See What We Cannot See: A Geo-guided Reasoning Benchmark for Object Counting under Adverse Earth Observation Conditions

Supplementary Material

Figure 1. Examples of annotations in the GROC dataset. Different categories of objects are labeled with center points of different colors (airplane: blue; truck: orange; boat: green; car: red).

A. Additional Details of the Geospatial Data Collector

We provide additional implementation details of the geospatial data collector used to build GROC. The collector sources remote sensing imagery and aligned geo-data from public geospatial repositories, including PDOK [5] for RGB, near-infrared (NIR), digital surface model (DSM), and map layers, and an official land-use product [4] for land-use information. These original data sources are distributed under the Creative Commons Attribution Share-Alike license (CC BY-SA 4.0), which permits editing and redistribution under proper attribution and share-alike conditions.

Different from conventional counting datasets that often discard geo-coordinates after image cropping or retain only visual imagery without the underlying geospatial context, our collection pipeline preserves the full geospatial linkage. Starting from scene queries such as "parking lot", "airport", and "port", we localize candidate regions of interest and retrieve the corresponding geo-referenced tiles. Each tile is then cropped into 1024×1024 patches while retaining the exact geo-coordinates of every patch, which enables future enrichment from external geodatabases.

RGB and NIR imagery are acquired simultaneously, while DSM, land-use, and map layers are aligned within the same year and co-registered to a common coordinate frame, so that each pixel across the multimodal stack corresponds to the same physical location on the ground. To balance usability and fidelity, each patch is exported both as a lightweight PNG for visualization and as the original TIFF for preserving radiometric richness and geospatial metadata. The resulting clear and spatially aligned RGB/NIR observations further serve as the input to the degradation generator for synthesizing adverse observation conditions in a controlled manner.

B. Dataset Annotation Visualization

To support large-scale dense object counting, GROC provides point-level annotations for four object categories: airplane, truck, boat, and car. As shown in Fig. 1, each instance is labeled by a single center point with a category-specific color. The dataset spans diverse geographic environments

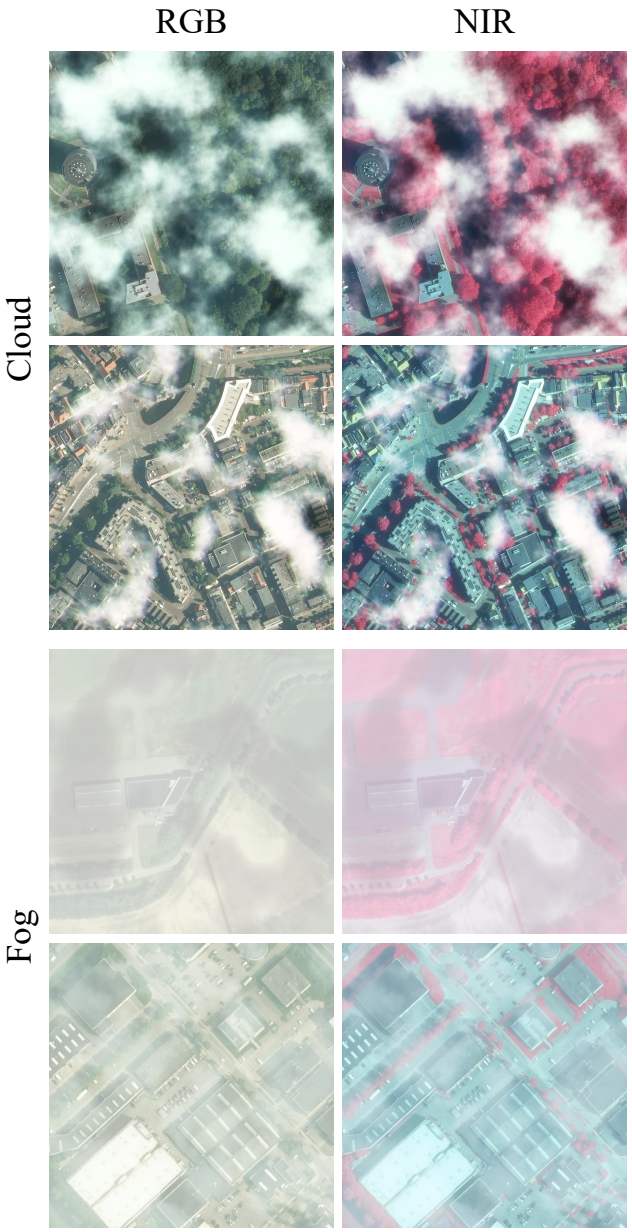

Figure 2. Synthetic Adverse Weather.

