MRGeo: Robust Cross-View Geo-Localization of Corrupted Images via Spatial and Channel Feature Enhancement

Le Wu, Lv Bo, Songsong Ouyang, Yingying Zhu*

College of Computer Science and Software Engineering, Shenzhen University, China
2300271073@email.szu.edu.cn, 2250271009@email.szu.edu.cn, 2400101027@mails.szu.edu.cn, zhuyy@szu.edu.cn

Abstract

Cross-view geo-localization (CVGL) aims to accurately localize street-view images through retrieval of corresponding geo-tagged satellite images. While prior works have achieved nearly perfect performance on certain standard datasets, their robustness in real-world corrupted environments remains under-explored. This oversight causes severe performance degradation or failure when images are affected by corruption such as blur or weather, significantly limiting practical deployment. To address this critical gap, we introduce MRGeo, the first systematic method designed for robust CVGL under corruption. MRGeo employs a hierarchical defense strategy that enhances the intrinsic quality of features and then enforces a robust geometric prior. Its core is the Spatial-Channel Enhancement Block, which contains: (1) a Spatial Adaptive Representation Module that models global and local features in parallel and uses a dynamic gating mechanism to arbitrate their fusion based on feature reliability; and (2) a Channel Calibration Module that performs compensatory adjustments by modeling multi-granularity channel dependencies to counteract information loss. To prevent spatial misalignment under severe corruption, a Region-level Geometric Alignment Module imposes a geometric structure on the final descriptors, ensuring coarse-grained consistency. Comprehensive experiments on both robustness benchmark and standard datasets demonstrate that MRGeo not only achieves an average R@1 improvement of 2.92% across three comprehensive robustness benchmarks (CVUSA-C-ALL, CVACT_val-C-ALL, and CVACT_test-C-ALL) but also establishes superior performance in cross-area evaluation, thereby demonstrating its robustness and generalization capability.

Code — https://github.com/WLHASH/MRGeo

1 Introduction

Cross-view geo-localization (CVGL) aims to retrieve the corresponding satellite image with GPS coordinates given a street-view query image. It serves as an auxiliary means or an effective supplement for providing geo-localization information in fields such as autonomous driving (Häne et al. 2017; Kim and Walter 2017) and robot navigation (McManus et al. 2014) when encountering urban canyons, GPS signal degradation, or absence, showcasing significant application potential (Zhu et al. 2023; Li and Zhu 2025).

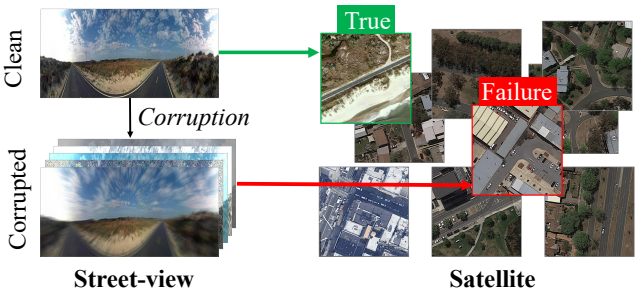

Figure 1: The fragility of existing CVGL models under real-world corruptions. While a model can easily match a clean street-view image to its correct satellite counterpart (top), its performance collapses when the query is affected by common corruptions like weather or blur, often leading to a complete failure in localization (bottom).

In recent years, with the advancement of CVGL, existing methods (Ye et al. 2024) achieved near-perfect performance on some standard datasets (*e.g.*, CVUSA). However, an undeniable fact is that these datasets are collected under clean conditions and do not fully reflect the common corrupted scenes (*e.g.*, rain) prevalent in the real world. This distribution shift between training data and real-world data causes a drastic performance degradation or even failure when models trained on clean images are applied to corrupted data, limiting the reliability and robustness of models in practical applications (Zhang and Zhu 2024), as illustrated in Fig. 1. Therefore, enhancing model robustness in corrupted environments is crucial for building truly reliable and deployable CVGL systems.

To comprehensively and fairly assess the robustness of CVGL methods under corruption, Zhang *et al.* first proposed a robustness benchmark for CVGL (Zhang and Zhu 2024). However, to the best of our knowledge, no prior work has proposed a systematic method to address the performance degradation of CVGL methods on corrupted images. To narrow down this gap, we investigate the impact of corrupted images on feature representation capabilities and conclude that the impact of corruption on feature representation is

*Corresponding author.

multi-level. Specifically, in the spatial domain, fine-grained CVGL on clean images heavily relies on fine-grained local details (*e.g.*, textures). Yet, these details are the most fragile and unreliable features under corruption (Hendrycks and Dietterich 2019; Yin et al. 2019). Conversely, global semantic information (*e.g.*, road layouts) is more robust but often too coarse to distinguish between visually similar but geographically distinct locations. In the channel domain, corruption introduces spatially heterogeneous perturbations at the pixel level. These perturbations accumulate through the network (Cui and Knoll 2023), distorting channel-wise information (Shi et al. 2023; Lee et al. 2024). But standard feature extractors and attention mechanisms, often employing a one-size-fits-all approach to channel re-weighting, are ill-suited to handle such complex, multi-granularity distortions (Hu, Shen, and Sun 2018; Wang et al. 2019).

To address the aforementioned challenges, this paper introduces MRGeo, a systematic, hierarchical method designed to counteract the significant performance degradation of CVGL models under image corruption. Our key insight is that achieving true robustness necessitates a two-level strategy: 1) enhancing the intrinsic quality of core features and 2) imposing a strong geometric structural prior.

