TEST-TIME DENOISING FOR NOISY CROSS-VIEW GEO-LOCALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-view geo-localization (CVGL) aims to localize a ground-level query by retrieving its corresponding image from a geo-tagged satellite gallery. Although recent methods achieve strong performance on clean benchmarks, their accuracy can degrade when query images are affected by sensor, acquisition, or transmission noise. Existing noise-robustness solutions typically rely on training strategies or specialized model architectures, which require modifying or retraining the downstream localization system. In this work, we adopt a training-free and plug-and-play strategy from a shared denoising perspective. We distinguish noise-dependent observation statistics from the shared underlying clean-image target and prepend a frozen pretrained denoiser to existing CVGL models while keeping the localization model unchanged. The same pretrained denoiser is applied across heterogeneous query-side noise without noise-specific adaptation. Extensive experiments across multiple CVGL architectures, noise types, and severity levels show that pretrained denoising can improve localization robustness, with the gains becoming more pronounced under stronger noise. These results demonstrate the potential of pretrained denoising as a simple test-time strategy for improving the noise robustness of frozen CVGL models without retraining or modifying the localization network.

1 INTRODUCTION

Cross-view geo-localization (CVGL) aims to determine the geographic location of a ground-level query image by retrieving its corresponding match from a geo-tagged satellite image gallery (Bansal et al., 2011; Hu et al., 2022; Li et al., 2024; Toker et al., 2021). Recent methods have achieved remarkable retrieval performance by learning discriminative cross-view representations from large-scale datasets (Shi et al., 2019; 2020a;b; Yang et al., 2021; Liu & Li, 2019; Zhang et al., 2024; Deuser et al., 2023). These advances make image-based geo-localization a promising complement to conventional positioning systems in applications such as autonomous drivingHäne et al. (2017) and robotic navigationMcManus et al. (2014), particularly in GPS-denied environments.

Despite this progress, existing CVGL methods are predominantly developed and evaluated under the assumption that test queries have visual quality comparable to the clean images used during training. In practical deployments, however, ground-view queries can be affected by sensor noise, photon-limited acquisition, transmission errors, and other acquisition-related factors. Such noise can distort local appearance and structural cues important for cross-view matching, introducing a query-side distribution shift that can degrade localization performance. As illustrated in Fig. 1(a), noisy queries are directly used for cross-view geo-localization in the conventional setting.

Existing approaches to improving robustness to noisy inputs commonly rely on noise-aware training, data augmentation, or specialized localization architectures, typically requiring modification or retraining of the localization model. We instead consider noisy CVGL from a shared denoising perspective. Although Gaussian, shot, speckle, and impulse noise arise from different stochastic processes (Foi et al., 2008; Goodman, 1976; Wang & Zhang, 1999), they can all be viewed as perturbations of the same underlying clean query. Their noise-dependent observation statistics therefore differ, while the desired denoising target remains the same latent clean image. This motivates us to ask whether a single pretrained denoiser can provide a shared clean-image prior across heterogeneous query-side noise.

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.

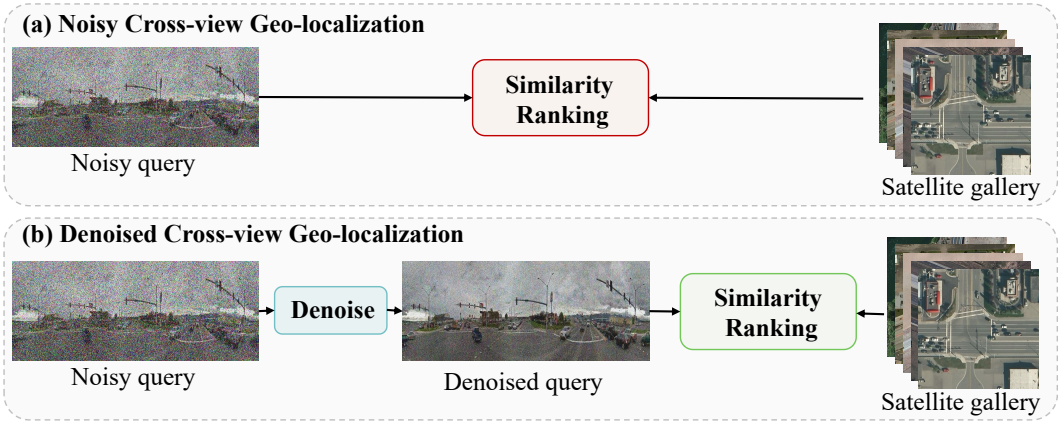


Figure 1: Overview of the standard noisy CVGL setting and our denoising-enhanced strategy. (a) A noisy ground-view query is directly matched against the satellite gallery, making retrieval vulnerable to query-side noise. (b) A frozen pretrained denoiser is inserted before cross-view matching, while the original CVGL model and satellite gallery remain unchanged.

Based on this perspective, we adopt a simple training-free denoising strategy by prepending a frozen pretrained denoiser to a frozen CVGL model, as illustrated in Fig. 1(b). The same denoiser is applied across different noise types and CVGL architectures without noise-specific adaptation or retraining of the localization network. Our experiments show that an appropriate pretrained denoiser can provide substantial overall robustness gains across heterogeneous query-side noise, with the benefit becoming more pronounced as noise severity increases. These results indicate that a denoising prior learned independently of the localization task can transfer beyond its nominal training noise and serve as a practical test-time strategy for noisy CVGL.

We conduct systematic experiments across multiple CVGL architectures, noise types, and severity levels. Our contributions are threefold. First, we introduce a shared denoising perspective for heterogeneous query-side noise in cross-view geo-localization, treating different noise processes through their common underlying clean-image target. Second, we adopt a simple training-free and plug-and-play denoising strategy in which a single pretrained denoiser is directly applied to frozen CVGL models across different noise types, without noise-specific adaptation or retraining of the localization model. Third, extensive cross-noise and cross-architecture evaluations demonstrate substantial overall robustness gains and show that the benefit of denoising becomes increasingly pronounced as noise severity increases.

