

Method	Normal		Fog		Rain		Snow		Fog+Rain		Fog+Snow		Rain+Snow		Dark		Over-exp		Wind		Mean	
	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP
Sample4Geo [1] [ICCV'23]	74.93	78.76	72.58	76.44	34.60	41.56	28.95	35.02	35.10	41.47	12.95	17.90	20.05	25.95	34.18	38.99	38.40	43.68	67.80	72.41	41.95	47.22
Safe-Net [2] [TIP'24]	45.75	52.59	12.52	19.01	17.00	22.04	8.47	12.29	4.65	8.31	2.83	5.39	12.58	16.75	21.42	26.52	22.62	29.11	69.93	73.56	21.78	26.56
CCR [3] [TCSVT'24]	73.22	74.53	70.95	73.14	60.14	64.95	50.31	53.12	45.87	49.14	45.80	47.87	31.25	32.94	31.03	34.36	59.97	61.07	52.02	53.33	52.06	53.46
MuSe-Net [4] [PR'24]	59.65	64.24	55.62	60.42	54.70	59.50	51.42	56.39	51.92	56.83	54.70	59.50	52.47	57.50	44.98	50.25	48.47	53.57	55.92	60.88	52.99	57.91
Ours	63.40	68.59	65.07	70.10	63.85	68.93	58.55	64.14	61.85	67.04	53.80	59.40	60.97	66.32	54.92	60.72	49.45	55.69	60.58	65.91	59.24 (+6.25)	64.68 (+6.77)
Satellite → Drone																						
Sample4Geo [1] [ICCV'23]	87.50	79.57	83.75	71.14	42.50	25.24	40.00	21.59	38.75	23.22	30.00	10.58	26.25	16.44	56.25	29.75	58.75	30.38	83.75	69.36	54.75	37.76
Safe-Net [2] [TIP'24]	52.50	42.55	53.75	28.04	50.00	29.18	35.00	13.55	40.00	13.76	13.75	4.04	43.75	15.64	47.50	20.91	36.25	24.08	50.00	36.50	42.25	22.83
CCR [3] [TCSVT'24]	90.59	80.45	82.99	70.62	43.39	45.90	39.81	40.88	42.63	39.46	29.32	30.65	25.89	26.94	26.01	30.40	58.01	59.13	83.09	61.05	52.17	48.55
MuSe-Net [4] [PR'24]	67.50	53.24	60.00	48.71	56.25	42.52	56.25	45.71	57.50	34.15	53.75	34.15	53.75	45.15	52.50	35.96	58.75	41.25	62.50	47.51	58.50	44.40
Ours	77.50	67.03	75.00	63.87	71.25	61.62	67.50	57.03	75.00	59.71	68.75	49.98	63.75	57.94	76.25	54.26	67.50	52.77	75.00	61.25	71.75 (+13.25)	58.55 (+14.15)

Table 1: **Cross-dataset retrieval performance.** Models are trained on University-1652 and evaluated on SUES-200. * denotes the use of official pretrained on University-1652 weights. Best results are highlighted in bold.

Method	Dark+Rain		Fog+Rain+Snow		Winter Night		Freezing Rain		Blizzard	
	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP
Drone → Satellite										
Sample4Geo [1] [ICCV'23]	1.89	3.02	0.94	1.79	25.40	28.60	37.83	41.95	36.00	40.42
Safe-Net [2] [TIP'24]	1.82	3.05	0.15	0.47	3.25	4.87	2.64	4.30	11.46	15.43
CCR [3] [TCSVT'24]	14.56	18.33	0.04	0.41	36.05	38.79	47.29	51.70	25.73	30.07
MuSe-Net [4] [PR'24]	13.18	17.37	29.93	36.30	39.45	44.36	44.09	49.01	33.44	38.51
Ours	36.42	19.45	70.87	74.57	58.67	63.08	61.52	65.78	59.76	64.15
Satellite → Drone										
Sample4Geo [1] [ICCV'23]	6.70	1.16	3.00	0.48	58.35	21.33	65.62	32.76	66.90	32.47
Safe-Net [2] [TIP'24]	30.67	9.17	4.14	0.63	56.92	8.28	41.08	8.15	57.92	22.05
CCR [3] [TCSVT'24]	57.90	32.95	10.84	2.26	69.97	44.08	66.16	55.20	60.88	34.23
MuSe-Net [4] [PR'24]	28.75	13.41	56.25	31.48	73.32	40.19	76.03	43.51	69.76	35.21
Ours	74.61	40.65	86.02	71.43	82.60	59.43	85.31	61.17	71.90	41.94

Table 2: **Quantitative evaluation study of extreme weathers on the University-1652 dataset.** * denotes the use of official pretrained on University-1652 weights. Best results are highlighted in bold.

