FOG-ROBUST CROSS-VIEW GEO-LOCALIZATION: DATASETS AND FRAMEWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-view geo-localization (CVGL) aims to match ground-view queries with satellite images, yet existing works largely assume clear visibility and rarely consider adverse weather conditions such as fog. Fog usually obscures textures and structural cues essential for cross-view correspondence. To enable controlled investigation, we construct two synthetic foggy CVGL datasets, CVUSA-Fog and CVACT-Fog, using the atmospheric scattering model while preserving the original ground–satellite correspondence. Beyond synthetic evaluation, we manually curate 1,441 real-fog ground-view images and pair them with satellite images to construct CVFog. Additionally, we further propose a plug-and-play Wavelet-guided Attenuation Structural Adapter (WASA) module, which couples low-frequency degradation cues with structural responses to selectively calibrate fog-affected yet informative regions in ground-view features. Experiments across multiple backbones, datasets, and fog intensities show that synthetic-fog training substantially improves both fog-domain retrieval and synthetic-to-real performance on CVFog, while WASA provides consistent further gains. We further study semi-supervised real-fog adaptation with limited pairing supervision. Under the same real-fog pairing budget, confidence-filtered hard pseudo-pairs improve retrieval over supervised adaptation by exploiting pair-unlabeled CVFog images.

1 INTRODUCTION

Cross-view geo-localization (CVGL) estimates the geographic location of a ground-view query by retrieving its corresponding satellite image. On datasets such as CVUSA (Zhai et al. (2017)) and CVACT (Liu & Li (2019)), recent methods (Chen et al. (2026b); Mi et al. (2024); Yang et al. (2021); Shugaev et al. (2024)) have substantially advanced cross-view matching through convolutional feature alignment and Transformer-based representation learning. However, most are developed for clear weather and assume favorable visibility and stable structural cues in ground-view images.

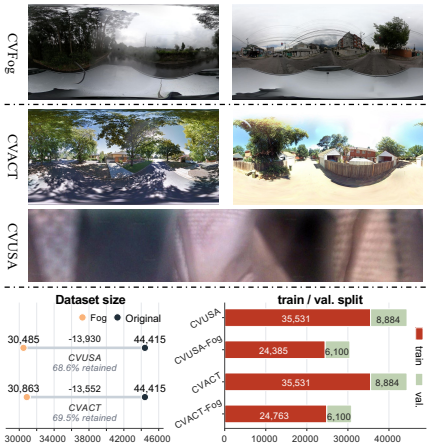


Figure 1: CVFog examples (top), ground-view images excluded from CVACT and CVUSA (middle), and statistics of the retained train and validation (val.) splits (bottom).

In real-world environments, however, ground-view images usually are susceptible to adverse weather. Among these conditions, fog is particularly disruptive, reducing image contrast, attenuating local textures, and obscuring distant structures. The loss of these cues weakens ground-satellite correspondence, further widening the inherent cross-view discrepancy and destabilizing retrieval. Nevertheless, existing methods (Chen et al. (2026a); Zhu et al. (2022); Zhao et al. (2024)) are generally developed under clear-weather assumptions and rarely account explicitly for the loss and distortion of structural cues caused by fog.

The atmospheric scattering model offers a controllable means of synthesizing depth-dependent fog while preserving the original ground–satellite pairings. Synthetic fog (Tsai et al. (2025); Wang et al.

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.

(2026)) has been studied in single-view tasks such as object detection (Mohamed & Huang (2026)) and image dehazing (Lyu et al. (2026); Li et al. (2026a)), and yet its effect on cross-view ground–satellite matching remains underexplored. To address this gap, we employ this model to construct synthetic fog datasets based on CVUSA and CVACT. To ensure high-quality fog synthesis, we manually screen ground-view images for fog synthesis based on exposure, image integrity, and panorama completeness, then repartition the retained pairs into training and validation sets (*Appendix A*). Fig. 1 shows excluded examples (middle) and datasets statistics (bottom). Fog is then applied only to the retained ground-view images while preserving their satellite counterparts, yielding CVUSA-Fog and CVACT-Fog datasets. Beyond the controlled synthetic evaluation, we also manually collect and curate 1,441 real-fog ground-view images from Mapillary. Each image is paired with a satellite image covering the same geographic location to construct CVFog, as shown in Fig. 1 (top).

To mitigate the fog-induced loss of structural cues essential for ground–satellite correspondence, we selectively calibrate fog-degraded features in the feature space rather than reconstructing photorealistic appearance. Motivated by the predominantly low-frequency manifestation of fog degradation in the wavelet domain (Sun et al. (2026); Zhang et al. (2026a)), we propose a plug-and-play Wavelet-guided Attenuation Structural Adapter (WASA) module. Its wavelet branch estimates low-frequency degradation cues, while fixed gradient operators branch guide the extraction of structural responses that remain observable under fog. Their interaction generates a spatial gate that selectively applies residual calibration to fog-affected yet structurally informative regions.