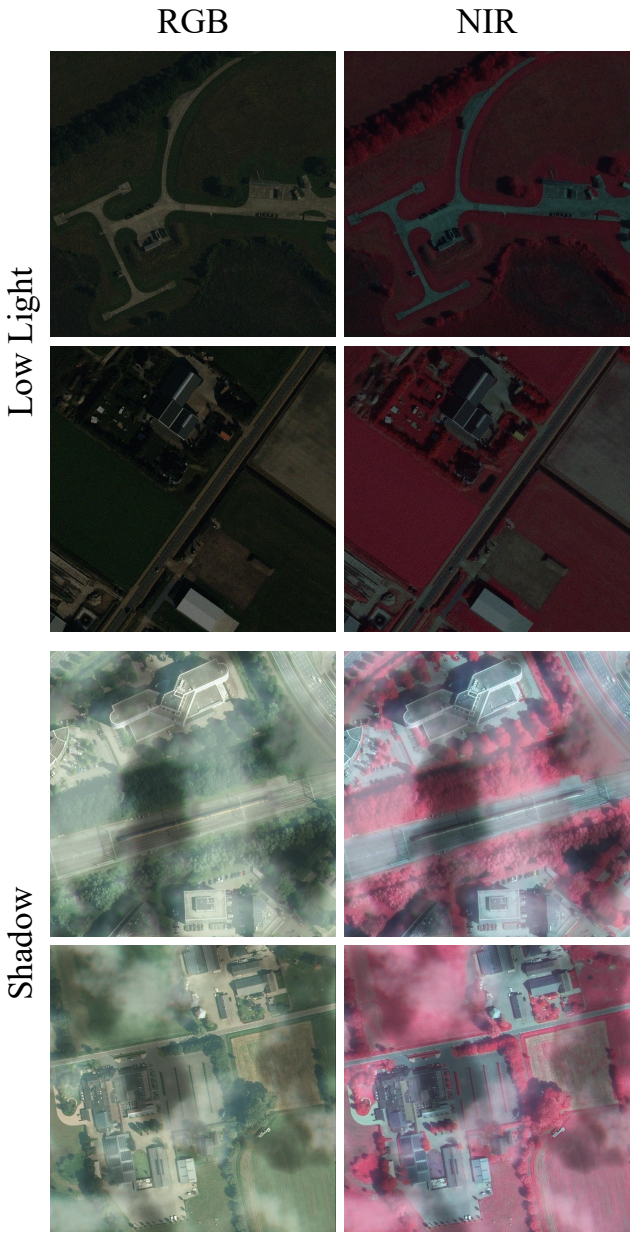

Figure 3. Synthetic Adverse Illumination.

and object configurations, making annotation and counting highly challenging. Objects vary greatly in scale, from large aircraft to small vehicles, and appear over complex backgrounds such as airports, ports, highways, industrial areas, and urban regions. GROC also exhibits a clear long-tail distribution, with dense and frequent categories such as cars coexisting with sparse and isolated categories such as airplanes. In addition, many scenes contain crowded layouts with small inter-object distances and frequent occlusions, making precise center-point annotation difficult. These factors together highlight the difficulty of building a reliable large-scale benchmark for object counting under realistic adverse earth-observation conditions.

C. More Details about Controllable Degradation Generator

To realistically simulate adverse conditions, we design a physics-guided degradation module that couples a cloud radiative model [6] with controllable procedural synthesis [1]. Instead of treating clouds as purely visual overlays, we adopt a simplified radiative transfer formulation in which the observed radiance is modeled as a combination of background transmittance and cloud scattering. For each spec-

tral band i, the degraded signal satisfies

$$I_i^{\text{obs}} = \tau_i B_i + \rho_i, \tag{1}$$

$$\tau_i + \rho_i + \alpha_i = 1, \tag{2}$$

where B_i is the clear-sky radiance, τ_i is the two-way transmittance, ρ_i is the cloud-scattered radiance, and α_i denotes absorption. In our implementation, a procedural cloud mask M approximates $1 - \tau$, while a cloud-color field C represents ρ. The final degraded image is obtained using the cloud compositing equation

$$I^{\text{obs}} = (1 - M)I^{\text{clear}} + MC. \tag{3}$$

Cloud fields are generated via multi-scale Perlin turbulence [9], with controls on locality, opacity range, and thickness to produce thin, broken, or dense structures. We additionally synthesize cloud shadows and small per-channel spatial shifts to mimic sensor parallax. The same cloud and shadow masks are applied to both RGB and CIR images to preserve multimodal occlusion consistency, which is important for geo-guided reasoning tasks.