1 A Additional Experimental Results

2 A.1 Quantitative Experiments

3 Cross-City Transfer Results

4 The model is trained on University-1652 [5] images and evaluated on SUES-200 [6]. All experimental
5 configurations follow those described in the main paper. For all methods, we use the official pretrained
6 weights released by the original authors, without any additional retraining or fine-tuning. The SUES-
7 200 test set is constructed by randomly sampling and combining images from four different altitude
8 levels to ensure comprehensive evaluation across varying viewpoints.

9 As shown in Table 1, the proposed method achieves the highest mean R@1 and AP in both
10 Drone → Satellite and Satellite → Drone tasks. In the Drone → Satellite task, our method attains a
11 mean R@1 of 59.24% outperforming the previous state-of-the-art method MuSe-Net [4] (52.99%),
12 with a +6.25% gain, and achieves an improvement of +6.77% in mean AP. For the Satellite → Drone
13 task, the proposed method attains a mean R@1 of 71.75%, outperforming MuSe-Net (58.05%) by
14 13.25% and achieving a 14.15% improvement in AP. Furthermore, previous methods, such as Sam-
15 ple4Geo [1] and Safe-Net [2], exhibit satisfactory R@1 performance under clear or mildly perturbed
16 conditions (e.g. Normal, Wind), but validate limited robustness under complex weather scenarios such
17 as fog+rain, rain+snow, and dark. These results verify that the proposed method maintains stability
18 and generalization in diverse weather conditions and challenging scenarios in the cross domain.

19 Generalization to Unseen Weather

20 To further evaluate the robustness of each method in out-of-distribution scenarios, we conduct
21 experiments on previously unseen extreme weather conditions, specifically, "Dark + Rain", "Fog
22 + Rain + Snow", "Winter Night", "Freezing Rain" and "Blizzard". The results are reported in
23 Table 2. For the Drone → Satellite task, existing approaches remain sensitive to extreme weather
24 and show limited cross-weather robustness. Using MuSe-Net as a representative baseline, the cross-
25 weather macro averages are 32.02% R@1 and 37.11% AP, whereas our method achieves 57.45%
26 and 57.41%, yielding gains of +25.43% and +20.30%. The margins are particularly evident under
27 Fog+Rain+Snow and Blizzard, where our method attains 70.87%/74.57% and 59.76%/64.15%,
28 while the corresponding baseline results are 29.93%/36.30% and 36.00%/40.42%, respectively. For
29 the Satellite → Drone task, our method achieves the top R@1 and AP across all weather settings, with
30 cross-weather macro averages of 80.09% R@1 and 54.92% AP. The corresponding prior averages
31 are 60.82% R@1 for MuSe-Net and 33.74% AP for CCR, indicating improvements of +19.27% and
32 +21.18%. These results validate that the proposed semantic-guided method is effective in generalizing
33 to complex, unseen weather distributions, maintaining high retrieval accuracy even when all other
34 methods experience substantial drops. We attribute this improvement to the explicit integration of

Method	D2S		S2D	
	Mean R@1(%)	Mean AP(%)	Mean R@1(%)	Mean AP(%)
GLM	76.52	79.85	86.62	73.65
Ours	77.14	80.20	87.72	76.39

Table 3: **Difference Description Generator Results on the University-1652 Dataset.** Best results are highlighted in bold.

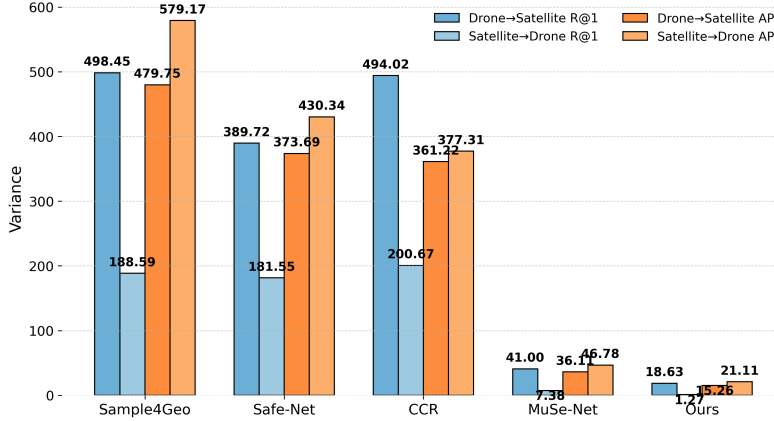


Figure 1: **Variance comparison of R@1 and AP across different models for cross-view localization on the University-1652 dataset.**

weather and scene information, which enables the model to extract robust and discriminative features invariant to challenging visual perturbations.