To evaluate the proposed datasets and framework, we conduct experiments across multiple representative CVGL backbones. Training on CVUSA-Fog and CVACT-Fog substantially improves retrieval under fog compared with clear-only training, demonstrating the value of controlled synthetic fog data. WASA provides consistent further gains across architectures, datasets, and degradation levels. Evaluation on CVFog further shows that both synthetic-fog training and WASA improve synthetic-to-real generalization under naturally occurring fog. Finally, we investigate semi-supervised real-fog adaptation under limited pairing supervision, showing that confidence-filtered hard pseudo-pairs effectively exploit pair-unlabeled images.

Our main contributions are summarized as follows:

1. We construct three foggy CVGL datasets: CVUSA-Fog and CVACT-Fog with controllable synthetic fog degradation, and CVFog, a manually curated real-fog dataset containing 1,441 ground–satellite pairs.
2. We propose a plug-and-play WASA module, which jointly models low-frequency degradation cues and structural responses to selectively calibrate fog-affected yet discriminative regions in ground-view features.
3. Extensive experiments demonstrate that synthetic-fog training substantially improves fog-domain retrieval and synthetic-to-real generalization on CVFog. WASA provides further consistent gains across multiple backbones, datasets, and fog intensities. Confidence-filtered hard pseudo-pairs further improve retrieval in semi-supervised real-fog adaptation under limited pairing supervision.

2 Construction of Foggy Cross-View Datasets

Fog synthesis models the depth-dependent mapping from clear-scene radiance to the radiance observed under fog. Given a clear ground-view image J, we estimate relative depth and atmospheric light to synthesize the foggy image I_β.

2.1 Atmospheric Scattering Model

We synthesize foggy ground-view images using the atmospheric scattering model:

$$I_\beta(x) = J(x)t(x) + A\big(1 - t(x)\big), \tag{1}$$

where A denotes atmospheric light and $t(x)$ is the transmittance at pixel x, defined as

$$t(x) = \exp\big(-\beta d(x)\big). \tag{2}$$

Here, $d(x)$ denotes scene depth, and β controls fog intensity, with larger values producing stronger attenuation. Since CVUSA and CVACT datasets do not provide metric depth, we estimate and normalize relative depth for fog synthesis.

2.2 Depth and Atmospheric Light Estimation

Relative depth estimation. We employ Depth Anything V2-Base (Yang et al. (2024)) to estimate an affine-invariant inverse-depth map D_{raw}. To obtain the relative-depth convention required for fog synthesis, we reverse its depth ordering as

$$\widetilde{D}_{\mathrm{raw}}(x) = D_{\max} + D_{\min} - D_{\mathrm{raw}}(x), \tag{3}$$

where $D_{\max}$ and $D_{\min}$ denote the maximum and minimum responses of D_{raw}, respectively.

To reduce the influence of outliers, we use the 2% and 98% depth quantiles (Sakaridis et al. (2018a)), $v_l = Q_{0.02}(\widetilde{D}_{\mathrm{raw}})$ and $v_u = Q_{0.98}(\widetilde{D}_{\mathrm{raw}})$, for percentile clipping:

$$\begin{aligned} D_c(x) &= \mathrm{clip}\left(\widetilde{D}_{\mathrm{raw}}(x), v_l, v_u\right), \\ d(x) &= \frac{D_c(x) - v_l}{\max(v_u - v_l, \epsilon)}. \end{aligned} \tag{4}$$

The normalized depth satisfies $d(x) \in [0, 1]$, and ϵ is a numerical stability term.

Atmospheric light estimation. We use SegFormer (Xie et al. (2021)) to extract the sky region Ω_{sky} and apply mild erosion to remove boundary contamination. The brightest 5% of sky pixels (Sakaridis et al. (2018b)), denoted by Ω_{top}, are selected according to luminance:

$$\begin{aligned} Y(x) &= 0.299R(x) + 0.587G(x) + 0.114B(x), \\ A_{\mathrm{sky}} &= \frac{1}{|\Omega_{\mathrm{top}}|} \sum\nolimits_{x \in \Omega_{\mathrm{top}}} J(x). \end{aligned} \tag{5}$$

To suppress color casts caused by sky-color variations, we construct a gray-white atmospheric light component using the clipped luminance L, whose computation is detailed in the *Appendix C*:

$$\begin{aligned} A_{\mathrm{gray}} &= [L, L, L], \\ A &= \alpha A_{\mathrm{sky}} + (1 - \alpha) A_{\mathrm{gray}}, \end{aligned} \tag{6}$$

where $\alpha = 0.7$ controls the amount of sky color retained.