2 Related Work

2.1 Cross-view Geo-localization

Early deep methods mainly improve geometric alignment and feature aggregation. SAFA (Shi et al., 2019) introduces spatial-aware feature aggregation to capture position-sensitive correspondences, while CVFT (Shi et al., 2020b) and DSM (Shi et al., 2020a) further exploit cross-view feature transformation and spatial matching. L2LTR (Yang et al., 2021) adopts Transformer-based representation learning to capture long-range dependencies across views. TransGeo (Zhu et al., 2022) further demonstrates the effectiveness of pure Transformer architectures for CVGL, while GeoDTR+ (Zhang et al., 2024) explicitly models geometric layout information to improve cross-view correspondence. More recently, Sample4Geo (Deuser et al., 2023) achieves strong performance using a simple dual-encoder architecture together with contrastive learning and effective negative sampling. Despite these advances, existing CVGL methods are primarily designed for clean-query settings and do not explicitly account for query-side noise, leading to substantial performance degradation under corrupted deployment queries.

2.2 Noisy Cross-View Geo-Localization

Zhang & Zhu (2024) introduced corruption robustness benchmarks for CVGL that include four representative noise types, namely Gaussian, shot, impulse, and speckle noise, each evaluated at five severity levels, and further explored training-time augmentation strategies for improving robustness. MRGeo (Wu et al., 2026) further proposed a dedicated robust CVGL architecture and improved overall corruption robustness, but did not separately report noise-specific performance. In contrast, we focus on query-side noise and study test-time denoising without modifying or retraining the localization model.

2.3 Image Denoising

Deep image denoising methods have achieved strong performance by learning mappings from noisy observations to clean images. DnCNN (Zhang et al., 2017) learns residual noise prediction with convolutional networks, while FFDNet (Zhang et al., 2018) introduces an explicit noise-level map to support flexible denoising strength. DRUNet (Zhang et al., 2021) further combines a U-Net architecture with residual blocks and explicit noise-level conditioning, providing a strong pretrained denoising model. Rather than developing a new denoiser, we study whether these pretrained models can transfer to noisy CVGL at test time.

3 Method

3.1 Problem Formulation

Let $\mathcal{G} = \{s_j\}_{j=1}^{N}$ denote a geo-tagged satellite gallery containing N images. A CVGL model consists of a ground-view encoder $f_g(\cdot;\phi_g)$ and a satellite-view encoder $f_s(\cdot;\phi_s)$. Given a clean ground-view query x, the query and gallery representations are

$$q_x = f_g(x;\phi_g), \qquad g_j = f_s(s_j;\phi_s). \tag{1}$$

Localization retrieves the satellite image with the highest feature similarity

$$j^* = \arg\max_{j\in\{1,\dots,N\}} \mathrm{sim}(q_x, g_j), \tag{2}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes the similarity function used by the corresponding CVGL model.

During deployment, the ground-view query may be affected by image noise. We denote a noisy observation under noise type k and severity η as

$$y_k = \mathcal{N}_k(x;\eta), \tag{3}$$

where $\mathcal{N}_k$ denotes the corresponding noise process. Noise can distort local appearance and structural cues used for cross-view matching, causing the resulting feature representation to deviate from its clean-query counterpart.

Our goal is to improve robustness to query-side noise without retraining or modifying the underlying CVGL model. We first characterize the heterogeneous observation statistics induced by different noise processes and distinguish them from the clean-image target shared by the denoising problem. Based on this distinction, we introduce a shared pretrained denoising strategy before the frozen CVGL model.

3.2 Heterogeneous Noise Statistics and Shared Clean Target

The Gaussian, shot, speckle, and impulse noise processes considered in this work arise from different stochastic mechanisms (Foi et al., 2008; Goodman, 1976; Wang & Zhang, 1999). We therefore do not assume that they share a common physical generation mechanism or probability distribution.

For a clean pixel intensity x_i and its noisy observation $y_{k,i}$, we define the noise residual as

Table 1: Conditional mean and variance of the query-side noise processes considered in this work.

Noise	$\mu_k(x)$	$v_k(x)$
Gaussian	0	σ_g^2
Shot	0	x/c
Speckle	0	$\sigma_s^2 x^2$
Impulse	$p(1/2 - x)$	$\frac{p}{4} + p(1-p)(x-1/2)^2$

$$r_{k,i} = y_{k,i} - x_i. \tag{4}$$

For noise type k at severity η, the conditional mean and variance of the residual are

$$\mu_k(x_i;\eta) = \mathbb{E}\left[r_{k,i} \mid x_i\right], \qquad v_k(x_i;\eta) = \mathrm{Var}\left[r_{k,i} \mid x_i\right]. \tag{5}$$

The noisy observation can consequently be expressed as

$$y_{k,i} = x_i + \mu_k(x_i;\eta) + \sqrt{v_k(x_i;\eta)}\epsilon_{k,i}, \tag{6}$$

where the standardized residual satisfies

$$\mathbb{E}\left[\epsilon_{k,i} \mid x_i\right] = 0, \qquad \mathrm{Var}\left[\epsilon_{k,i} \mid x_i\right] = 1. \tag{7}$$

Equation equation 6 is a conditional moment decomposition rather than a unified physical noise model. The dependence on k and η preserves the distinct conditional mean, variance, and residual distribution of each noise process.

Under the noise definitions used in our evaluation, their conditional statistics are summarized in Table 1. Detailed derivations of these expressions are provided in Appendix.

Gaussian noise has signal-independent variance, whereas shot and speckle noise exhibit different forms of signal-dependent variance. Impulse noise follows a replacement-based process and, under the noise definition used in our evaluation, introduces an intensity-dependent conditional mean. These differences show that the noisy observations presented to the CVGL model can follow substantially different statistics.

The heterogeneity of the observation statistics should be distinguished from the target of the denoising problem. For a noise type k at severity η, a noise-specific denoising estimator can be conceptually written as

$$D_{k,\eta}^* = \arg\min_D \mathbb{E}_{\substack{x\sim p_X,\\ y\sim p_k(\cdot\mid x,\eta)}} \left[\ell\left(D(y), x\right)\right], \tag{8}$$

where p_X denotes the clean-image distribution, $p_k(y \mid x, \eta)$ denotes the observation distribution induced by noise type k at severity η, and $\ell(\cdot,\cdot)$ denotes a generic denoising loss. Equation 8 is introduced only to distinguish the noise-dependent observation distribution from the shared clean-image target and is not optimized in our method.

Across different noise types and severity levels, the observation distribution $p_k(y \mid x, \eta)$ changes, consistent with the heterogeneous statistics described above. The target in Equation 8, however, remains the corresponding clean image x. The commonality across noise conditions therefore lies in the underlying clean-image target rather than in their observation statistics.