Impact of the Captioning LVL. To assess independence from a particular LVL, we replaced the caption generator with GLM-4.1V-9B-Thinking [7], retained the same prompts and structured output schema, regenerated captions, retrained from scratch, and re-evaluated. R@1 and mAP remained consistent with the original configuration (see Table 3), indicating that the gains primarily stem from the normalized caption schema and training objectives rather than a specific LVL.

Variance-based Analysis of Weather Robustness

Figure 1 presents the variance of the retrieval accuracy (R@1) and the average precision (AP) under different weather conditions for both Drone→Satellite and Satellite→Drone tasks. As illustrated, previous methods exhibit considerable fluctuations in variance, with Sample4Geo and CCR showing R@1 variances as high as 498.45 and 494.02, respectively, for the Drone→Satellite direction. Even the most stable competitive method (MuSe-Net) yields an R@1 variance of 41.00 in this task.

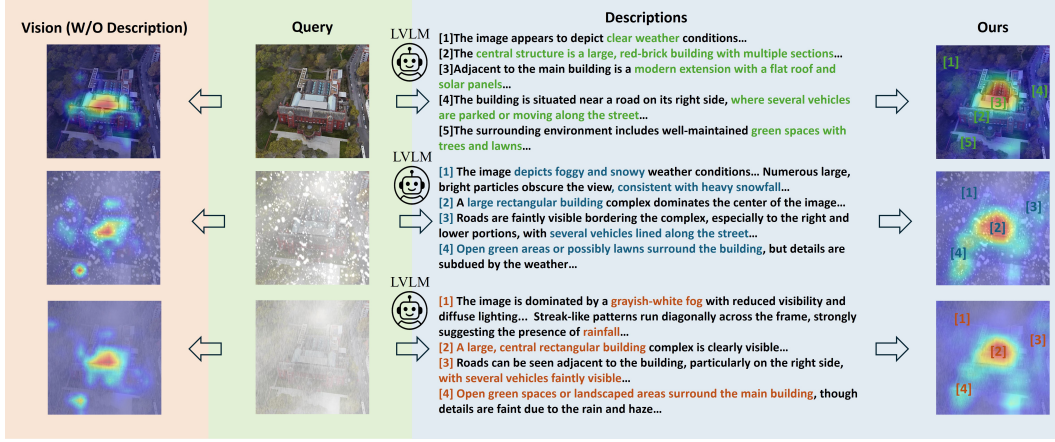
In contrast, the proposed method achieves substantially lower variance values 18.63 for Drone→Satellite R@1 and only 1.27 for Satellite→Drone R@1. A similar trend holds for AP, where our method records the lowest variance among all methods (15.26 and 21.11, respectively), while previous methods typically exceed 300 or even 500. This significant reduction in variance validates that our method maintains consistent retrieval performance under a wide range of weather conditions, effectively mitigating the instability and performance drops observed in existing methods.

A.2 Qualitative Result

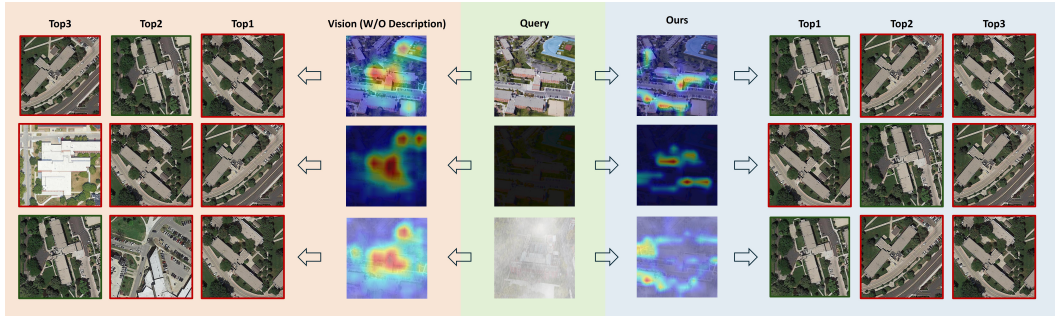
Grad-CAM Visualizations

Understanding the regions a model attends to during cross-view retrieval is crucial for diagnosing whether it effectively captures discriminative, weather-invariant cues or merely depends on superficial visual features. Therefore, we utilize Grad-CAM [8] to visualize and contrast the attention maps produced by a purely visual encoder and our proposed text-guided model.