2.3 Transmittance Refinement and Fog Synthesis

Near-field depth values may approach zero, resulting in negligible fog effects. We therefore introduce a minimum depth offset $d_{\min}$ and refine the initial transmittance using guided filtering (Zhang et al. (2026b); He et al. (2013)):

$$\begin{aligned} d'(x) &= d_{\min} + (1 - d_{\min}) d(x), \\ \widehat{t}(x) &= \exp\big(-\beta d'(x)\big), \\ t &= \mathcal{G}\left(\widehat{t}; J\right), \end{aligned} \tag{7}$$

where $\mathcal{G}(\cdot; J)$ denotes guided filtering with J as guidance. The depth offset introduces moderate scattering in near-field regions, while guided filtering suppresses depth-estimation artifacts and preserves structural boundaries.

With $d(x)$ representing normalized relative depth, β controls attenuation in the normalized depth space. Substituting the refined transmittance into Eq. 1 then produces the final foggy image, with the choices of $d_{\min}$ and β described in *Appendices D* and *E*. We use $\beta \in \{1.5, 2.0, 2.5, 3.0, 3.5\}$, with $\beta = 2.5$ as the default setting. Fig. 2 presents examples from CVUSA-Fog and CVACT-Fog.

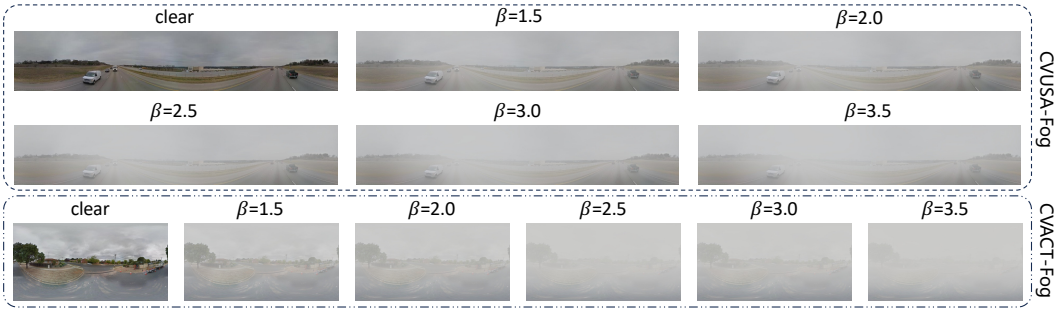


Figure 2: Examples under different attenuation coefficients on CVUSA-Fog and CVACT-Fog.

2.4 CVFog: Real-Fog Dataset Construction

Ground-view imagery. Beyond the controlled synthetic datasets, we construct CVFog using real-fog ground-view imagery collected from Mapillary. We first collect approximately 150,000 candidate records, focusing on mountainous roads, suburban areas, and transition zones between natural and built environments, etc.. We then manually inspect the candidate pool and retain images that contain clearly observable fog and scene content for cross-view matching.

Cross-view pairing. For each retained ground-view image, we retrieve a satellite image covering the same geographic location and manually verify their cross-view correspondence. This manual curation and pairing process yields 1,441 ground–satellite pairs, with representative examples shown in Fig. 1 (top). By complementing synthetic datasets with manually verified real-fog image pairs, CVFog enables evaluation of synthetic-to-real generalization and label-efficient real-fog adaptation. *Appendix E* illustrates the image statistics considered for comparing synthetic and real fog.

3 Wavelet-Guided Attenuation Structural Adapter

Given a shallow ground-view feature $F_g \in \mathbb{R}^{C\times H\times W}$, WASA produces an enhanced feature $\widehat{F}_g$ for subsequent processing. As shown in Fig. 3, its wavelet and structural branches estimate fog degradation and matching-relevant structures, whose interaction guides gated residual calibration.

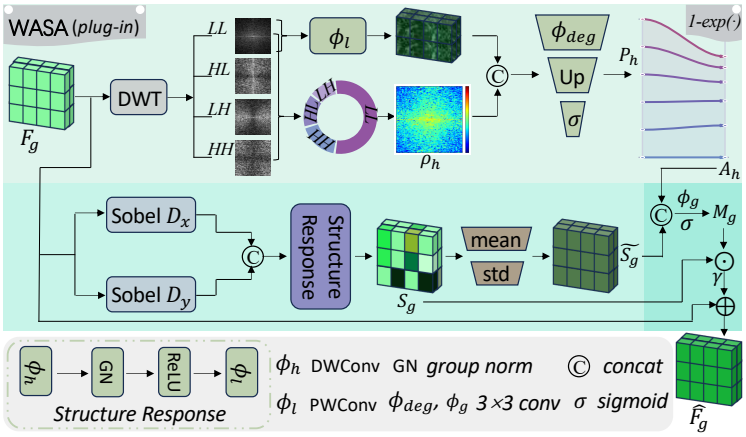


Figure 3: Overview of WASA. The wavelet and Sobel-guided branches produce the degradation cue A_h and structural responses $(S_g, \widetilde{S}_g)$; their interaction forms M_g for gated residual calibration.

