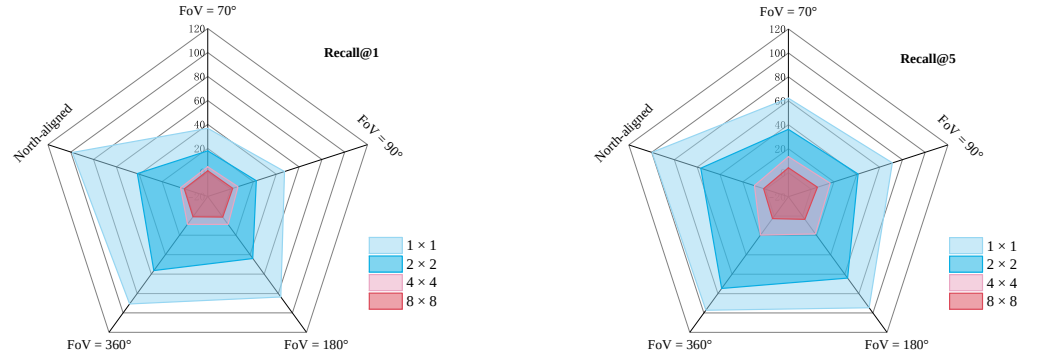
In this benchmark, all generative models and computer vision algorithms function strictly as offline preprocessing tools to separate geometric structure from raw photographic appearance. Downstream geo-localization networks are trained and evaluated on these preprocessed images, without joint optimization or backpropagation into the upstream feature extractors. While choosing alternative generation models, prompt formulations, or algorithm thresholds can yield slightly different line densities, our findings demonstrate that cross-view geo-localization operates reliably on derived structural cues under the representation-generation pipelines and evaluation settings considered here.



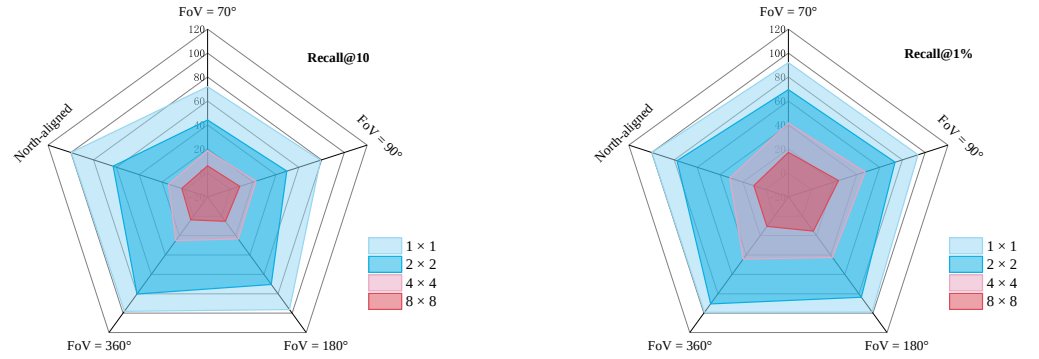
(a) Region occlusion: FoV 90° (b) Region occlusion: FoV 180° (c) Region occlusion: FoV 360°



(d) Gaussian blur: Recall@5 (e) Gaussian blur: Recall@10 (f) Gaussian blur: Recall@1%



(g) Patch shuffling: Recall@1 (h) Patch shuffling: Recall@5



(i) Patch shuffling: Recall@10 (j) Patch shuffling: Recall@1%

Figure 4: **Sensitivity evaluations to spatial perturbations across viewing conditions on CVUSA-Cue.** All evaluations are conducted using Sample4Geo under unknown orientation (spanning $70°, 90°, 180°$, and $360°$ FoVs) and north-aligned panoramas. (a)–(c) Spatial sensitivity under directional region occlusion across $90°$, $180°$, and $360°$ FoVs with masking ratios $\rho \in \{0.10, 0.33, 0.50\}$. (d)–(f) Sensitivity to high-frequency attenuation via Gaussian low-pass filtering across kernel sizes $k \in \{1, 3, 5, 9, 15\}$, reported in Recall@5, Recall@10, and Recall@1% ($k = 1$ indicates the unblurred baseline). (g)–(j) Sensitivity to spatial topology disruption via patch shuffling across four grid partitions ($1 \times 1, 2 \times 2, 4 \times 4, 8 \times 8$), reported in Recall@1, Recall@5, Recall@10, and Recall@1% (1×1 indicates the intact spatial baseline).