This distinction does not imply that the optimal estimators $D_{k,\eta}^*$ are identical. Instead, it motivates the empirical question studied in this work. We investigate whether a fixed pretrained denoiser can improve downstream CVGL robustness across heterogeneous query-side noise without estimating the noise type or severity of each test query.

3.3 Shared Pretrained Denoising Strategy

Let D_θ denote a pretrained image denoiser with fixed parameters θ. Rather than training a separate noise-specific estimator $D^*_{k,\eta}$, we apply the same pretrained denoiser to every noisy query

$$\hat{x}_k = D_\theta(y_k; \gamma_D), \tag{9}$$

where γ_D denotes optional denoiser-specific test-time conditioning. For denoisers that do not require explicit conditioning, γ_D is omitted.

For a given denoiser, both its parameters θ and its test-time configuration γ_D remain fixed across noise types and severity levels. The denoiser is therefore not adapted according to the unknown noise type or severity of an individual query. No noisy CVGL data are used to retrain or fine-tune D_θ.

The statistical formulation in Sec. 3.2 characterizes the noise process at the pixel level. In contrast, D_θ is an image-level operator

$$D_\theta \colon \mathbb{R}^{H\times W\times C} \to \mathbb{R}^{H\times W\times C}, \tag{10}$$

which processes the complete query image and can exploit spatial context during denoising. We do not assume that a fixed pretrained denoiser approximates every noise-specific optimum $D^*_{k,\eta}$ equally well. Its effectiveness across different noise types and severity levels is evaluated empirically.

The proposed strategy is independent of a particular denoising architecture. In the main experiments, we instantiate D_θ with DRUNet (Zhang et al., 2021). We additionally evaluate other pretrained denoisers to examine how the choice of D_θ affects downstream CVGL robustness.

3.4 Integration with Frozen Cross-View Geo-Localization

The denoised query is directly processed by the original ground-view encoder

$$q_{\hat{x}_k} = f_g(\hat{x}_k; \phi_g). \tag{11}$$

The satellite gallery remains unchanged and its representations are

$$g_j = f_s(s_j; \phi_s), \qquad j = 1, \ldots, N. \tag{12}$$

Localization follows the original retrieval rule

$$j^* = \arg\max_{j\in\{1,\ldots,N\}} \operatorname{sim}\left(q_{\hat{x}_k}, g_j\right). \tag{13}$$

Substituting Equation 9 into Equation 11, the complete inference process under the proposed strategy is

$$j^* = \arg\max_{j\in\{1,\ldots,N\}} \operatorname{sim}\left(f_g\left(D_\theta(y_k; \gamma_D); \phi_g\right), g_j\right). \tag{14}$$

The denoiser parameters θ and the CVGL parameters ϕ_g and ϕ_s remain fixed throughout evaluation. Neither the satellite gallery, the similarity function, nor the CVGL architecture is modified.

The proposed strategy therefore modifies only the query-side preprocessing stage. It requires no CVGL fine-tuning, no joint denoising-localization training, and no noise-specific denoiser trained on the target CVGL data.

Table 2: Main results for cross-view geo-localization under varying sensor noise. We report **the average R@1 (%)** across five severity levels for Gaussian (Gauss.), Shot, Impulse (Imp.), and Speckle (Speck.) noise on the CVUSA-C and CVACT_val-C datasets. "Raw" denotes the baseline performance on noisy queries, while "Denoise" indicates queries pre-processed by **DRUNet** (Zhang et al., 2021). **Bold** values indicate the better performance. Small arrows next to the Avg. columns denote the absolute performance change from the Raw baseline.

Model	Setting	CVUSA-C					CVACT_val-C				
		Gauss.	Shot	Imp.	Speck.	Avg.	Gauss.	Shot	Imp.	Speck.	Avg.
SAFA	Raw	33.04	28.51	30.38	37.59	32.38	48.66	43.68	48.82	51.61	48.19
	Denoised	**56.60**	**50.92**	**60.42**	**62.20**	**57.54** ↑25.16	**50.07**	**46.80**	**53.69**	50.45	**50.25** ↑2.06
CVFT	Raw	21.83	19.19	20.57	26.25	21.96	35.68	30.80	36.32	36.84	34.91
	Denoised	**29.03**	**26.78**	**31.76**	**30.85**	**29.61** ↑7.65	**47.74**	**48.85**	**46.95**	**47.30**	**47.71** ↑12.80
DSM	Raw	49.55	46.40	48.84	60.83	51.41	58.34	53.05	58.18	63.06	58.16
	Denoised	**72.43**	**69.27**	**75.74**	**74.90**	**73.08** ↑21.67	**73.23**	**73.96**	**72.40**	**72.92**	**73.13** ↑14.97
L2LTR	Raw	81.32	80.29	82.88	86.54	82.76	82.20	81.19	**82.98**	82.23	82.15
	Denoised	**90.04**	**89.37**	**90.75**	**90.79**	**90.24** ↑7.48	**82.62**	**82.72**	82.48	**82.53**	**82.59** ↑0.44
TransGeo	Raw	66.00	61.80	66.99	77.65	68.11	55.87	48.77	57.65	60.03	55.58
	Denoised	**84.61**	**82.49**	**88.37**	**89.83**	**86.33** ↑18.22	**67.71**	**61.25**	**71.28**	**73.38**	**68.40** ↑12.82
GeoDTR+	Raw	73.19	72.41	74.89	83.86	76.09	**82.78**	**81.62**	**83.62**	**84.73**	**83.19**
	Denoised	**85.65**	**84.83**	**86.99**	**86.91**	**86.10** ↑10.01	81.09	81.54	81.18	80.83	81.16 ↓2.03
Sample4Geo	Raw	80.34	77.26	80.51	88.96	81.77	71.11	67.13	73.73	77.91	72.47
	Denoised	**85.75**	**82.03**	**86.42**	**90.14**	**86.16** ↑4.39	**82.96**	**80.59**	**84.82**	**84.40**	**83.19** ↑10.72

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and noise benchmarks. We use the noise subsets of the CVUSA-C and CVACT_val-C corruption benchmarks introduced in (Zhang & Zhu, 2024), which contain Gaussian, shot, impulse, and speckle noise at five severity levels. Only ground-view queries are corrupted, while the satellite gallery remains clean.