For low-light scenes, we emulate the imaging pipeline by applying exposure reduction, gamma darkening, and signal-dependent noise with variance

$$\sigma^2(x) = ax + b, \tag{4}$$

which approximates typical nighttime sensor behavior.

Together, these physically grounded operators yield consistent, controllable, and cross-modality-aligned degradations that approximate real remote sensing conditions and provide a more principled alternative to purely stylistic augmentations.

D. Synthetic Degradation Visualization

We provide additional examples of the synthetic degradations to demonstrate their realism. As shown in Fig. 2 and Fig. 3, our synthetic degradations preserve the underlying geospatial layout while introducing realistic impairments, including cloud occlusion, fog, shadows, and low-light effects.

We use the same degradation synthesis for red-green-blue (RGB) and near-infrared (NIR) false-color images because the NIR imagery used here is still reflective rather than thermal. Similar to visible bands, it depends on solar illumination, remains strongly affected by cloud occlusion and atmospheric scattering, and cannot provide heat-distribution cues under low-light conditions as thermal infrared (TIR) can. As a result, its degradation mechanism is highly similar to that of RGB, which justifies the use of the same synthesis strategy across the two modalities. This is also consistent with real degraded observations in NWPU-MOC [3].

Quantitative evaluation with Qwen3-VL-8B as an independent scorer further supports the realism of our synthetic degradations. Real degradations receive slightly higher realism scores than synthetic ones (0.982 vs. 0.948), but the gap remains small (AUC=0.63, $p < 10^{-2}$). Additional qualitative and quantitative evidence is provided in the appendix.

E. Additional Comparison with MLLM-Only, Tool-Only, and Agent Baselines

To further disentangle the contributions of the *MLLM backbone*, *tool access*, and *explicit agentic reasoning*, we compare three types of baselines on the degraded GROC test split, as shown in Table 1. First, we evaluate an *MLLM-only* baseline, where Qwen3-VL-8B-Instruct directly predicts object counts without calling any external tools. Second, we consider several *tool-only* baselines that use the same external experts as GROC Agent, including *Average*, which directly averages all tool outputs, *Median*, which is more robust to outlier predictions, and *Manual Workflow*, which mimics a hand-crafted coordination strategy without agentic reasoning. Finally, we report *GROC Agent*, which uses the same Qwen3-VL-8B-Instruct backbone but further incorporates geo-modalities and adaptive tool use.

Table 1. Comparison of MLLM-only, tool-only, and agent-based baselines on the degraded GROC test split. The MLLM-only baseline and GROC Agent both use Qwen3-VL-8B-Instruct [11] as the backbone. All tool-based methods use the same external experts (DINO-X [10], PSGCNet [2], BL [8], and FIDTM [7]).

Method	Overall		Weather		Low-light	
	MAE	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**
Backbone: Qwen3-VL-8B-Instruct						
MLLM only	40.84	69.44	41.43	70.32	40.21	68.50
Tools: DINO-X, PSGCNet, BL, FIDTM						
Average	20.77	39.24	19.28	37.04	22.36	41.45
Median	19.49	37.52	17.63	34.35	21.47	40.63
Manual Workflow	17.37	35.71	**16.73**	**32.80**	18.06	38.58
GROC Agent	**16.58**	**34.22**	17.75	34.96	**15.34**	**33.40**

Several observations can be made. First, the MLLM-only baseline performs substantially worse than all tool-based variants, indicating that the backbone alone struggles to produce reliable count estimates under degraded conditions. This suggests that direct multimodal reasoning without external experts remains insufficient for this benchmark.

Second, simple tool aggregation already yields a large improvement over MLLM-only prediction, confirming that expert models provide strong complementary evidence once appearance cues are impaired. Among the tool-only baselines, *Median* consistently outperforms *Average*, suggesting that expert predictions can vary considerably across challenging scenes and that robust aggregation is beneficial. *Manual Workflow* further improves performance, indicating

that even hand-crafted coordination among tools is more effective than static ensembling.

Finally, GROC Agent achieves the best overall and low-light performance, while remaining competitive under adverse weather. Since it uses the same Qwen3-VL-8B-Instruct backbone as the MLLM-only baseline and the same external tools as the tool-only baselines, its gains cannot be attributed solely to either the backbone or tool availability. Instead, the results suggest that adaptive tool selection and reasoning over multimodal evidence provide additional benefits beyond fixed aggregation rules. At the same time, the relatively small gap between GROC Agent and the best tool-only baseline under weather degradation also indicates that stronger geo-guided reasoning strategies are still needed.