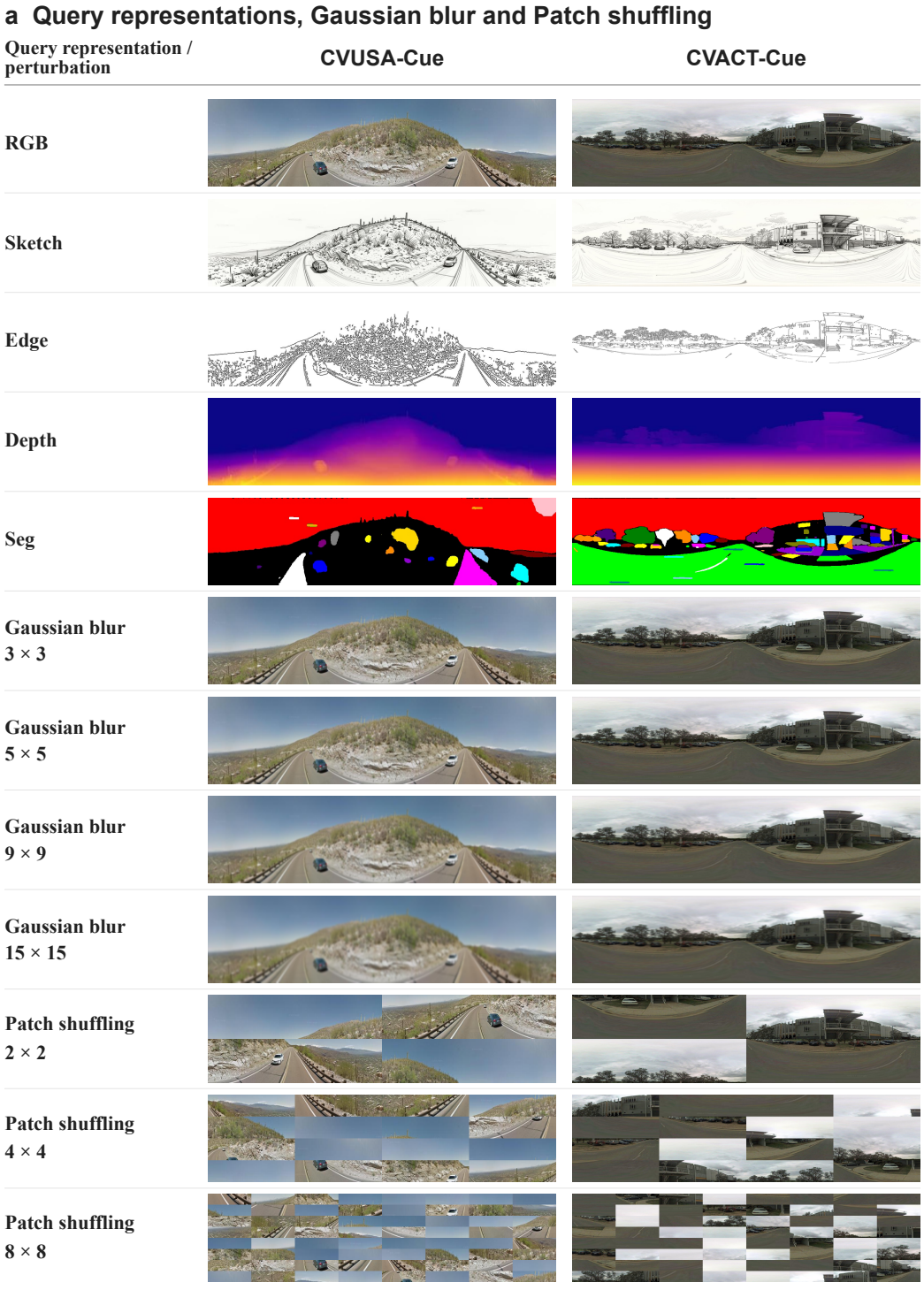


Figure 5: **Visual examples of CVUSA-Cue and CVACT-Cue. (a)** Query representations and spatial perturbations, comprising the five benchmark representations (RGB, Sketch, Edge, Depth, and Seg), four Gaussian blur kernel sizes, and three patch-shuffling grids. The figure continues on the following pages.