CVGL models. To evaluate whether test-time denoising generalizes across localization architectures, we consider representative CVGL models spanning different network designs, including SAFA (Shi et al., 2019), CVFT (Shi et al., 2020b), DSM (Shi et al., 2020a), L2LTR (Yang et al., 2021), TransGeo (Zhu et al., 2022), GeoDTR+ (Zhang et al., 2024), and Sample4Geo (Deuser et al., 2023). All localization models use their original pretrained weights and remain fully frozen throughout evaluation.

Test-time denoising. Unless otherwise specified, we use a publicly available pretrained DRUNet (Zhang et al., 2021) as the default denoiser. Each noisy query is first restored by the frozen denoiser and then passed to the corresponding frozen CVGL model for retrieval. No noisy images are used for training or adaptation, and neither the denoiser nor the localization model is updated during testing. We additionally evaluate DnCNN (Zhang et al., 2017) and FFDNet (Zhang et al., 2018) to assess the generality of test-time denoising.

Evaluation metrics. We evaluate localization performance using Recall@1 (R@1), Recall@5 (R@5), Recall@10 (R@10), and Recall@1% (R@1%), which measure whether the ground-truth satellite image is retrieved within the top-ranked candidates.

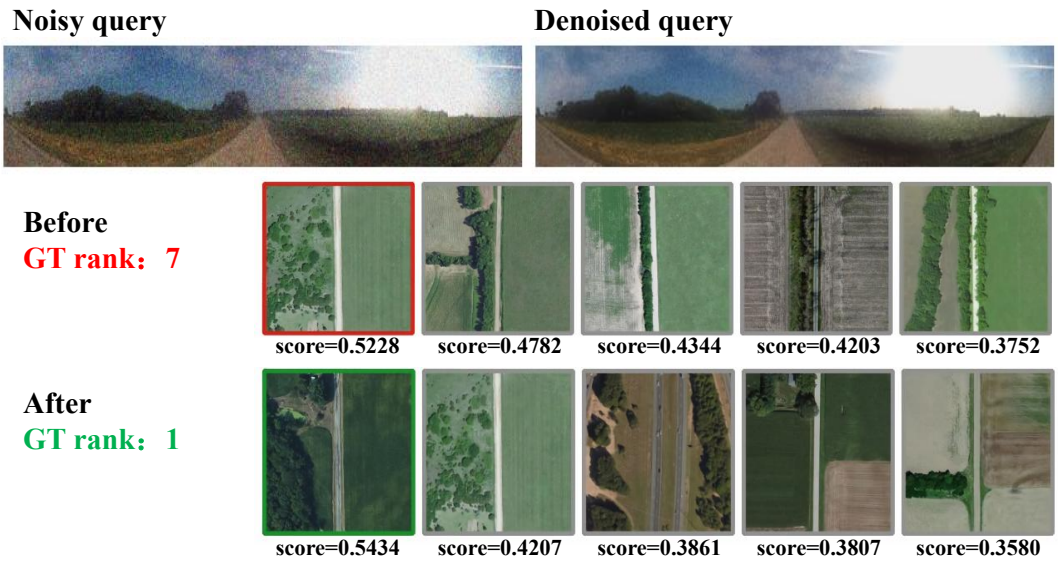


Figure 2: **Qualitative comparison of cross-view retrieval before and after query denoising.** The top row shows the raw corrupted query and its restored version. Below, we display the top retrieved satellite images for each setting. The ground-truth match is highlighted with a green box, while incorrect candidates are outlined in red. Pretrained denoising effectively corrects the retrieval failure, promoting the ground truth from rank 7 to rank 1.

4.2 Main Results

Table 2 reports the average R@1 over five corruption severity levels across seven CVGL models and four noise types. For the majority of model–dataset combinations, applying the pretrained denoiser improves robustness, demonstrating that query-side denoising can serve as an effective test-time intervention without modifying the localization model. On CVUSA-C denoising consistently increases the average R@1 for all evaluated models, with gains ranging from 4.32 points for Sample4Geo (Deuser et al., 2023) to 25.16 points for SAFA. These improvements are observed across all four corruption types, suggesting that the performance gain is not restricted to a particular noise.

The results on CVACT_val-C exhibit a more heterogeneous pattern. Most models still benefit from denoising—including DSM (Shi et al., 2020a), TransGeo (Zhu et al., 2022), CVFT, and Sample4Geo (Deuser et al., 2023), whose average R@1 improves by 14.97, 12.82, 12.80, and 10.72 points, respectively. In contrast, the gain for L2LTR (Yang et al., 2021) is marginal (+0.44), while GeoDTR+ (Zhang et al., 2024) decreases by 2.03 points. These variations highlight that the interaction between denoising and localization is highly model-dependent. Although denoising can suppress corruption-induced distortions, it may simultaneously introduce smoothing effects or alter original image details that certain highly robust models might otherwise rely on. Consequently, these findings establish pretrained denoising as a broadly effective robustness strategy, though its ultimate impact remains tied to the specific characteristics of the downstream localization architecture.

4.3 Qualitative Analysis

Figure 2 illustrates how query restoration affects downstream cross-view retrieval. Before denoising, the ground-truth satellite image is ranked 7th, while several incorrect candidates with similar coarse spatial layouts receive higher similarity scores. After applying the pretrained denoiser to reduce the visual degradation, the ground-truth image is promoted to rank 1. Notably, this improvement is accompanied by a reordering of the top retrieved candidates rather than a uniform increase in similarity scores. This example suggests that denoising can modify the relative similarity ordering among competing candidates, making the correct match more distinguishable under noisy observations. Additional qualitative results, including both success and failure cases, are provided in the supplementary material.

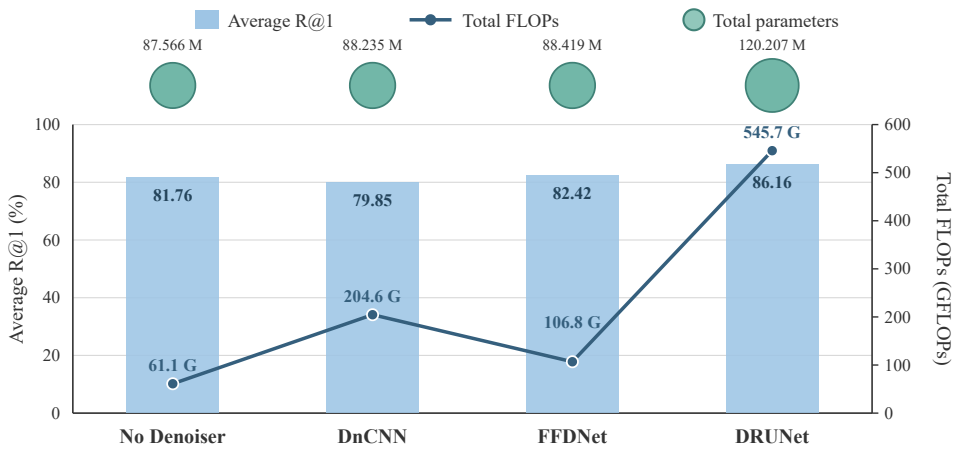


Figure 3: **Robustness–efficiency trade-off of different pretrained denoisers on Sample4Geo (Deuser et al., 2023) over CVUSA-C.** Bars report the average R@1 across the evaluated corruption types and severity levels, the line denotes the total FLOPs of the complete denoising–localization pipeline, and bubble size represents the total number of model parameters.