Specifically, in the spatial domain of features, the corruption creates a critical conflict: corruptions like blur primarily destroy fine-grained local details, while leaving coarse global structures relatively intact. Conversely, corruptions like fog and snow can obscure larger regions, degrading global semantic information. In the channel domain of features, corruptions such as contrast or JPEG directly perturb the pixel values, leading to distorted channel-wise statistics and a subsequent loss of semantic information.

To counteract this degradation at the feature level, we propose the Spatial-Channel Enhancement Block (SCEB). This block comprises two synergistic sub-modules engineered to fundamentally improve core feature quality: 1) Spatial Adaptive Representation Module (SARM) as a dynamic arbitration mechanism that explicitly models global semantics and local details in parallel. It features a learned gating system that acts as a reliability estimator for local features, adaptively suppressing their contribution in the presence of corruption while leveraging them in clean conditions. 2) We introduce a Channel Calibration Module (CCM) to counteract channel-level information loss. Moving beyond simplistic channel-wise scaling, CCM performs a more sophisticated compensatory calibration. It achieves this by decoupling and modeling multi-granularity channel dependencies, allowing it to dynamically correct for channel distortions at spatial location with a richer, more comprehensive context. However, feature-level enhancement alone is insufficient. Under severe corruption, features can become highly ambiguous, causing the model to erroneously match semantically similar but geographically disparate regions. To mitigate the risk of spatial misalignment, we introduce the Region-level Geometric Alignment Module (RGAM). Based on the inherent spatial correspondence between the two views, RGAM partitions the feature map into a fixed grid and concatenates the resulting regional features in a consistent, predefined order. This structural constraint compels the model to perform matching within corresponding coarse-grained geographic areas, which fundamentally ensures geometric consistency, thereby reinforcing the robustness of the entire localization system. Our main contributions are as follows:

- To the best of our knowledge, this paper proposes a novel and systematic method for the performance degradation of CVGL methods on corrupted images, significantly enhancing the model's robustness and generalization under corruption.
- To address the multi-level impact of corruption on feature representation, a hierarchical defense strategy is proposed. At the feature level, **SARM** enhances spatial representations by dynamically arbitrating between global and local information, while **CCM** calibrates channel-wise distortions to counteract information loss. At the structural level, **RGAM** imposes a rigid geometric prior, ensuring robust matching even under severe corruption.
- Extensive experiments on CVGL robustness benchmark and standard datasets comprehensively demonstrate that MRGeo not only achieves excellent performance on robustness benchmark but also achieves superior performance on cross-area tasks, fully demonstrating its effectiveness and robustness.

2 Related Work

2.1 Cross-View Geo-Localization

Early CVGL research on datasets like CVUSA (Zhai et al. 2017) and CVACT (Liu and Li 2019) focused on improving retrieval accuracy using Siamese networks (Shi et al. 2019), attention mechanisms (Yang, Lu, and Zhu 2021), and Transformers (Zhu, Shah, and Chen 2022). Subsequent efforts addressed geometric complexities in datasets such as VIGOR (Zhu, Yang, and Chen 2021) with sophisticated techniques like BEV rectification (Ye et al. 2024) and advanced sampling (Deuser, Habel, and Oswald 2023). While these methods progressively pushed performance boundaries, they shared a common trait: a heavy reliance on precise, fine-grained visual cues. This pursuit of precision has inadvertently created a critical vulnerability, as these fragile cues are easily compromised by common image corruptions.

2.2 Corrupted Image Robustness in CVGL

The performance degradation of models on corrupted images was a known challenge in computer vision (Hendrycks and Dietterich 2019). In the context of CVGL, this issue was first systematically quantified by Zhang *et al.* (Zhang and Zhu 2024), who constructed robustness benchmarks and revealed drastic performance drops across existing methods. However, their work diagnosed the problem without offering a methodological solution. To fill this critical gap, we introduce MRGeo, a novel and systematic method designed to enhance the intrinsic robustness of CVGL models. Unlike prior work focused on geometric precision, MRGeo directly confronts feature degradation caused by corruption, making it a crucial step towards reliable real-world deployment.