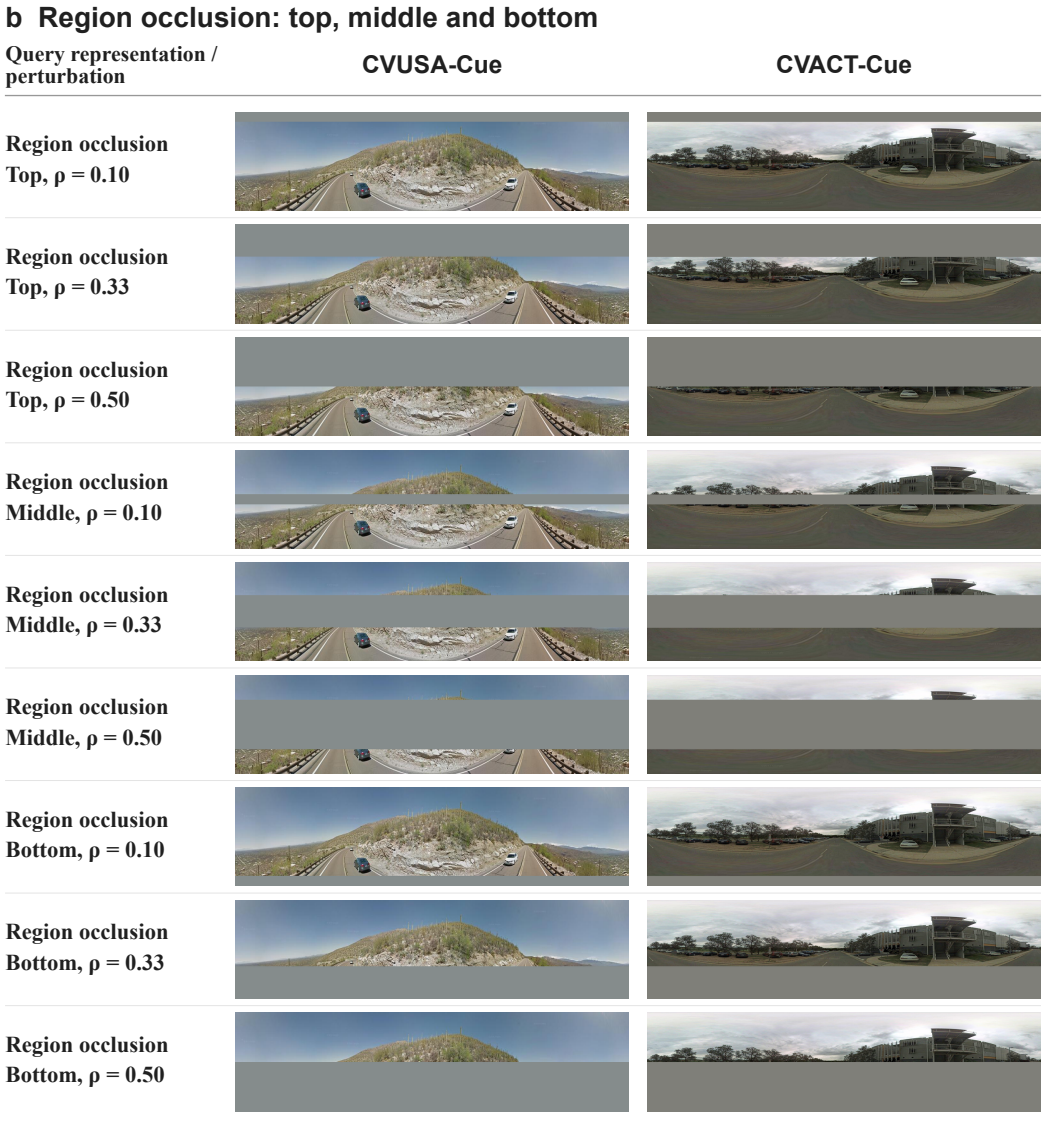


Figure 5: **Visual examples of CVUSA-Cue and CVACT-Cue (continued). (b)** Vertical region occlusion (top, middle, and bottom) under masking ratios $\rho \in \{0.10, 0.33, 0.50\}$, denoting the approximate fraction of the image area masked. Each mask spans the complete horizontal width and is filled with the pre-mask channel means of the input image.