4.4 Efficiency Analysis

Since the denoising prior is prepended to the frozen CVGL pipeline, improved robustness inevitably introduces additional inference cost. We therefore evaluate different pretrained denoisers on Sample4Geo (Deuser et al., 2023) over CVUSA-C by comparing the average R@1, total FLOPs, and total parameter count of the complete denoising–localization pipeline. As shown in Figure 3, directly processing corrupted queries without denoising requires only 61.1 GFLOPs and achieves an average R@1 of 81.76%. Among the evaluated denoising priors, DRUNet (Zhang et al., 2021) provides the strongest robustness, improving the average R@1 to 86.16%, but incurs a computational overhead with 545.7 GFLOPs and 120.207 M parameters. We adopt DRUNet (Zhang et al., 2021) as the default denoiser in our main experiments as it yields the highest empirical robustness among the evaluated models, allowing us to clearly demonstrate the potential benefits of query-side denoising. In comparison, FFDNet (Zhang et al., 2018) offers a computationally lighter option, achieving an average R@1 of 82.42% with 106.8 GFLOPs. Interestingly, although DnCNN (Zhang et al., 2017) increases the total computation to 204.6 GFLOPs, it reduces the average R@1 to 79.85%. This performance drop demonstrates that introducing a denoising module alone does not guarantee improved downstream localization; rather, the specific characteristics of the chosen denoising prior influence the final performance. Consequently, these findings highlight a clear robustness–efficiency trade-off, where different denoisers occupy distinct positions on the spectrum of computational cost and localization accuracy.

4.5 Effect of Corruption Severity

To better understand how pretrained denoising affects robustness under different corruption strengths, we evaluate Sample4Geo (Deuser et al., 2023) across five severity levels under Gaussian, Shot, Impulse, and Speckle noise. As shown in Figure 4, the performance of raw queries drops rapidly as corruption becomes stronger, whereas denoised queries exhibit a more stable degradation trend. The advantage of denoising becomes increasingly evident at higher corruption levels. From Severity 3 onward, the denoised queries consistently outperform the raw baseline across all four corruption types, with particularly large gains at Severity 4 and 5. Under the mildest corruption level, only a small decrease in R@1 is observed after denoising. We hypothesize that when corruption is limited, the original query already preserves most discriminative cues, while the restoration process may slightly perturb some features useful to the frozen localization model. As corruption increases, however, removing corruption becomes more beneficial, resulting in substantially improved downstream localization robustness. Additional feature-level analysis is provided in the supplementary material, where clean-reference feature similarity and retrieval margin exhibit trends consistent with the downstream results. We also provide complete severity-wise results for additional CVGL

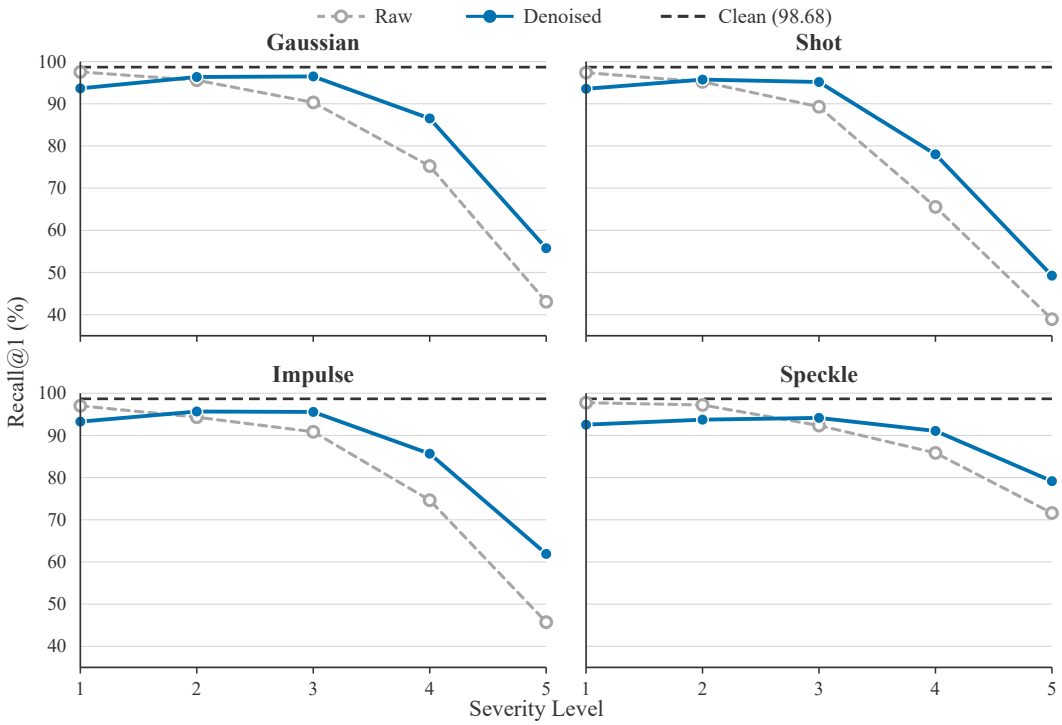


Figure 4: **R@1 of Sample4Geo (Deuser et al., 2023) on CVUSA-C under increasing corruption severity.** We compare direct geo-localization from corrupted queries (Raw) with geo-localization after pretrained denoising (Denoised) under Gaussian, Shot, Impulse, and Speckle noise. The dashed horizontal line denotes R@1 on clean queries.

models on both CVUSA-C and CVACT_val-C in the supplementary material, further examining the behavior of pretrained denoising across architectures and datasets.