3.1 Wavelet-Based Low-Frequency Degradation Perception

Fog-induced loss of visible detail motivates using relative low-frequency energy as a degradation cue (Sun et al. (2026)). We therefore apply a single-level Haar DWT and compute the low-frequency energy concentration at each wavelet location u:

$$\{F_{LL}, F_{LH}, F_{HL}, F_{HH}\} = \mathrm{DWT}(F_g),$$
$$\rho_h(u) = \frac{\sum_{c=1}^{C} |F_{LL}^c(u)|^2}{\sum_{b\in\mathcal{B}}\sum_{c=1}^{C} |F_b^c(u)|^2 + \varepsilon}, \quad s.t. \quad \mathcal{B} = \{LL, LH, HL, HH\}. \tag{8}$$

Here, F_{LL} represents slowly varying content, whereas the remaining subbands encode directional details. A large ρ_h indicates low-frequency dominance, but may also arise in naturally smooth

regions. We therefore combine it with the low-frequency context and map the result to a bounded calibration cue:

$$\begin{aligned} P_h &= \sigma(\mathrm{Up}(\phi_{\deg}([\phi_l(F_{LL}), \rho_h]))) \,, \\ A_h(x) &= 1 - \exp(-\zeta_r P_h(x)) \,, \quad s.t. \quad \zeta_r = \mathrm{Softplus}(1) > 0. \end{aligned} \tag{9}$$

Here, ϕ_l and $\phi_{\deg}$ denote pointwise and 3×3 convolutions, respectively. The prompt P_h contextualizes the energy statistic, while the bounded cue A_h represents relative calibration demand.

3.2 Attenuation-Guided Structural Response Calibration

Fog can obscure building boundaries and other cross-view structural cues. To exploit the remaining cues for calibration, a structural branch complements A_h using fixed horizontal (D_x) and vertical (D_y) Sobel operators as gradient priors to form the structural candidate S_g:

$$S_g = \mathrm{PWConv}_{1\times1}(\mathrm{ReLU}(\mathrm{GN}(\mathrm{DWConv}_{3\times3}([\mathcal{D}_x(F_g), \mathcal{D}_y(F_g)])))) \,. \tag{10}$$

The learnable layers aggregate the generic Sobel responses under cross-view matching supervision, yielding geo-localization-relevant structures. Their relative spatial saliency is obtained by channel aggregation and sample-wise standardization:

$$\overline{S}_g = \mathrm{Mean}_c(|S_g|), \quad \widetilde{S}_g = \sigma\left(\frac{\overline{S}_g - \mu_s}{\sigma_s + \varepsilon}\right), \tag{11}$$

where μ_s and σ_s are the spatial mean and standard deviation of $\overline{S}_g$, respectively.

We fuse the individual cues and their explicit interaction to generate the spatial calibration gate, which modulates S_g through a residual update:

$$\begin{aligned} M_g &= \sigma\Big(\phi_{\mathrm{g}}\Big([A_h, \widetilde{S}_g, A_h \odot \widetilde{S}_g]\Big)\Big) \,, \\ \widehat{F}_g &= F_g + \gamma M_g \odot S_g. \end{aligned} \tag{12}$$

Here, ϕ_{g} is a 3×3 convolution, γ is initialized to 0.01 and is learnable, and the single-channel M_g is broadcast over channels. The interaction $A_h \odot \widetilde{S}_g$ highlights regions that are both degraded and structurally informative. WASA is optimized end-to-end using each backbone's original cross-view retrieval objective; further implementation details are provided in *Appendix B*.

4 Experiments

To verify that the constructed synthetic fog data can effectively support supervised training of CVGL models under foggy conditions, we conduct supervised learning experiments on CVUSA-Fog and CVACT-Fog. We evaluate our approach using four representative CVGL models: GeoDTR (Zhang et al. (2023)), Sample4Geo (Deuser et al. (2023)), ConGeo (Mi et al. (2024)), and SinGeo (Chen et al. (2026a)). Each model is trained on the corresponding synthetic fog training set and evaluated on the synthetic fog val. set. Models trained exclusively on the original clear-weather data are used as baselines, allowing us to analyze both the influence of fog degradation on cross-view matching and the ability of fog-domain supervision to adapt existing models to degraded scenes. We further evaluate synthetic-to-real generalization on the real-fog CVFog dataset, examining whether robustness learned from synthetic fog transfers to naturally occurring fog.

Experimental setting. Our experiments were conducted using four V100 GPUs. For Sample4Geo, ConGeo, and SinGeo, the number of epochs was set to 40, whereas for GeoDTR, it was set to 200; the AdamW optimizer was used with a learning rate of 0.0001 and a batch size of 32. For consistency, "Clear" denotes the filtered CVUSA or CVACT subset used to construct the corresponding synthetic fog dataset.

4.1 Results on the CVUSA-Fog and CVACT-Fog datasets

Table 1 shows that fog degradation substantially affects cross-view representation learning. When clear-trained models are directly evaluated on foggy images, all four methods exhibit pronounced R@1 degradation, indicating that fog weakens the ground–satellite correspondences learned from clear scenes. Training on our synthetic fog data consistently recovers

Table 1: Performance under clear and foggy conditions, as well as with the proposed WASA, on CVUSA-Fog and CVACT-Fog.