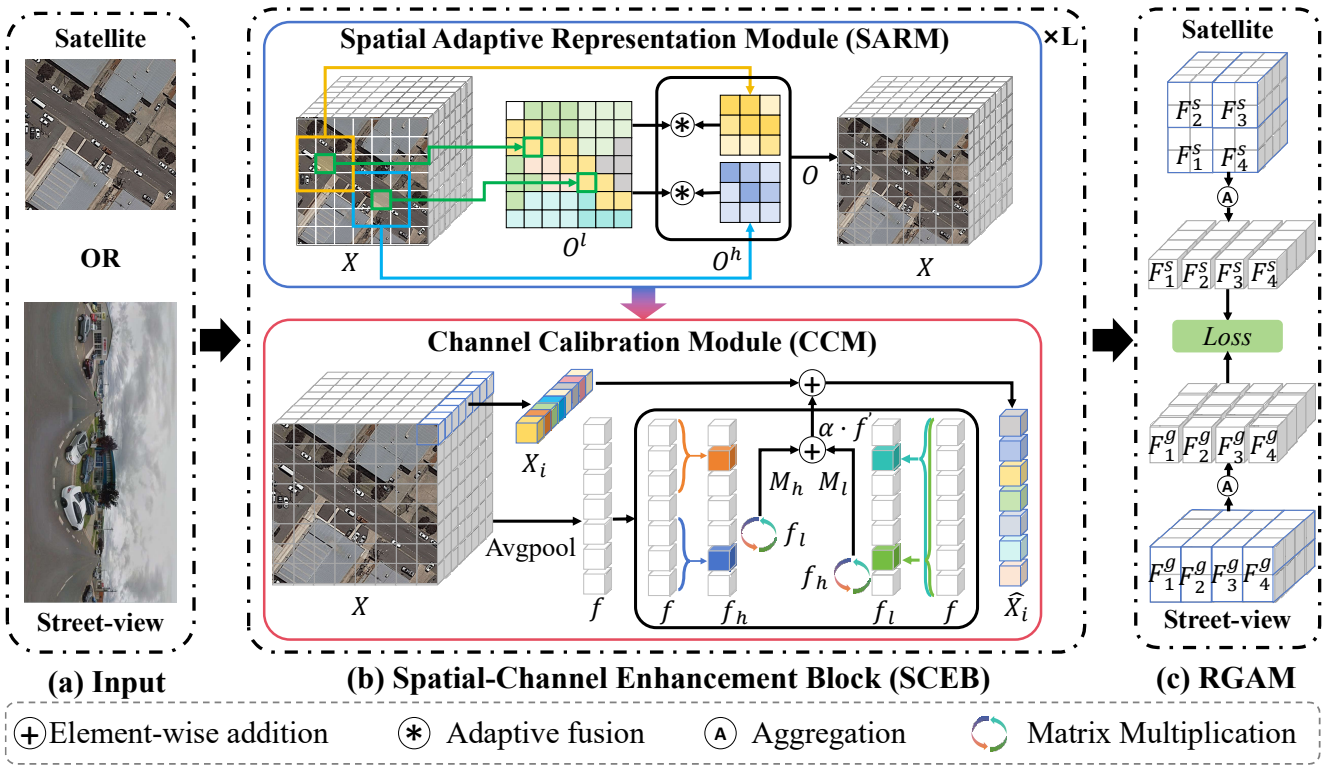

Figure 2: Overview of our MRGeo architecture. The framework processes street-view and satellite images through a shared-weight backbone containing our proposed **Spatial-Channel Enhancement Block (SCEB)**. SCEB enhances feature quality via its two sub-modules: the **SARM** and the **CCM**. Finally, the **Region-level Geometric Alignment Module (RGAM)** imposes a structural constraint on the enhanced features to generate robust final descriptors for retrieval.

3 Method

3.1 Problem Definition

Given a set of street-satellite image pairs $\{I_i^g, I_i^s\}_{i=1}^N$, where N denotes the number of image pairs, g and s represent the street-view and satellite respectively. Assume the image encoders f^s and f^g are trained on the ideal distribution $\mathcal{D}$. In the ideal scenario, we find matching pairs by minimizing the distance between features from the two views, expressed as:

$$\mathcal{D} \sim \{d(f^g(I_i^g), f^s(I_i^s)) < d(f^g(I_i^g), f^s(I_j^s))\}, i \neq j \quad (1)$$

where $d(\cdot)$ denotes the $\mathcal{L}_2$ distance. However, real-world data is often affected by various corruptions $\mathcal{C}$ (*e.g.*, rain, snow, blur, noise), causing the data distribution to deviate from $\mathcal{D}$ and follow a more complex corrupted data distribution $\mathcal{P}$. On corrupted data, models trained on distribution $\mathcal{D}$ experience a significant performance degradation or even failure. Therefore, this paper primarily studies the task of accurate localization on corrupted data, and formulated it as:

$$\mathcal{P} \sim \{d(f^g(\mathcal{C}(I_i^g)), f^s(I_i^s)) < d(f^g(\mathcal{C}(I_i^g)), f^s(I_j^s))\} \quad (2)$$

3.2 Spatial Adaptive Representation Module

The Spatial Adaptive Representation Module (SARM) is designed to resolve a fundamental conflict in robust feature extraction: corruption primarily degrades fine-grained local details, while robust global semantics remain relatively stable. To address this, SARM treats them as distinct, complementary information streams and by introducing a mechanism for dynamic, content-aware arbitration.

Specifically, for an input feature map $X \in \mathbb{R}^{C\times H\times W}$, where C,H, and W denote the number of channels, height, and width respectively, we first capture global semantic structure $O^l \in \mathbb{R}^{C\times H\times W}$, which are less affected by noise or local occlusions. This is achieved via a standard self-attention mechanism where each spatial feature vector $x_i \in \mathbb{R}^C$ attends to all other positions in the map. The process is formulated as:

$$\begin{aligned} O_i^l &= \mathcal{S}(x_i, X) \cdot (XW_v^l) \\ \mathcal{S}(x_i, X) &= (x_i W_q^l) \cdot (XW_k^l)^T \end{aligned} \quad (3)$$

Here, W_q^l, W_k^l, and W_v^l are learnable projection matrices. $\mathcal{S}(x_i, X)$ computes the similarity scores. To extract corruption-sensitive, fine-grained local context $O^h \in \mathbb{R}^{C\times H\times W}$, we apply a similar attention mechanism but restrict the interaction of x_i to its local $k \times k$ neighborhood, denoted as $\mathcal{N}_k(x_i) \in \mathbb{R}^{k^2\times C}$:

$$\begin{aligned} O_i^h &= \mathcal{S}(x_i, \mathcal{N}_k(x_i)) \cdot (\mathcal{N}_k(x_i) W_v^h) \\ \mathcal{S}(x_i, \mathcal{N}_k(x_i)) &= (x_i W_q^h) \cdot (\mathcal{N}_k(x_i) W_k^h)^T \end{aligned} \quad (4)$$

where W_q^h, W_k^h, W_v^h are the corresponding learnable matrices. Finally, the O^l and O^h are aggregated using a gating mechanism that acts as a learned reliability estimator to form the output feature O:

$$O = O^l \circledast O^h = O^l + \sigma(FC([O^l, O^h])) \odot O^h \quad (5)$$

Here, $\circledast$ denotes an adaptive fusion operation, $[O^l, O^h]$ denotes channel-wise concatenation of the O^l and O^h. A gating vector is generated by passing the concatenated features through a fully connected layer (FC) and a sigmoid activation function (σ). This gate, with values in $[0, 1]$, dynamically scales the contribution of the local features O^h via element-wise multiplication ($\odot$) before they are added to the global features O^l. This adaptive fusion allows the model to learn a dynamic trade-off: in the presence of corruption, it can suppress the unreliable O^h by producing low gate values and rely on the robust O^l. Conversely, for clean images, it can leverage O^h to achieve higher precision.