5 Conclusion and Limitations

In this work, we study query-side noise in cross-view geo-localization (CVGL) from a shared denoising perspective. Rather than modifying or retraining the localization model, we apply a pretrained denoiser prior to feature extraction, keeping the downstream CVGL model frozen. By separating noise-dependent observation statistics from the underlying clean-image target, our approach accommodates Gaussian, shot, speckle, and impulse noise within a single framework. Experiments across multiple architectures and noise severities demonstrate that this plug-and-play strategy effectively mitigates performance drops, particularly under severe degradation. These results suggest that pretrained denoising serves as a simple yet practical test-time baseline for improving the noise-robustness of existing CVGL systems.

Our study also has a few limitations. The effectiveness of this plug-and-play denoising relies heavily on the specific combination of the denoiser and the downstream CVGL architecture, meaning performance gains can vary across different noise severities. Additionally, our evaluation is based on benchmark-generated query corruptions, and future work should further validate the findings on naturally degraded or real sensor imagery. Finally, relying on a fixed denoising configuration—though deployment-friendly—limits the method's potential under varying noise intensities. Future work could investigate adaptive, yet training-free, denoising strategies to handle a wider spectrum of real-world corruptions.

6 AI USE STATEMENT

Generative AI tools were used solely for language editing and polishing to improve the grammar, readability, and clarity of the manuscript. They were not used to conduct experiments, generate experimental results, or make scientific conclusions. We take responsibility for the final content of this work, including the AI-assisted text.

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.

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

Alessandro Foi, Mejdi Trimeche, Vladimir Katkovnik, and Karen Egiazarian. Practical poissonian-gaussian noise modeling and fitting for single-image raw-data. *IEEE transactions on image processing*, 17(10):1737–1754, 2008.

Joseph W Goodman. Some fundamental properties of speckle. *JOSA*, 66(11):1145–1150, 1976.

Christian Häne, Lionel Heng, Gim Hee Lee, Friedrich Fraundorfer, Paul Furgale, Torsten Sattler, and Marc Pollefeys. 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, 2017.

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*, pp. 6155–6164, 2022.

Ling Li, Yu Ye, Yao Zhou, Bingchuan Jiang, and Wei Zeng. Georeasoner: Geo-localization with reasoning in street views using a large vision-language model. *arXiv preprint arXiv:2406.18572*, 2024.

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

Colin McManus, Winston Churchill, Will Maddern, Alexander D Stewart, and Paul Newman. Shady dealings: Robust, long-term visual localisation using illumination invariance. In *2014 IEEE international conference on robotics and automation (ICRA)*, pp. 901–906. IEEE, 2014.

Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. *Advances in Neural Information Processing Systems*, 32, 2019.

Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. Where am i looking at? joint location and orientation estimation by cross-view matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4064–4072, 2020a.

Yujiao Shi, Xin Yu, Liu Liu, Tong Zhang, and Hongdong Li. Optimal feature transport for cross-view image geo-localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 11990–11997, 2020b.

Aysim Toker, Qunjie Zhou, Maxim Maximov, and Laura Leal-Taixé. Coming down to earth: Satellite-to-street view synthesis for geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6488–6497, 2021.

Zhou Wang and David Zhang. Progressive switching median filter for the removal of impulse noise from highly corrupted images. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 46(1):78–80, 1999.

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 evolving transformer. *arXiv preprint arXiv:2107.00842*, 2021.

Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. Predicting ground-level scene layout from aerial imagery. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 867–875, 2017.

Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26 (7):3142–3155, 2017.

Kai Zhang, Wangmeng Zuo, and Lei Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE Transactions on Image Processing*, 27(9):4608–4622, 2018.

Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6360–6376, 2021.

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, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1162–1171, 2022.

A Derivations of Query Noise Statistics

This section provides the detailed derivations of the conditional bias $\mu_k(x)$ and conditional variance $v_k(x)$ used in Sec. 3. These derivations characterize the distinct statistics of Gaussian, shot, speckle, and impulse noise while preserving the common conditional-moment representation

$$y_{k,i} = x_i + \mu_k(x_i) + \sqrt{v_k(x_i)}\epsilon_{k,i}, \tag{15}$$

where

$$\mathbb{E}[\epsilon_{k,i} \mid x_i] = 0, \qquad \mathrm{Var}[\epsilon_{k,i} \mid x_i] = 1. \tag{16}$$

A.1 Gaussian Noise

Additive Gaussian noise is defined as

$$y_{\mathrm{G},i} = x_i + \sigma_g \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0,1). \tag{17}$$

Therefore,

$$\mathbb{E}[y_{\mathrm{G},i} \mid x_i] = x_i, \tag{18}$$

and

$$\mathrm{Var}[y_{\mathrm{G},i} \mid x_i] = \sigma_g^2. \tag{19}$$

Hence,

$$\mu_{\mathrm{G}}(x) = 0, \qquad v_{\mathrm{G}}(x) = \sigma_g^2. \tag{20}$$

Gaussian noise therefore corresponds to the signal-independent case

$$(\alpha_{0,\mathrm{G}}, \alpha_{1,\mathrm{G}}, \alpha_{2,\mathrm{G}}) = \left(\sigma_g^2, 0, 0\right). \tag{21}$$

A.2 SHOT NOISE

Shot noise is modeled through a signal-dependent counting process. Let

$$n_i \sim \text{Poisson}(cx_i), \tag{22}$$

and define

$$y_{\text{P},i} = \frac{n_i}{c}, \tag{23}$$

where $c > 0$ controls the noise level.

Since

$$\mathbb{E}[n_i \mid x_i] = cx_i, \qquad \text{Var}[n_i \mid x_i] = cx_i, \tag{24}$$

we obtain

$$\mathbb{E}[y_{\text{P},i} \mid x_i] = x_i, \tag{25}$$

and

$$\text{Var}[y_{\text{P},i} \mid x_i] = \frac{x_i}{c}. \tag{26}$$

Define the standardized residual

$$\epsilon_{\text{P},i} = \frac{n_i - cx_i}{\sqrt{cx_i}}. \tag{27}$$

Then

$$\mathbb{E}[\epsilon_{\text{P},i} \mid x_i] = 0, \qquad \text{Var}[\epsilon_{\text{P},i} \mid x_i] = 1. \tag{28}$$

Substituting this expression yields

$$y_{\text{P},i} = x_i + \sqrt{\frac{x_i}{c}}\epsilon_{\text{P},i}. \tag{29}$$

Thus,

$$\mu_{\text{P}}(x) = 0, \qquad v_{\text{P}}(x) = \frac{x}{c}. \tag{30}$$

Shot noise therefore corresponds to

$$(\alpha_{0,\text{P}}, \alpha_{1,\text{P}}, \alpha_{2,\text{P}}) = \left(0, \frac{1}{c}, 0\right). \tag{31}$$

This is a moment-standardized representation rather than a Gaussian approximation. The standardized residual remains Poisson-derived.