Methods	Setting		CVUSA-Fog				CVACT-Fog			
	Training	Testing	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
GeoDTR	Clear	Clear	94.11	98.66	99.34	99.89	92.03	98.89	99.23	99.67
	Clear	Syn. Fog	82.23	94.31	96.57	99.34	75.02	89.66	95.21	99.08
	Syn. Fog	Syn. Fog	87.69	94.79	97.85	99.75	88.09	97.85	98.77	99.57
	Syn. Fog + WASA	Syn. Fog	92.59	96.59	99.18	99.89	91.61	98.92	99.20	99.67
Sample4Geo	Clear	Clear	98.18	99.54	99.67	99.87	95.23	99.05	99.28	99.56
	Clear	Syn. Fog	81.25	94.03	95.49	99.83	80.49	92.81	95.47	99.78
	Syn. Fog	Syn. Fog	94.31	98.45	99.32	99.85	92.43	98.78	99.03	99.22
	Syn. Fog + WASA	Syn. Fog	97.93	99.44	99.59	99.87	94.80	98.87	99.10	99.57
ConGeo	Clear	Clear	98.13	99.47	99.67	99.89	93.47	98.67	99.00	99.51
	Clear	Syn. Fog	88.80	97.95	98.66	99.41	82.48	93.61	95.46	98.49
	Syn. Fog	Syn. Fog	95.54	98.92	99.44	99.80	92.38	98.67	99.08	99.61
	Syn. Fog + WASA	Syn. Fog	97.93	99.51	99.72	99.77	93.41	98.81	99.11	99.83
SinGeo	Clear	Clear	98.77	99.51	99.85	99.87	93.53	98.51	99.10	99.66
	Clear	Syn. Fog	87.79	92.00	97.41	99.52	85.44	94.85	96.33	98.30
	Syn. Fog	Syn. Fog	95.25	98.73	99.37	99.83	91.83	97.57	98.10	98.93
	Syn. Fog + WASA	Syn. Fog	96.58	99.10	99.43	99.80	92.75	98.19	98.73	99.01

performance; for example, Sample4Geo improves from 81.25% to 94.31% on CVUSA-Fog. This validates the constructed foggy datasets as effective resources for learning fog-adaptive cross-view representations. Under the same fog-domain training setting, WASA further raises Sample4Geo to 97.93% on CVUSA-Fog and yields consistent gains across all backbones and datasets, demonstrating the effectiveness of degradation-aware structural calibration.

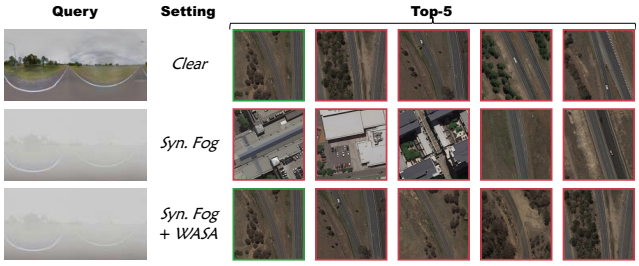


Figure 4: Qualitative Top-5 retrieval results under different settings on CVACT-Fog dataset.

Fig. 4 compares Top-5 retrieval results under clear, foggy, and WASA-enhanced foggy settings. The model retrieved the correct satellite image at Top-1 under clear conditions but failed under fog as weakened textures and distant structures reduced cross-view discriminability. With WASA, the correct Top-1 match was recovered, showing that feature calibration mitigated fog-induced representation degradation and detected discriminative cross-view cues.

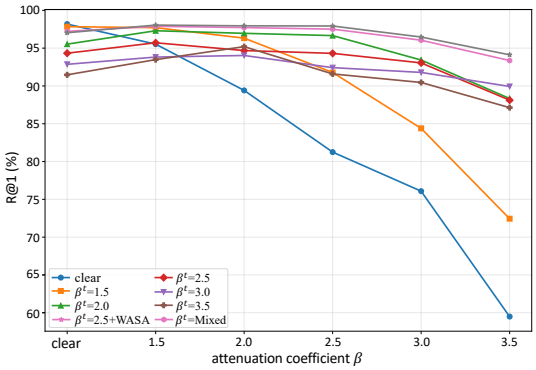


Figure 5: Cross-intensity R@1 performance on CVUSA-Fog. Sample4Geo is trained on clear images, a fixed attenuation coefficient β^t, or a uniform mixture $\beta^t \in \{1.5, 2.5, 3.5\}$, and evaluated at different test fog intensities. WASA is trained with $\beta^t = 2.5$.