This gating mechanism is conceptually analogous to those in LSTMs and GRUs, where the model learns to control the flow of information based on the input context. Here, the "context" is the level of corruption implicitly encoded in the concatenated features $[O^l, O^h]$. The sigmoid function ensures the gate values are in a stable $[0, 1]$ range, acting as a soft switch that dynamically arbitrates between global robustness and local precision. The effectiveness of this design is validated in our ablation studies (Tab 3, #5-#7).

3.3 Channel Calibration Module

While SARM primarily focuses on spatial robustness, the channel calibration module (CCM) is designed to mitigate the multi-faceted impact of corruption on feature channels. Previous studies have shown that not all channel information is highly relevant to the task (Chen et al. 2023; Yun and Ro 2024; Fu and Zhu 2025), but high-level channel semantic features can enhance the model's discriminative ability (Lu et al. 2024; Woo et al. 2018). Therefore, we aim to enhance key channel representations by constructing fine-grained global channel dependencies and dynamically adjusting channel representations at spatial location, thereby strengthening the expression of key channels.

The process begins by extracting a global channel representation $f \in \mathbb{R}^C$ from the input feature X via global average pooling. To explore the dynamic, complementary relationships between channels, we decouple f into two components. Specifically, we use a 1D convolutional layer (to capture local dependencies between adjacent channels) and a linear layer (to capture the overall pattern of global channels) to process f, obtaining global structural features f_l and local detail features f_h, respectively. Then, we calculate their correlation to achieve mutual strengthening and correction. The specific process is as follows:

$$f' = \underbrace{\mathcal{B}(f_h \cdot f_l^T)}_{\mathcal{M}_h} + \underbrace{\mathcal{B}(f_l \cdot f_h^T)}_{\mathcal{M}_l} \quad (6)$$

Here, the matrix multiplication (*e.g.*, $f_l \cdot f_h^T$) computes a cross-channel correlation matrix, where each element quantifies the interaction strength between global structural channel and local detail channel. The function $\mathcal{B}$ performs a summation along the horizontal axis of this matrix, transforming the correlation scores into a vector. The intermediate terms, $\mathcal{M}_h \in \mathbb{R}^C$ and $\mathcal{M}_l \in \mathbb{R}^C$, represent the mutual enhancement between the structural (f_l) and detail (f_h) channel features. In essence, this step allows the global channel patterns to calibrate the local channel responses, and vice-versa, leading to a more refined and robust channel representation.

The final result, a highly informative and robust channel correction signal f', is used to perform a dynamic residual correction on the feature map at spatial location. After an initial projection of the input features X to $\hat{X} \in \mathbb{R}^{C\times H\times W}$, we calibrate each feature vector $\hat{x_i}$ as follows: $\hat{x_i}' = \hat{x_i} + \alpha f'$, where α is a learnable parameter controlling the injection strength of f'. By making α learnable, the network can autonomously determine the optimal degree of calibration needed at different layers and for different tasks, avoiding a manually tuned hyperparameter. It allows global channel statistics to enhance salient features and suppress noise at each spatial location, effectively mitigating the channel-level perturbations caused by corruption and improving overall feature robustness.

3.4 Region-level Geometric Alignment Module

Inspired by prior work, we recognize the importance of geometric consistency between views for the CVGL task. To further address the feature loss and interference caused by corruption, we introduce the region-level geometric alignment module (RGAM). As illustrated in Fig. 2(c), the RGAM operates on the feature map X from SCEB. It first spatially partitions X into a 2 $\times$ 2 grid for satellite (and a 1 $\times$ 4 grid for street-view) into four non-overlapping regions ($F_1^i, F_2^i, F_3^i, F_4^i$), where the superscript $i \in \{g, s\}$ indicates the view type (street-view or satellite). Then each region is aggregated through average pooling $\mathcal{A}(\cdot)$. Finally, these regional vectors are concatenated in a fixed order to construct the final descriptor f^i, formulated as:

$$f^i = [\mathcal{A}(F_1^i), \mathcal{A}(F_2^i), \mathcal{A}(F_3^i), \mathcal{A}(F_4^i)] \quad (7)$$

This strategy enforces a consistent spatial layout, ensuring that matching occurs between corresponding local regions. By providing robust structural cues, this approach effectively counters the loss of fine-grained details caused by corruption and enhances overall model robustness.