A.3 SPECKLE NOISE

Speckle noise is multiplicative and can be expressed as

$$y_{\text{S},i} = x_i + \sigma_s x_i \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, 1). \tag{32}$$

Therefore,

$$\mathbb{E}[y_{\text{S},i} \mid x_i] = x_i, \tag{33}$$

and

$$\text{Var}[y_{\text{S},i} \mid x_i] = \sigma_s^2 x_i^2. \tag{34}$$

Hence,

$$\mu_{\text{S}}(x) = 0, \qquad v_{\text{S}}(x) = \sigma_s^2 x^2. \tag{35}$$

Speckle noise therefore corresponds to

$$(\alpha_{0,\text{S}}, \alpha_{1,\text{S}}, \alpha_{2,\text{S}}) = \left(0, 0, \sigma_s^2\right). \tag{36}$$

A.4 IMPULSE NOISE

Impulse noise randomly replaces pixel values rather than continuously perturbing them. For symmetric salt-and-pepper noise with replacement probability p,

$$y_{\mathrm{I},i} = \begin{cases} x_i, & \text{with probability } 1-p, \\ 0, & \text{with probability } p/2, \\ 1, & \text{with probability } p/2. \end{cases} \tag{37}$$

Its conditional expectation is

$$\mathbb{E}[y_{\mathrm{I},i} \mid x_i] = (1-p)x_i + \frac{p}{2}. \tag{38}$$

Therefore, the conditional bias is

$$\mu_{\mathrm{I}}(x) = p\left(\frac{1}{2} - x\right). \tag{39}$$

The conditional second moment is

$$\mathbb{E}[y_{\mathrm{I},i}^2 \mid x_i] = (1-p)x_i^2 + \frac{p}{2}. \tag{40}$$

Hence,

$$\begin{aligned} v_{\mathrm{I}}(x_i) &= \mathrm{Var}[y_{\mathrm{I},i} \mid x_i] \\ &= \frac{p}{4} + p(1-p)\left(x_i - \frac{1}{2}\right)^2. \end{aligned} \tag{41}$$

Equivalently,

$$v_{\mathrm{I}}(x) = \frac{p(2-p)}{4} - p(1-p)x + p(1-p)x^2. \tag{42}$$

A standardized residual can then be defined as

$$\epsilon_{\mathrm{I},i} = \frac{y_{\mathrm{I},i} - x_i - \mu_{\mathrm{I}}(x_i)}{\sqrt{v_{\mathrm{I}}(x_i)}}. \tag{43}$$

It follows that

$$\mathbb{E}[\epsilon_{\mathrm{I},i} \mid x_i] = 0, \qquad \mathrm{Var}[\epsilon_{\mathrm{I},i} \mid x_i] = 1. \tag{44}$$

Although impulse noise remains discrete and heavy-tailed, its conditional first- and second-order statistics are compatible with the unified conditional-moment representation used in the main text.

The four noise processes exhibit distinct conditional statistics. Gaussian noise has signal-independent variance, shot noise has variance linear in the clean intensity, speckle noise has quadratic variance, and impulse noise introduces both an intensity-dependent bias and quadratic variance. These differences characterize the noise-dependent observation statistics, while the common underlying clean-image target motivates the shared denoising strategy described in the main paper.

B DETAILED RESULTS ACROSS CORRUPTION SEVERITIES

This section provides the complete R@1 results underlying the noise-severity analysis in the main paper, with detailed results reported in Tables 3–5.

Table 3: **Detailed R@1 (%) results of Sample4geo (Deuser et al., 2023) under increasing corruption severity.** We report performance on CVUSA-C and CVACT_val-C for Gaussian, Shot, Impulse, and Speckle noise across five severity levels, comparing direct localization from corrupted queries (Raw) with localization after pretrained denoising (Denoised). The Avg. column denotes the mean Recall@1 over the four corruption types at each severity level.

Severity	Setting	CVUSA-C					CVACT_val-C				
		Gaussian	Shot	Impulse	Speckle	Avg.	Gaussian	Shot	Impulse	Speckle	Avg.
Severity 1	Raw	97.55	97.38	97.02	97.76	97.43	89.23	88.99	89.25	89.31	89.19
	Denoised	93.65	93.55	93.27	92.55	93.26	88.51	88.54	88.17	88.20	88.36
Severity 2	Raw	95.58	95.20	94.32	97.22	95.58	87.82	86.89	87.55	88.43	87.67
	Denoised	96.35	95.73	95.67	93.74	95.37	89.13	88.67	88.35	88.43	88.64
Severity 3	Raw	90.31	89.28	90.86	92.35	90.70	83.41	80.99	84.94	82.81	83.04
	Denoised	96.48	95.16	95.57	94.16	95.34	87.00	85.96	88.03	85.96	86.74
Severity 4	Raw	75.23	65.54	74.64	85.85	75.32	68.26	54.42	71.43	74.71	67.21
	Denoised	86.52	78.01	85.67	91.06	85.32	82.06	79.09	83.94	82.51	81.90
Severity 5	Raw	43.05	38.92	45.72	71.60	49.82	26.85	24.34	35.46	54.29	35.23
	Denoised	55.76	49.25	61.91	79.16	61.52	68.13	60.67	75.61	76.91	70.33

Table 4: **Detailed R@1 (%) results of TransGeo (Zhu et al., 2022) under increasing corruption severity.** We report performance on CVUSA-C and CVACT_val-C for Gaussian, Shot, Impulse, and Speckle noise across five severity levels, comparing direct localization from corrupted queries (Raw) with localization after pretrained denoising (Denoised). The Avg. column denotes the mean Recall@1 over the four corruption types at each severity level.