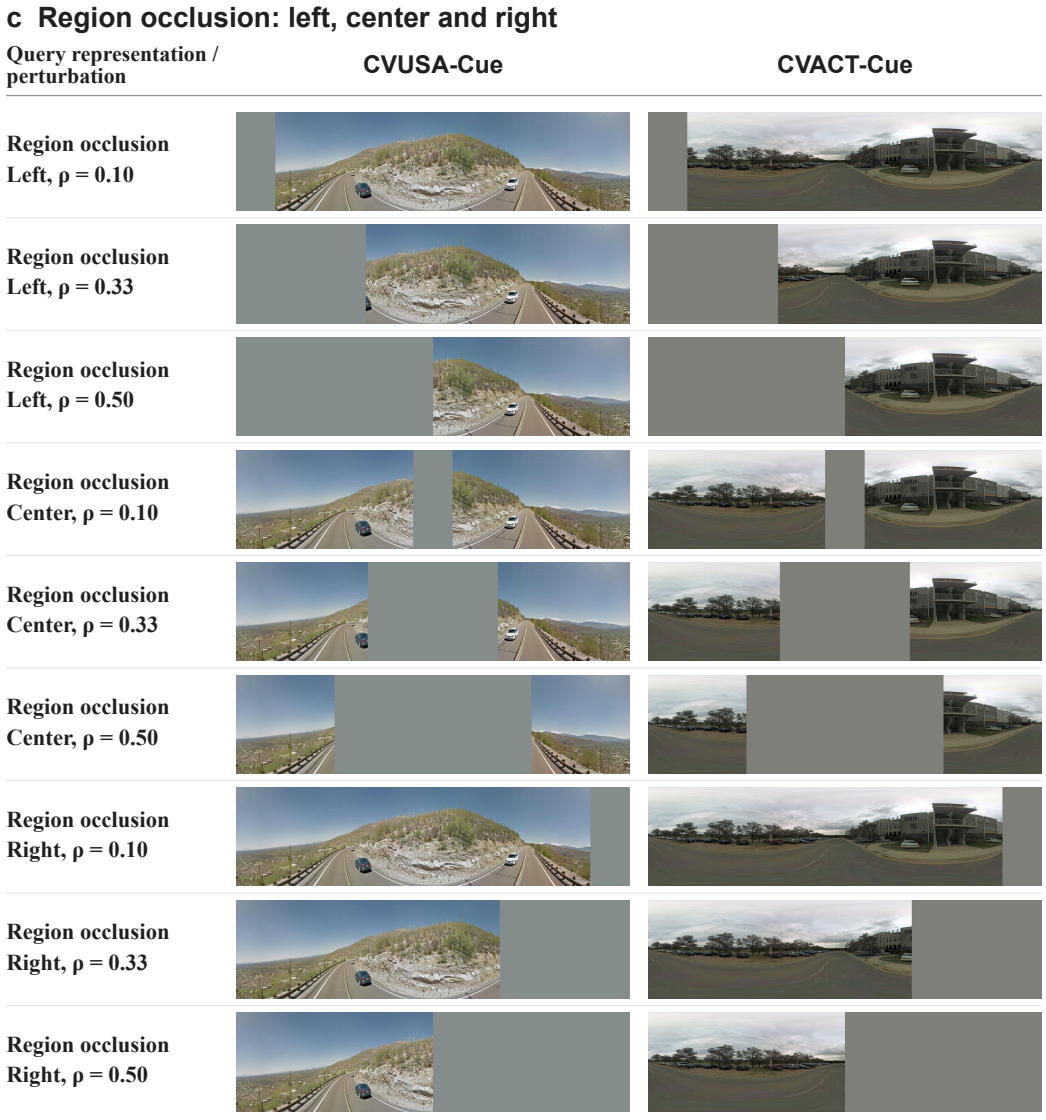


Figure 5: **Visual examples of CVUSA-Cue and CVACT-Cue (continued). (c)** Horizontal region occlusion (left, center, and right) under masking ratios $\rho \in \{0.10, 0.33, 0.50\}$, denoting the approximate fraction of the image area masked. Each mask spans the complete vertical height and is filled with the pre-mask channel means of the input image.