4.2 Cross-Intensity Generalization Analysis under Synthetic Fog

As shown in Fig. 5, the clear-trained model degrades rapidly as the test fog intensity increases. Training with a fixed β^t improves robustness around its training severity but generalizes less effectively to stronger fog. Uniform mixed-β training produces a more stable performance across intensities, confirming the benefit of severity-diverse augmentation. Nevertheless, WASA trained only with $\beta^t = 2.5$ consistently outperforms mixed-β training under all fog levels, with a larger advantage under severe fog, while remaining competitive on clear images. These results show WASA provides stronger cross-intensity robustness than mixed-intensity training, supporting the benefit of its structure-aware calibration.

Table 3: Ablation of the degradation and structural cues in WASA on CVUSA-Fog using Sample4Geo. DC., SC., and Inter. indicate whether A_h, $\widetilde{S}_g$, and $A_h \odot \widetilde{S}_g$ are used to construct the calibration gate M_g, respectively.

Variants	DC.	SC.	Inter.	R@1	R@5	R@10	R@1%
Sample4Geo-*fog*	–	–	–	94.31	98.45	99.32	99.85
DC. Only	✓	–	–	95.42	98.93	99.43	99.85
SC. Only	–	✓	–	95.87	99.25	99.53	99.85
w/o Inter.	✓	✓	–	97.31	99.36	99.56	99.86
Full WASA	✓	✓	✓	**97.93**	**99.44**	**99.59**	**99.87**

4.3 Results on the CVFog dataset

Table 2 evaluates the generalization of models trained on clear or synthetic-fog data to the real-fog CVFog dataset. Training with synthetic fog consistently improves all retrieval metrics across the three backbones. Specifically, compared with clear-image training, synthetic-fog training increases R@1 by 3.59%, 6.55%, and 7.81% for Sample4Geo, ConGeo, and SinGeo, respectively. These consistent gains indicate that the synthetic fog degradations provide transferable robustness to real fog and effectively narrow the synthetic-to-real domain gap. Incorporating WASA further improves R@1 over synthetic-fog training alone. In particular, SinGeo equipped with WASA achieves the best performance, reaching 35.27% R@1. Overall, synthetic-fog training establishes an effective basis for real-fog generalization, while WASA provides additional gains through structure-guided feature calibration. Their complementary effects substantially improve retrieval robustness on real fog without using CVFog images for training.

4.4 Ablation Study and Visualization Analysis

Table 2: Zero-shot generalization performance on the real-fog CVFog dataset, without training on CVFog.

Methods	Training	CVFog			
		R@1	R@5	R@10	R@1%
Sample4Geo	Clear	13.67	34.28	45.10	60.12
	Syn. Fog	17.26	41.34	58.83	75.26
	Syn. Fog + WASA	20.08	50.69	67.74	78.67
ConGeo	Clear	19.29	47.53	61.76	71.89
	Syn. Fog	25.84	57.94	70.51	83.60
	Syn. Fog + WASA	30.65	63.92	77.02	91.12
SinGeo	Clear	20.60	45.14	56.16	79.89
	Syn. Fog	28.41	56.02	67.69	86.96
	Syn. Fog + WASA	35.27	64.60	82.24	93.30

Ablation of the WASA components. Table 3 evaluates the individual and joint effects of the degradation and structural cues. Using DC. or SC. alone improves R@1 over the fog-trained baseline, with SC. producing the larger gain, indicating that structural responses provide stronger standalone evidence for cross-view matching. Combining both cues without their interaction further raises R@1 to 97.31%, confirming their complementarity. The full model achieves the best R@1 of 97.93%, showing that explicitly modeling their spatial co-occurrence further improves the calibration of fog-affected structural regions.

Qualitative analysis of WASA responses. Fig. 6 compares model responses and Top-1 retrieval results with and without WASA under fog degradation. Without WASA, spatially dispersed responses led to an incorrect satellite match. WASA concentrated responses around structures relevant to cross-view correspondence and discovered the correct match, indicating that it selectively calibrates fog-degraded ground-view features to exploit discriminative structural information more reliably.

Comparison with generic modules. Fig. 7 compares WASA with CBAM (Woo et al. (2018)), a Conv Adapter (Chen et al. (2024)) and a DehazeFormer-based (Song et al. (2023)) dehazing module. WASA achieves the best retrieval performance with comparable parameter and computational costs. Compared with DehazeFormer, which focuses on restoring image appearance, WASA obtains higher accuracy with lower latency. This suggests that, in the evaluated setting, task-oriented structural calibration is more effective for fog-robust CVGL than image dehazing before feature extraction.

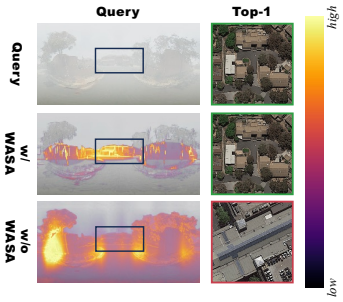


Figure 6: Qualitative visualization of WASA responses and associated Top-1 retrieval results.