3.5 Optimization Objective

Prior studies (Deuser, Habel, and Oswald 2023) have established the effectiveness of contrastive learning for cross-view geo-localization. Based on this, we employ the InfoNCE loss (Oord, Li, and Vinyals 2018; Hadsell, Chopra, and LeCun 2006) with refined formulation:

$$\mathcal{L}_{\text{InfoNCE}}(q, R) = -\log\left[\frac{\exp\left(\mathbf{q}^\top \mathbf{r}_+/\tau\right)}{\sum_{i=0}^{K}\exp\left(\mathbf{q}^\top \mathbf{r}_i/\tau\right)}\right] \quad (8)$$

where $\mathbf{q} \in \mathbb{R}^D$ denotes the ℓ_2-normalized query embedding from street-view, $R = \{\mathbf{r}_+, \mathbf{r}_1, ..., \mathbf{r}_K\}$ represents a

Method	Publication	CVUSA-C-ALL				CVACT_val-C-ALL				CVACT_test-C-ALL			
		R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
L2LTR	NIPS'21	87.93	95.45	97.01	99.01	82.13	93.34	94.93	98.10	57.20	82.59	87.23	98.09
TransGeo	CVPR'22	82.72	91.95	94.03	97.92	74.04	86.19	89.10	94.98	52.18	74.35	78.99	95.03
GeoDTR	AAAI'23	84.64	93.29	95.01	98.24	77.40	88.95	91.28	95.91	52.87	78.84	83.17	95.84
Sample4G	ICCV'23	93.36	97.32	98.08	99.22	86.71	94.35	95.60	98.04	66.33	88.05	90.75	98.07
EP-BEV	ECCV'24	84.33	94.38	96.06	98.39	82.07	92.19	94.09	97.52	59.25	83.12	87.02	97.49
DReSS	JPRS'25	92.42	97.50	98.33	99.25	87.39	94.44	95.61	98.00	68.12	88.81	91.24	98.07
MRGeo	–	**95.09**	**97.99**	**98.48**	**99.38**	**90.45**	**96.50**	**97.33**	**98.91**	**71.16**	**92.24**	**94.21**	**98.80**

Table 1: Experimental results of cross-view geo-localization methods on comprehensive corruption robustness benchmarks.

Method	Clean	CVACT_val-C												
		Weather					Blur				Digital			
		Snow	Frost	Fog	Bright	Spatter	Defocus	Glass	Motion	Zoom	Contrast	Pixel	JPEG	$R@1_{cor}$
L2LTR	84.89	71.03	77.93	83.50	81.17	73.78	83.98	85.07	84.00	49.79	**79.15**	85.07	83.40	78.16
TransGeo	84.95	47.65	58.51	32.91	72.67	67.13	81.43	84.83	81.80	36.34	22.18	84.92	83.74	62.84
GeoDTR	86.21	48.24	71.74	83.26	84.60	61.39	79.11	85.51	73.44	8.26	55.48	86.01	85.19	68.52
Sample4G	90.81	78.69	84.68	86.97	88.80	84.97	89.17	90.22	89.03	50.00	51.99	90.26	89.52	81.20
EP-BEV	88.91	69.43	78.67	81.55	86.08	79.77	87.52	88.55	87.36	38.15	49.09	88.77	86.92	76.82
DReSS	91.32	81.23	86.50	88.14	89.97	85.48	90.02	90.89	90.12	49.98	52.78	90.95	89.95	82.17
MRGeo	**92.67**	**88.75**	**90.97**	**91.25**	**91.95**	**90.22**	**91.47**	**92.35**	**91.79**	**65.31**	72.71	**92.34**	**91.67**	**87.56**

Table 2: Experimental results of 7 cross-view geo-localization methods on the CVACT_val-C datasets. We report the R@1 performance of each method under different corruptions (obtained by averaging the 5 corruption severities), as well as the average $R@1_{cor}$ under all corruption types.

batch containing one positive satellite embedding $\mathbf{r}_+$ and K negative samples $\mathbf{r}_i$, with $\tau > 0$ being a temperature hyperparameter. The loss minimizes the negative log-likelihood of identifying the correct geospatial correspondence, effectively pulling $\mathbf{q}$ toward $\mathbf{r}_+$ in the embedding space while pushing it away from all $\mathbf{r}_i$.

4 Experiment

4.1 Datasets and Evaluation Metrics

Datasets. We evaluate on CVGL databases—CVUSA (Zhai et al. 2017), CVACT (Liu and Li 2019)—and their robust variants: the fine-grained robustness benchmarks and the comprehensive robustness benchmark (Zhang and Zhu 2024). All datasets contain precisely center-aligned street-view-satellite image pairs. The fine-grained robustness benchmarks (CVUSA-C and CVACT-C) include 16 different corruption types, each with 5 severity levels (totaling 80 evaluation sets), to systematically evaluate model robustness; the comprehensive robustness benchmark (CVUSA-C-ALL, CVACT_val-C-ALL, and CVACT_test-C-ALL) aggregate all corruption types into a single evaluation set. **Metrics.** Following previous work (Zhu et al. 2023), retrieval performance is evaluated by R@K (K $\in \{1, 5, 10, 1\% \}$), which represents the probability of correctly identifying the matching image within the top K retrieved reference images based on the query image. To specifically quantify robustness, we also report the average R@1 across all 16 corruption types, denoted as $R@1_{cor}$.

4.2 Implementation Details

MRGeo is designed based on ViT architecture and pre-trained on LVD-142M, with an output feature dimension of 768. After processing by RGAM, the final output dimension is 3072 (4×768). For optimization, consistent with prior work (Ye et al. 2024), we use the AdamW optimizer (Kingma 2014) combined with a cosine decay learning rate scheduler (initial learning rate $1e^{-4}$). The hyperparameter k for the SARM is set to 3, a value justified by our ablation studies (see Section 4.4). Training is conducted on 4 NVIDIA V100 GPUs with a batch size of 16 for 20 epochs, with the first epoch serving as a warm-up phase.

4.3 Comparison with Existing Methods

We benchmark our method against 6 state-of-the-art (SOTA) methods, including L2LTR (Yang, Lu, and Zhu 2021), TransGeo (Zhu, Shah, and Chen 2022), GeoDTR (Zhang et al. 2023), Sample4Geo (Deuser, Habel, and Oswald 2023), EP-BEV (Ye et al. 2024), and DReSS (Xia et al. 2025). The comparison is performed on CVUSA-C, CVACT_val-C, CVUSA-C-ALL, CVACT_val-C-ALL, and CVACT_test-C-ALL datasets. To ensure a fair and direct comparison, results for EP-BEV and DReSS are based on official public code on these benchmarks.