Figure 2a presents a comparison between the attention maps generated by the vision-only baseline (red block) and our method (blue block) under varying weather conditions (normal, fog+snow,



(a) Impact of Text-guided Descriptions on Attention Allocation.



(b) Impact of Semantic-guided Attention on Cross-view Retrieval Results.

Figure 2: (a) Compares the attention distributions of the vision-only encoder and the text-guided model, highlighting the role of semantic descriptions in directing the model’s focus. With semantic guidance, the model attends to semantically meaningful and weather-resilient key regions (as indicated by the numbered annotations), whereas the vision-only method tends to concentrate on visually prominent but less stable areas. (b) Validates the improvement in retrieval performance brought by semantic-guided attention. Under various challenging weather conditions, the text-guided model consistently retrieves the correct results, while the vision-only baseline is more prone to errors. The green boxes indicate the correct matches, and the images in the red boxes are the wrong matches.

62 fog+rain). The vision-only baseline model typically focuses on visually salient regions, often
 63 missing contextually meaningful cues that are crucial for localization. By incorporating weather and
 64 scene-aware textual descriptions, our model dynamically re-allocates attention to cover semantically
 65 relevant structures (*e.g.* building outlines, vehicle clusters, or vegetation), even under fog occlusion or
 66 degradation. The numbers in the attention maps correspond to key semantic elements extracted from
 67 language descriptions, providing direct evidence of semantic alignment between vision and text.

68 In Figure 2b, we examine the impact of semantic guidance on retrieval under complex weather
 69 conditions by directly comparing the attention maps and top-3 retrieval results. Under normal weather,
 70 the vision-only baseline predominantly attends to regions with high local contrast or color intensity
 71 (*e.g.* buildings or reflective surfaces), which are visually salient but not uniquely informative for
 72 localization. As a result, the vision-only baseline sometimes retrieves areas with similar appearance
 73 but different spatial context. When text guidance is incorporated, the attention of our model shifts to
 74 stable, semantically grounded structures (*e.g.* the outlines of buildings and road intersections). In fog
 75 and rain scenes, the attention of vision-only baseline further drifts toward blurred or occluded regions,
 76 reducing retrieval reliability. In contrast, the text-guided model continues to highlight robust elements
 77 (*e.g.* the unique L-shaped building and adjacent open space in row 2), which are specifically mentioned
 78 or implied in the text description. This allows the model to maintain precise matching even when the
 79 visual signal is weak. At night, while all models face significant difficulty, the text-guided method
 80 still prioritizes the locations of the major structures, likely due to cues (*e.g.* "adjacent to the large

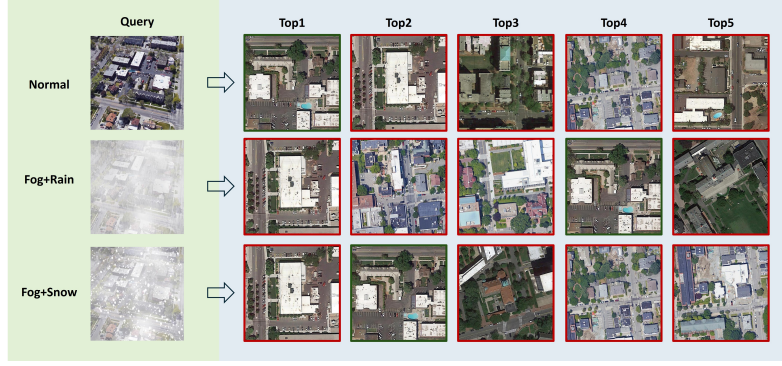


Figure 3: **Retrieval Performance and Typical Errors Under Adverse Weather.** The green boxes indicate the correct matches, and the images in the red boxes are the wrong matches.

square"), allowing more competitive retrieval performance. These examples validate that, without textual guidance, visual attention is easily misled by non-distinctive features or transient artifacts. In contrast, semantic prompts help anchor the model's focus on task-relevant, weather-resilient regions, providing both improved robustness and interpretability.