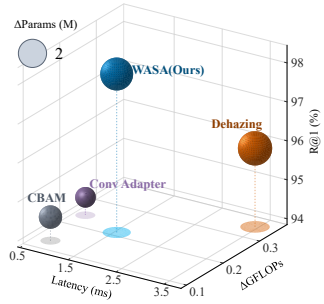


Figure 7: Comparison with generic modules and WASA on CVUSA-Fog.

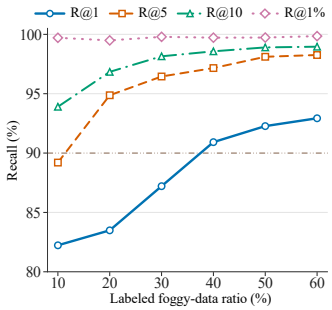


Figure 8: Effect of CVUSA-Fog data ratio in fine-tuning a clear Sample4Geo model.

Fog-domain adaptation under different labeled-data ratios. Fig. 8 reports supervised adaptation using different proportions of synthetic fog data. As the labeled ratio increases from 10% to 60%, R@1 steadily improves, confirming that the constructed data provides effective supervision for learning fog-adaptive cross-view representations. Notably, with only 40% of the labeled fog-domain data, R@1 already exceeds 90%. These results validate the training value of synthetic fog data.

Response-performance correlation analysis. Fig. 9 analyzes the internal behavior of WASA by examining the relationship between its response statistics and improvements in retrieval performance. As shown in Fig. 9 (Left), Δm exhibits an overall increasing trend as fog intensity increases, indicating that WASA provides more pronounced gains under severe degradation. Fig. 9 (Right) reports the correlations between the response statistics and Δm. The interaction term and spatial gate M_g exhibit substantially stronger correlations with Δm than either A_h or $\widetilde{S}_g$ alone. The stronger association between the interaction response and retrieval-margin improvement suggests WASA enhances cross-view matching by identifying matching-relevant structural regions through interactive modeling.

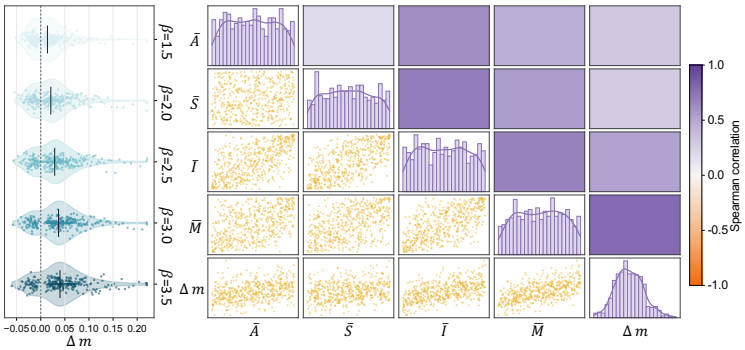


Figure 9: Left: Distribution of Δm under different fog intensities. Right: Spearman correlation matrix between response statistics and Δm. $\Delta m = m_{\text{WASA}} - m_{\text{Base}}$ denotes the change in retrieval margin after applying WASA, where m is defined as the similarity gap between the GT satellite image and the most similar negative satellite image.

4.5 Semi-Supervised Real-Fog Adaptation

Synthetic-fog training improves real-fog generalization, but domain differences remain and real-fog pairings are costly to obtain. We investigate whether pair-unlabeled real-fog images can further improve adaptation under limited pairing supervision, using a fixed 7:1:2 CVFog train/validation/test split. For each budget $r \in \{5\%, 10\%, 20\%\}$, only r of the real-fog training correspondences is exposed, while the remaining ground images form a pair-unlabeled pool whose satellite images stay in the shuffled training gallery. Starting from a common Sample4Geo+WASA model trained on the complete paired synthetic-fog set, we adapt a teacher T_r using the synthetic pairs and the exposed real-fog pairs, with the satellite encoder frozen. For each pair-unlabeled query, the fixed teacher scores the complete real-fog training gallery under two weakly augmented views. Let $\hat{k}_j^{(v)}$ and $m_j^{(v)}$ denote the Top-1 satellite index and its Top-1–Top-2 similarity margin for view v. A pseudo-pair is retained only if

$$c_j = \mathbb{I}\left[\hat{k}_j^{(1)} = \hat{k}_j^{(2)}\right] \mathbb{I}\left[\min\left(m_j^{(1)}, m_j^{(2)}\right) \geq \delta\right]. \tag{13}$$

Here, $\delta = 0.006$ is a similarity-margin threshold calibrated on the CVFog validation split and fixed across all label budgets. The teacher accesses neither withheld pair identities nor GPS co-

Table 4: Real-fog adaptation with supervised and semi-supervised settings across pairing-label budgets. All settings use the complete paired synthetic-fog set. Pair-Unlabeled Access denotes the candidate pool before confidence filtering; Real-Sup excludes it. Partial-budget results report mean $\pm$ standard deviation over three labeled-subset samplings.