F. Potential Applications

GROC is motivated by all-weather remote sensing scenarios where direct visual evidence becomes unreliable due to cloud occlusion, fog, shadows, or low-light conditions, yet downstream users still require timely object-level situational awareness. Rather than replacing high-confidence recognition in clear imagery, GROC focuses on counting under partial observability by leveraging stable geo-modalities and environmental context. This capability is useful for approximate vehicle activity estimation in parking areas, airport and port monitoring, and broad-area screening of transportation infrastructure when imagery quality is degraded. In such settings, count estimates can support regional monitoring and anomaly screening, especially when precise localization is infeasible. Beyond benchmarking, GROC also provides a testbed for multimodal reasoning under incomplete observations, uncertainty-aware counting, and agentic integration of foundation models with geospatial data.

G. Limitations and Future Work

Despite its utility, GROC still has several limitations. First, the current benchmark mainly relies on synthetically generated degradations for controlled evaluation. While this design enables paired clear/degraded analysis, it cannot fully capture the distribution shift introduced by real atmospheric effects or sensor artifacts. Second, the current release covers only four object categories, which limits semantic diversity. Third, although the annotation pipeline incorporates detector-assisted human verification, the current paper does not yet provide a systematic quantitative analysis of multi-round annotation effectiveness or cost-benefit. Fourth, comprehensive cross-dataset validation against existing counting benchmarks is still limited in the current version. Finally, while GROC Agent offers an interpretable baseline for geo-guided reasoning, it remains a preliminary benchmark rather than a definitive solution. Future work will therefore focus on expanding the dataset to more object categories and geographic regions, constructing real degraded evaluation sets with verified annotations, strengthening cross-dataset transfer evaluation, quantifying the effectiveness of multi-round verification, and developing more capable geo-guided reasoning baselines with stronger uncertainty modeling under severe visual impairment.

References

[1] Mikolaj Czerkawski, Robert Atkinson, Craig Michie, and Christos Tachtatzis. SatelliteCloudGenerator: Controllable cloud and shadow synthesis for multi-spectral optical satellite images. *Remote Sensing*, 15(17), 2023. Art. no. 4138. 2

[2] Guangshuai Gao, Qingjie Liu, Zhenghui Hu, Lu Li, Qi Wen, and Yunhong Wang. PSGCNet: A pyramidal scale and global context guided network for dense object counting in remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12, 2022. Art. no. 3153946. 3

[3] Junyu Gao, Liangliang Zhao, and Xuelong Li. NWPU-MOC: A benchmark for fine-grained multi-category object counting in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024. Art. no. 5606614. 3

[4] Gerard Hazeu, Jan Droesen, Daphne Thomas, Marian Vittek, and Igor Staritsky. Landelijk Grondgebruik Nederland 2023 (LGN2023), 2025. Dataset. 1

[5] Dorus Kruse. Public service on the map: A case history of PDOK; the public geodata portal in the netherlands! In *Proceedings of the 16th European Conference on e-Government (ECEG)*, page 335, 2016. 1

[6] Jun Li, Zhaocong Wu, Zhongwen Hu, Jiaqi Zhang, Mingliang Li, Lu Mo, and Matthieu Molinier. Thin cloud removal in optical remote sensing images based on generative adversarial networks and physical model of cloud distortion. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166:373–389, 2020. 2

[7] Dingkang Liang, Wei Xu, Yingying Zhu, and Yu Zhou. Focal inverse distance transform maps for crowd localization. *IEEE Transactions on Multimedia*, 25:6040–6052, 2023. 3

[8] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6141–6150, 2019. 3

[9] Ken Perlin. An image synthesizer. *ACM SIGGRAPH Computer Graphics*, 19(3):287–296, 1985. 3

[10] Tianhe Ren, Yihao Chen, Qing Jiang, Zhaoyang Zeng, Yuda Xiong, Wenlong Liu, Zhengyu Ma, Junyi Shen, Yuan Gao, Xiaoke Jiang, Xingyu Chen, Zhuheng Song, Yuhong Zhang, Hongjie Huang, Han Gao, Shilong Liu, Hao Zhang, Feng Li, Kent Yu, and Lei Zhang. DINO-X: A unified vision model for open-world object detection and understanding. *arXiv preprint arXiv:2411.14347*, 2024. 3

[11] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen

Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 3