Typical Failure Cases and Analysis

Figure 3 presents retrieval results for the same location under normal, fog+rain, and fog+snow conditions. Under normal weather, our model accurately identifies the ground-truth satellite image as the top-1 match. Under more adverse conditions (*e.g.* Fog+Rain and Fog+Snow) the visual degradation results in some imperfection. The correct target appears as top-4 and top-2, respectively.

Notably, even under these challenging conditions, our method validates significantly enhanced robustness compared to vision-only methods, which typically fail to retrieve the correct location within the top-5. Careful inspection shows that the false positive results most often correspond to areas with highly similar building layouts and rooftop features. This is largely due to the extreme reduction in visibility and the presence of severe occlusions and specular highlights, which obscure the fine-grained, discriminative cues necessary for identification. Our semantic-guided retrieval framework helps alleviate this problem by leveraging both textual and weather-aware context, enabling the model to adapt its attention towards stable, meaningful structures and maintain reliable performance even when detailed appearance cues are compromised. However, in densely built urban environments, where multiple regions may share similar global spatial patterns, there remains a risk of confusion under severe weather, as the model must sometimes rely on coarse-grained features.

B Societal Impact and Risks

The drone-based cross-view geo-localization method introduced in this work has potential benefits in several practical, real-world applications. For instance, accurate and robust drone localization under challenging weather conditions can greatly facilitate rapid emergency responses, enabling efficient mapping and evaluation of affected regions during natural disasters. Moreover, such localization methods may enhance routine urban management tasks, including traffic monitoring and incident detection, thereby contributing to improved public safety and optimized city planning.

Despite these practical benefits, deploying precise drone localization systems poses certain societal risks, primarily related to data privacy and responsible usage. High-resolution aerial and satellite images, necessary for effective geo-localization, might inadvertently capture sensitive information, potentially compromising individual privacy. Without careful management, there is also a possibility of misuse in unauthorized surveillance or tracking scenarios. Thus, ensuring responsible deployment requires clear operational guidelines, appropriate anonymization techniques, and transparent data-handling practices to balance technological benefits with privacy considerations.

C Evaluation protocol

We report two standard evaluation metrics widely used in image retrieval tasks: Recall@K (R@K) and Average Precision (AP). Specifically, Recall@K measures the proportion of queries whose corresponding ground-truth image is ranked within the top-K retrieved candidates, reflecting the retrieval accuracy at different ranks. Average Precision (AP) summarizes the precision-recall curve by computing the mean precision across all relevant retrieval positions, providing a comprehensive evaluation of the ranking quality.

References

- [1] Fabian Deuser, Konrad Habel, and Norbert Oswald. “Sample4Geo: Hard Negative Sampling for Cross-View Geo-Localization”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023, pp. 16847–16856.
- [2] Jinliang Lin et al. “A Self-Adaptive Feature Extraction Method for Aerial-View Geo-Localization”. In: *IEEE Transactions on Image Processing* (2024).
- [3] Haolin Du, Jingfei He, and Yuanqing Zhao. “CCR: A Counterfactual Causal Reasoning-Based Method for Cross-View Geo-Localization”. In: *IEEE Transactions on Circuits and Systems for Video Technology* (2024).
- [4] Tingyu Wang et al. “Multiple-environment Self-adaptive Network for Aerial-view Geo-localization”. In: *Pattern Recognition* 152 (2024), p. 110363. DOI: 10.1016/j.patcog.2024.110363.
- [5] Zhedong Zheng, Yunchao Wei, and Yi Yang. “University-1652: A Multi-view Multi-source Benchmark for Drone-based Geo-localization”. In: *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)*. 2020, pp. 1395–1403.
- [6] Runzhe Zhu et al. “SUES-200: A Multi-height Multi-scene Cross-view Image Benchmark Across Drone and Satellite”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 33.9 (2023), pp. 4825–4839.
- [7] V Team et al. *GLM-4.5V and GLM-4.1V-Thinking: Towards Versatile Multimodal Reasoning with Scalable Reinforcement Learning*. 2025. arXiv: 2507.01006 [cs.CV].
- [8] Ramprasaath R. Selvaraju et al. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 618–626.