Comprehensive Corruption Robustness Evaluation. We evaluate MRGeo's overall robustness against corruption on the comprehensive benchmarks, with the results presented in Tab 1. MRGeo establishes a new state-of-the-art (SOTA), achieving the best performance across

#	SARM	CCM	RGAM	CVACT_test-C-ALL			
				R@1	R@5	R@10	R@1%
1	×	×	×	65.84	88.32	91.03	98.30
2	×	×	✓	67.86	89.90	92.26	98.45
3	✓	✓	×	68.90	90.76	92.99	98.56
4	✓	✓	✓	**71.16**	**92.24**	**94.21**	**98.80**
5	2:1	✓	✓	70.40	91.37	93.38	98.50
6	1:1	✓	✓	69.16	90.82	93.86	98.58
7	1:2	✓	✓	59.55	83.51	86.91	97.24
8	✓	FC	✓	53.59	79.59	84.12	97.18
9	✓	XCiT	✓	69.87	90.89	93.09	98.62
10	✓	MFCM	✓	59.56	82.54	85.72	96.32

Table 3: The Ablation study of Components.

Method	CVUSA → CVACT			CVACT → CVUSA		
	R@1	R@5	R@1%	R@1	R@5	R@1%
SAFA†	30.40	52.93	85.82	21.45	36.55	69.83
DSM†	33.66	52.17	79.67	18.47	34.46	69.01
TransGeo	37.81	61.57	89.14	18.99	38.24	88.94
L2LTR†	52.58	75.81	93.51	37.69	57.78	89.63
GeoDTR†	53.16	75.62	93.80	44.07	64.66	90.09
Sample4G	56.62	77.79	94.69	44.95	64.36	90.65
OR-CVFI†	68.07	83.25	–	40.13	56.24	–
MRGeo						
1×1	82.05	93.09	97.67	43.65	62.29	88.23
3×3	81.57	92.84	97.56	**47.73**	**65.65**	**90.85**
5×5	**82.73**	**93.21**	**97.92**	45.00	63.38	88.63

Table 4: Cross-area evaluation when trained on the CVUSA dataset and evaluated on CVACT and vice versa. † denotes models that use the polar transformation. – indicates that the corresponding data are not provided. The bottom section shows the impact of the hyperparameter k in SARM on cross-area evaluation.

all three datasets. Specifically, it surpasses the SOTA method (DReSS) with R@1 performance gains of 2.67% on CVUSA-C-ALL and 3.06% on CVACT_val-C-ALL. This superiority is particularly evident on the more challenging CVACT_test-C-ALL dataset, where our method achieves a 3.04% improvement. These results demonstrate that MRGeo can effectively handle diverse types of corruption in the real world, significantly outperforming prior methods.

Fine-grained Corruption Robustness Evaluation. To conduct a detailed analysis of model's performance under different corruption types, we focus our discussion on the CVACT_val-C benchmark, whose complex scenes provide a more compelling testbed for robustness. The experimental results, presented in Tab 2, show that "Snow", "Contrast", and "Zoom" corruptions have the most significant impact on model performance. Even under these most impactful corruptions, our method demonstrates remarkable robustness, surpassing the SOTA method (DReSS) with performance gains of 7.52% on "snow" and 15.33% on "Zoom".

Notably, our performance on "Contrast" warrants a closer look. While achieving a substantial 19.93% gain compared to DReSS, its average performance is sub-optimal. This is because at severity level 5, extreme contrast reduction creates low-variance features, which flattens SARM's attention distributions and renders its adaptive gate ineffective. Concurrently, the resulting homogenization of channel-wise activations provides no distinct patterns for CCM to model, thus failing to generate a meaningful calibration signal to enhance key features. Nevertheless, our method still secures competitive performance in this type of corruption.

4.4 Ablation Studies

To ascertain the effectiveness of MRGeo's components, we conduct the following ablation experiments: 1) the hyperparameter k of SARM; 2) the adaptive fusion operation $\circledast$ of SARM; 3) the effectiveness of CCM; and 4) the effectiveness of RGAM.

Hyperparameter k of SARM. We find that the selection of the hyperparameter k in SARM is highly correlated with scene complexity. Consequently, we use the cross-area evaluation to best illustrate this dependency. As shown in Tab 4, when $k = 1$, the extracted local detail features are too fine-grained, causing local details to suppress the representational capacity of global structural features. When $k = 3$, it balances the capture of local details and global context perception well on the CVACT dataset, avoiding over-reliance on local details or omission of global information, thereby achieving optimal performance on the CVACT→CVUSA task. When $k = 5$, it achieves optimal performance on the CVUSA→CVACT task because the CVUSA dataset has fewer scene categories and sparser unique local detail features; a larger receptive field can more effectively capture and understand local feature distributions.

Adaptive fusion operation $\circledast$ of SARM. To verify the effectiveness of adaptive fusion operation $\circledast$, we set the ratio of O^l and O^h to a fixed value. As shown in Tab 3 (#4, #5, #6, and #7), as the proportion of O^h increases, the proportion of global semantics O^l, which is more robust to corruption, decreases, and model's performance also declines. In contrast, with adaptive fusion, the ratio changes with the increase of network depth, which aligns with the findings of existing studies (Rao et al. 2021), indicating MRGEO has different feature preferences in regions of different depths.