Severity	Setting	CVUSA-C					CVACT_val-C				
		Gaussian	Shot	Impulse	Speckle	Avg.	Gaussian	Shot	Impulse	Speckle	Avg.
Severity 1	Raw	92.64	92.86	92.31	93.24	92.76	82.64	82.45	82.27	82.86	82.56
	Denoised	90.38	90.60	90.56	90.36	90.48	83.17	83.17	83.10	83.12	83.14
Severity 2	Raw	90.95	89.89	88.92	92.51	90.57	80.11	77.86	79.23	81.30	79.63
	Denoised	90.59	91.04	90.89	90.39	90.73	83.44	83.13	83.03	83.26	83.22
Severity 3	Raw	81.56	79.85	83.17	85.56	82.54	71.17	65.32	74.62	69.71	70.21
	Denoised	91.46	91.77	91.48	91.11	91.46	81.97	79.93	82.22	80.54	81.17
Severity 4	Raw	53.40	37.40	55.45	72.85	54.78	42.39	16.91	46.92	50.21	39.11
	Denoised	90.88	86.72	91.55	91.07	90.06	70.06	49.85	74.03	72.79	66.68
Severity 5	Raw	11.43	9.00	15.07	44.09	19.90	3.04	1.29	5.22	16.06	6.40
	Denoised	59.76	52.32	77.40	86.22	68.93	19.93	10.15	34.02	47.17	27.82

C Feature-Level Analysis Across Corruption Severity

To further examine how pretrained denoising affects the representation space under different corruption strengths, we analyze two feature-level quantities using the frozen Sample4Geo (Deuser et al., 2023) encoder on CVUSA (Zhai et al., 2017): clean-reference feature similarity and retrieval margin. For each corrupted query, we compare the denoised representation with its raw noisy counterpart and report the denoising-induced gain.

As shown in Figure 5, the feature-level behavior is broadly consistent with the downstream retrieval trends reported in the main paper. Under the mildest corruption level, denoising produces small negative changes in both metrics for most corruption types. This suggests that when the original query remains largely reliable, restoration may slightly perturb feature patterns that are already useful for localization. As corruption becomes stronger, however, the effect reverses. From Severity 3 onward, both clean-reference feature similarity and retrieval margin improve across all four corruption types.

Table 5: **Detailed R@1 (%) results of GeoDTR+ (Zhang et al., 2024) under increasing corruption severity.** We report performance on CVUSA-C and CVACT_val-C for Gaussian, Shot, Impulse, and Speckle noise across five severity levels, comparing direct localization from corrupted queries (Raw) with localization after pretrained denoising (Denoised). The Avg. column denotes the mean Recall@1 over the four corruption types at each severity level.

Severity	Setting	CVUSA-C					CVACT_val-C				
		Gaussian	Shot	Impulse	Speckle	Avg.	Gaussian	Shot	Impulse	Speckle	Avg.
Severity 1	Raw	93.13	92.86	92.18	93.57	92.94	87.03	87.01	87.08	87.42	87.13
	Denoised	86.40	86.29	86.40	86.24	86.33	80.86	80.82	80.57	80.77	80.76
Severity 2	Raw	90.18	89.93	88.94	92.64	90.42	86.74	86.45	86.62	87.00	86.70
	Denoised	87.25	87.36	87.25	86.32	87.04	81.07	81.20	80.16	80.85	80.82
Severity 3	Raw	83.53	83.37	84.43	86.83	84.54	85.50	85.19	85.65	85.50	85.46
	Denoised	88.97	89.10	87.91	87.98	88.49	81.22	81.24	80.20	80.81	80.87
Severity 4	Raw	67.67	60.00	67.93	80.53	69.03	82.51	80.31	83.24	83.89	82.49
	Denoised	90.42	87.83	90.42	88.06	89.18	81.20	81.88	80.13	80.68	80.97
Severity 5	Raw	36.80	35.61	40.97	65.72	44.78	72.14	69.12	75.54	79.82	74.16
	Denoised	75.23	73.58	82.99	85.95	79.44	81.08	82.59	79.86	81.02	81.14

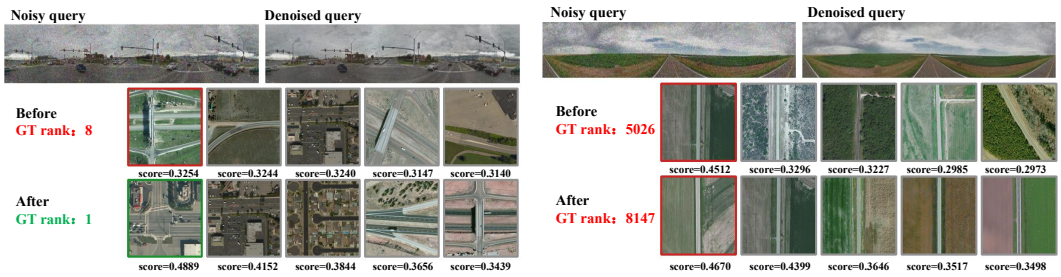


Figure 5: **Feature-level effects of pretrained denoising for Sample4Geo (Deuser et al., 2023) on CVUSA (Zhai et al., 2017) across corruption severity.** We report the denoising-induced gain over raw corrupted queries in (a) clean-reference feature similarity and (b) retrieval margin under Gaussian, Shot, Impulse, and Speckle noise. Positive values indicate that denoising moves the corrupted-query representation closer to its clean counterpart or increases the separation between the ground-truth match and competing negatives.

The transition is particularly clear under Gaussian, Shot, and Impulse noise, where the gains become strongly positive at intermediate severity levels. Speckle noise exhibits a similar overall trend, although the improvement emerges more gradually. These results provide additional feature-space evidence that the robustness gains observed in the main paper are accompanied by improved alignment with the clean-query representation and increased separation between the ground-truth match and competing negatives.

D Additional Qualitative Results

Figure 6 presents additional success and failure cases. In the urban example, denoising promotes the ground-truth satellite image from rank 8 to rank 1, accompanied by a substantial reordering of the retrieved candidates. In contrast, the example shows a failure case in which the ground-truth rank decreases from 5026 to 8147 after denoising. The retrieved candidates in this scene exhibit



(a) Success case. The ground-truth satellite image is promoted from rank 8 to rank 1 after denoising.

(b) Failure case. The ground-truth rank decreases from 5026 to 8147 despite a slight increase in its similarity score.

Figure 6: Additional qualitative results of denoising-assisted cross-view retrieval. Denoising can substantially improve the retrieval ordering, as in the urban example, but does not guarantee improvement for every query.

highly similar coarse field and road layouts, making the cross-view matching problem particularly ambiguous. This example shows that pretrained denoising does not guarantee improved retrieval for every individual query, even though it improves robustness on average.