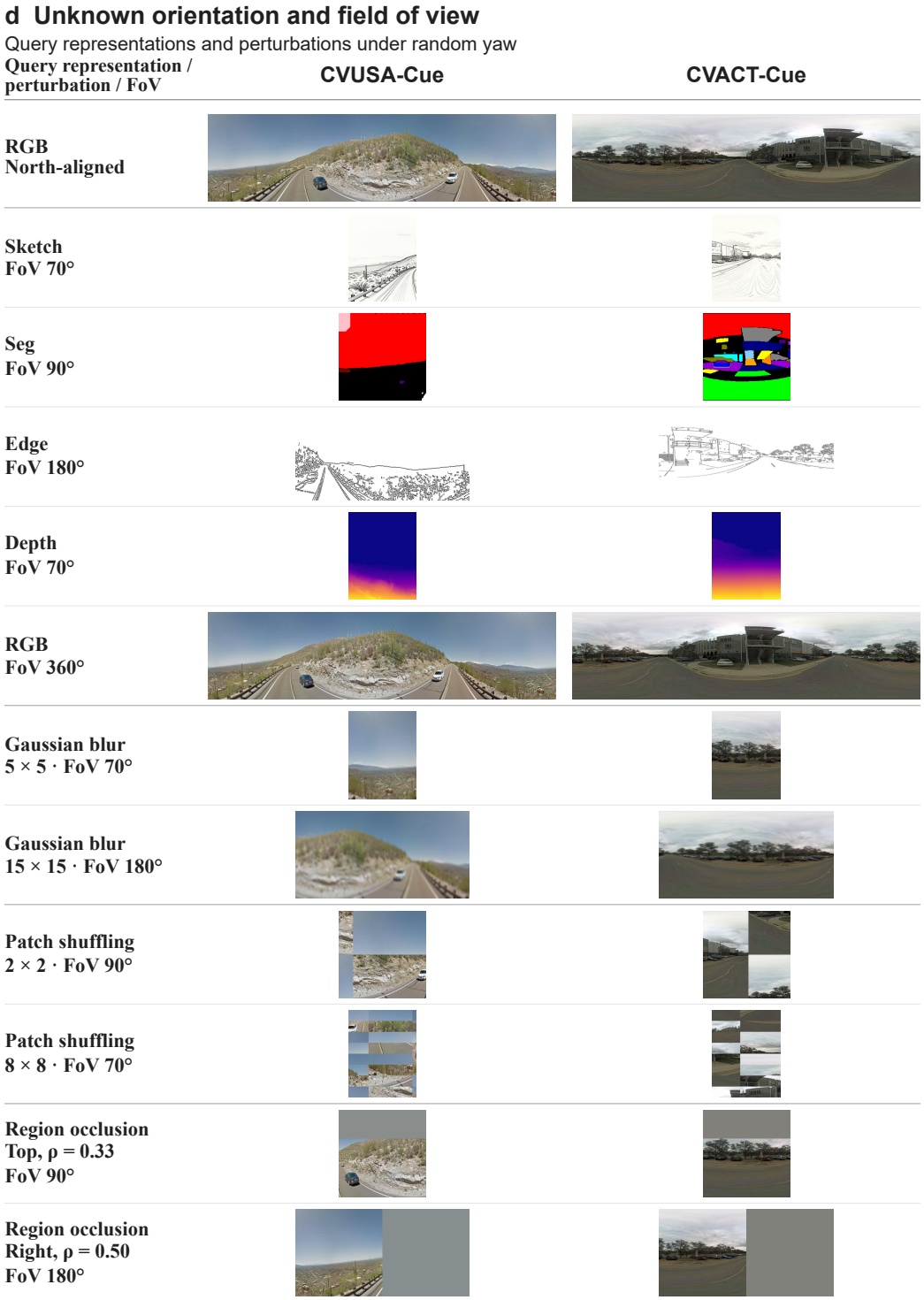


Figure 5: **Visual examples of CVUSA-Cue and CVACT-Cue (continued). (d)** Limited FoVs under unknown orientation. Below the north-aligned RGB reference, query representations and perturbations are depicted under restricted FoVs following random yaw rotation and cyclic cropping across the panorama boundary. Gaussian blur and patch shuffling precede cropping; region occlusion follows cropping and utilizes the pre-mask channel means of the visible query window. Random permutations are fixed solely for figure reproducibility.