Effectiveness of the CCM. We replace our channel calibration module with FC (Houlsby et al. 2019), XCiT (Touvron et al. 2021), and MFMC (Lu et al. 2024) respectively. The experimental results, shown in Tab 3 (#4, #8, #9, and #10), indicate that MFMC, as a multi-scale method, performs better than simple FC, demonstrating the importance of high-level channel features for enhancing feature representation capabilities. XCiT considers the positional information of different tokens and interacts with the channel information of different tokens, also achieving some improvement. Our method, by dynamically adjusting channel weights based on global channel features that are more robust to corruption, more effectively compensates for channel

information perturbations and losses caused by corruption, significantly enhancing feature robustness.

Effectiveness of the RGAM. As shown in Tab 3 (#1, #2, #3, and #4), we verify the effectiveness of the region-level geometric alignment module. The introduction of this module improved R@1 by 2.02% and 2.26%, respectively. This indicates that enforcing consistency of regional features between views can effectively enhance the model's robustness when facing corruption, further narrowing the performance gap with the ideal scenario. Although its reliance on center-alignment is a limitation, an ablation study on the VIGOR dataset confirms its importance, showing a performance drop from 77.99% to 73.03% upon its removal.

Furthermore, comparing experiments in Tab 3 (#2, #3, and #4) highlights a crucial synergistic effect. While adding only RGAM (#2) or only the SARM+CCM block (#3) improves performance, the complete model (#4) achieves a disproportionately larger gain (R@1 increases by 5.32% over the baseline). This super-additive boost confirms that our modules are complementary: the SARM and CCM tackle corruption at the feature level (spatial and channel), while the RGAM enforces geometric consistency at the descriptor level. This demonstrates that our holistic method, addressing the problem from multi-level, is substantially more effective than addressing any single aspect in isolation.

Cross-area Evaluation. To validate the model's generalization capability, we conduct a cross-area evaluation, with the results summarized in Tab 4. In the CVUSA $\rightarrow$ CVACT task, our method achieves a remarkable 13.5% R@1 improvement over the previous SOTA. Conversely, in the challenging CVACT$\rightarrow$CVUSA task, our model still surpasses the SOTA by 2.78%. The performance difference stems from the datasets' characteristics. Training on the simpler CVUSA encourages the model to learn highly generalizable geometric features, leading to a substantial performance leap on the complex CVACT. Conversely, training on CVACT causes the model to learn intricate, scene-specific cues that are less transferable to CVUSA's simpler context. Crucially, MRGeo's robustness ensures it still outperforms existing methods even when learning from complex data, underscoring its superior generalization capability.

Few-Shot Training. We evaluate MRGeo's learning capability via few-shot learning to test its performance with limited data. As shown in Fig. 3, MRGeo demonstrates exceptional efficiency. Using just 20% of the training data, our model achieves an R@1 of 95.2%. This result is not only significantly better than Sample4Geo using 80% of the data (R@1 91.5%) but also eclipses FRGeo's performance with 60% of the data (R@1 95.18%). This highlights MRGeo's ability to learn robust representations from limited information, making it a highly efficient and practical solution for scenarios where data acquisition is costly or challenging.

4.5 Visualization of Qualitative Results

To better demonstrate the robustness of MRGeo to corrupted images, we visualize heatmaps on clean and corrupted images. Under "Fog" and "Zoom" corruption (Severity-5),

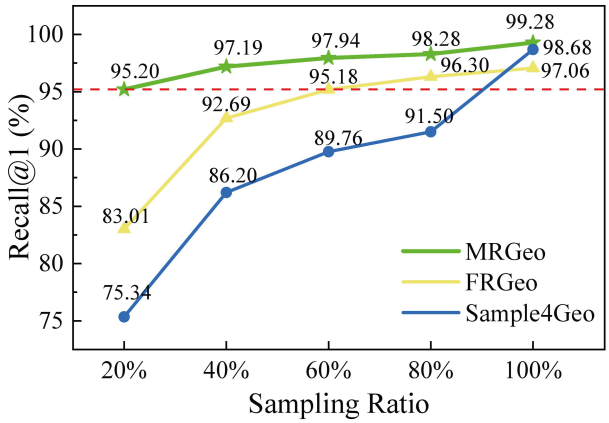

Figure 3: Few-shot training on CVUSA. Performance evaluation on the raw test set with progressively sampled training subsets (20%, 40%, 60%, 80%, 100%). The red dotted line is MRGeo's R@1 benchmark on 20% data.

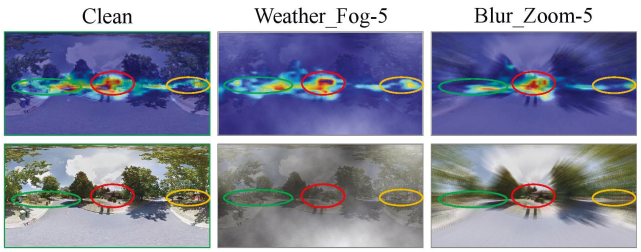

Figure 4: Heatmap visualization of MRGeo's feature focus on both clean and corrupted images. Best viewed on screen with zoom-in.

MRGeo still exhibits excellent feature extraction and analysis capabilities. Compared to focus areas on clean images, it fails only on some minor features affected by interference.

5 Conclusion and Limitations

This paper proposes **MRGeo**, the first systematic method to address the performance degradation of cross-view geo-localization models on corrupted images. By building a hierarchical defense—dynamically enhancing features at the spatial and channel level with our **SARM** and **CCM**, and enforcing structural consistency at the descriptor level with **RGAM**—MRGeo not only sets a new SOTA on robustness benchmarks but also exhibits superior generalization ability in cross-area evaluation and high sample efficiency in few-shot training. These results validate our core thesis that true robustness stems from a holistic, multi-level enhancement strategy. However, our method's performance degrades under extremely severe corruptions, as unstructured noise can violate modules' assumptions and RGAM's effectiveness relies on a strict center-alignment. Addressing these challenges remains an important direction for future research.

Acknowledgments

This work was supported in part by Shenzhen Science and Technology Program under Grant JCYJ20240813142510014 and Grant 20220810142553001, in part by the Key Project of Department of Education of Guangdong Province under Grant 2023ZDZX1016.

References

Chen, J.; Kao, S.-h.; He, H.; Zhuo, W.; Wen, S.; Lee, C.-H.; and Chan, S.-H. G. 2023. Run, don't walk: chasing higher FLOPS for faster neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12021–12031.

Cui, Y.; and Knoll, A. 2023. Exploring the potential of channel interactions for image restoration. *Knowledge-Based Systems*, 282: 111156.

Deuser, F.; Habel, K.; and Oswald, N. 2023. Sample4geo: Hard negative sampling for cross-view geo-localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16847–16856.

Fu, C.; and Zhu, Y. 2025. BGHR: Bridging the Gap Between HBox-Supervised and RBox-Supervised Oriented Object Detection via Adaptive Fine-Grained Sample Mining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 3022–3030.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, 1735–1742. IEEE.

Häne, C.; Heng, L.; Lee, G. H.; Fraundorfer, F.; Furgale, P.; Sattler, T.; and Pollefeys, M. 2017. 3D visual perception for self-driving cars using a multi-camera system: Calibration, mapping, localization, and obstacle detection. *Image and Vision Computing*, 68: 14–27.

Hendrycks, D.; and Dietterich, T. 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. arXiv:1903.12261.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-efficient transfer learning for NLP. In *International conference on machine learning*, 2790–2799. PMLR.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Kim, D.-K.; and Walter, M. R. 2017. Satellite image-based localization via learned embeddings. In *2017 IEEE international conference on robotics and automation (ICRA)*, 2073–2080. IEEE.

Kingma, D. P. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lee, S.; Chang, M.; Park, S.; and Seo, J. 2024. Assessing Graphical Perception of Image Embedding Models using Channel Effectiveness. In *2024 IEEE Visualization and Visual Analytics (VIS)*, 226–230. IEEE.

Li, Y.; and Zhu, Y. 2025. PLGeo: A Patch-level Framework to Overcome Orientation Discrepancies in Cross-view Geo-localization. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 6057–6065.

Liu, L.; and Li, H. 2019. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5624–5633.

Lu, F.; Lan, X.; Zhang, L.; Jiang, D.; Wang, Y.; and Yuan, C. 2024. CricaVPR: Cross-image Correlation-aware Representation Learning for Visual Place Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16772–16782.

McManus, C.; Churchill, W.; Maddern, W.; Stewart, A. D.; and Newman, P. 2014. Shady dealings: Robust, long-term visual localisation using illumination invariance. In *2014 IEEE international conference on robotics and automation (ICRA)*, 901–906. IEEE.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Rao, Y.; Zhao, W.; Zhu, Z.; Lu, J.; and Zhou, J. 2021. Global filter networks for image classification. *Advances in neural information processing systems*, 34: 980–993.

Shi, Y.; Liu, L.; Yu, X.; and Li, H. 2019. Spatial-aware feature aggregation for image based cross-view geo-localization. *Advances in Neural Information Processing Systems*, 32.

Shi, Y.; Yang, L.; An, W.; Zhen, X.; and Wang, L. 2023. Parameter-free channel attention for image classification and super-resolution. *arXiv preprint arXiv:2303.11055*.

Touvron, H.; Cord, M.; Sablayrolles, A.; Synnaeve, G.; and Jégou, H. 2021. Going deeper with image transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 32–42.

Wang, Z.; Lu, J.; Tao, C.; Zhou, J.; and Tian, Q. 2019. Learning channel-wise interactions for binary convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 568–577.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Xia, P.; Yu, L.; Wan, Y.; Wu, Q.; Chen, P.; Zhong, L.; Yao, Y.; Wei, D.; Liu, X.; Ru, L.; et al. 2025. Cross-view geo-localization with panoramic street-view and VHR satellite imagery in decentrality settings. *ISPRS Journal of Photogrammetry and Remote Sensing*, 227: 1–11.

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

Ye, J.; Lv, Z.; Li, W.; Yu, J.; Yang, H.; Zhong, H.; and He, C. 2024. Cross-view image geo-localization with Panorama-BEV Co-Retrieval Network. In *European Conference on Computer Vision*, 74–90. Springer.

Yin, D.; Gontijo Lopes, R.; Shlens, J.; Cubuk, E. D.; and Gilmer, J. 2019. A fourier perspective on model robustness

in computer vision. *Advances in Neural Information Processing Systems*, 32.

Yun, S.; and Ro, Y. 2024. Shvit: Single-head vision transformer with memory efficient macro design. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5756–5767.

Zhai, M.; Bessinger, Z.; Workman, S.; and Jacobs, N. 2017. Predicting ground-level scene layout from aerial imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 867–875.

Zhang, Q.; and Zhu, Y. 2024. Benchmarking the Robustness of Cross-View Geo-Localization Models. In *European Conference on Computer Vision*, 36–53. Springer.

Zhang, X.; Li, X.; Sultani, W.; Zhou, Y.; and Wshah, S. 2023. Cross-view geo-localization via learning disentangled geometric layout correspondence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3480–3488.

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

Zhu, S.; Yang, T.; and Chen, C. 2021. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3640–3649.

Zhu, Y.; Yang, H.; Lu, Y.; and Huang, Q. 2023. Simple, effective and general: A new backbone for cross-view image geo-localization. *arXiv preprint arXiv:2302.01572*.