Method	Labeled Real	Pair-Unlabeled Access	R@1	R@5	R@10
Real-Sup	5%	0%	30.56 ± 0.71	60.01 ± 0.56	71.92 ± 0.31
RealFog-SSL-Hard	5%	95%	31.28 ± 0.53	64.34 ± 0.91	76.34 ± 0.78
Real-Sup	10%	0%	31.83 ± 0.72	65.09 ± 0.78	75.04 ± 0.31
RealFog-SSL-Hard	10%	90%	34.75 ± 0.81	68.17 ± 0.33	83.55 ± 0.12
Real-Sup	20%	0%	35.45 ± 0.40	67.64 ± 0.46	80.82 ± 0.35
RealFog-SSL-Hard	20%	80%	42.87 ± 0.64	73.25 ± 0.52	83.64 ± 0.35
Full Real-Sup	100%	0%	53.40	79.18	87.86

ordinates and never inserts the ground-truth satellite during retrieval. The student is initialized from T_r and learns from the synthetic pairs, exposed real-fog pairs, and retained pseudo-pairs. Real-Sup and RealFog-SSL-Hard use identical supervised data, initialization, training iterations, and model-selection criteria; their only difference is the use of pair-unlabeled real-fog images. The complete protocol and more experiments are provided in *Appendices F, G* and *H*.

Main results. Table 4 compares semi-supervised adaptation under identical real-fog pairing-label budgets. At $r = 10\%$, RealFog-SSL-Hard improves R@1 from 31.83% to 34.75%, corresponding to an absolute gain of 2.92%. The corresponding gains under the 5% and 20% budgets are 0.72% and 7.42%, respectively. The increasing gains suggest a complementary effect: labeled pairs establish a real-fog adaptation anchor, while confidence-filtered pseudo-pairs provide additional supervision from pair-unlabeled images, improving retrieval under the same real-fog pairing-label budget.

5 Related Work

CVGL matches ground-view and satellite-view images by learning cross-view structural correspondences. Existing methods exploit spatial aggregation and geometric alignment (Chen et al. (2026c); Zhu et al. (2022); Zhang et al. (2023)), while recent work separates scene content from viewpoint variations (Li et al. (2025)). Despite their success on standard datasets, these methods largely rely on structural cues that fog can obscure, challenging reliable cross-view matching.

Related studies improve robustness to test-time corruptions (Wu et al. (2026)) or learn weather-aware drone–satellite representations through restoration (Xu et al. (2026); Feng et al. (2024)) and textual guidance (Wen et al. (2025)). We complement these efforts by studying explicit fog-domain training and real-fog transfer for ground–satellite matching. Using atmospheric scattering for controllable, depth-dependent synthesis (Yu et al. (2026); Li et al. (2026b); Huang et al. (2026)), we construct synthetic fog datasets for supervised training. Together with a curated real-fog dataset, these support our investigation of retrieval-oriented structural calibration and semi-supervised adaptation under limited real-fog pairing supervision.

6 Discussion and Conclusion

Mapillary sequences create near-duplicate ground views and overlapping satellite crops. Distributing them across partitions can cause spatial leakage and inflate retrieval performance. We remove consecutive or near-duplicate ground images and pairs with substantially overlapping satellite coverage before partitioning CVFog. The semi-supervised protocol uses the cleaned splits without geographic metadata or location-based candidate restrictions during training or retrieval.

This work studies fog-robust CVGL under synthetic and real fog. We construct CVUSA-Fog and CVACT-Fog with controllable fog and CVFog with manually verified real-fog pairs. Synthetic-fog training improves fog-domain retrieval and synthetic-to-real generalization, while WASA provides further gains across backbones, datasets, and fog intensities. Confidence-filtered hard pseudo-pairs also improve real-fog adaptation under limited supervision. Together, these results demonstrate the value of synthetic fog datasets, structure-aware adaptation, and pair-unlabeled real-fog data for robust CVGL under varying degradation levels and limited pairing supervision.

7 AI USE STATEMENT

In this work, we used generative AI tools for language editing and polishing to improve the grammar, readability, and linguistic clarity of the manuscript. We take responsibility for the final content of this work, including the AI-assisted text.

REFERENCES

Hao Chen, Ran Tao, Han Zhang, Yidong Wang, Xiang Li, Wei Ye, Jindong Wang, Guosheng Hu, and Marios Savvides. Conv-adapter: Exploring parameter efficient transfer learning for convnets. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1551–1561. IEEE, 2024.

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

Zhongwei Chen, Hai-Jun Rong, Longrun Xu, Jie Lu, Zhuo Zou, Zhao-Xu Yang, and Guoqi Li. Efficient spike-driven transformer for high-performance drone-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2026b. doi: 10.1109/TCSVT.2026.3716422.

Zhongwei Chen, Zhao-Xu Yang, Hai-Jun Rong, Jiawei Lang, and Guoqi Li. From limited labels to open domains: An efficient learning method for drone-view geo-localization. *IEEE Transactions on Multimedia*, pp. 1–13, 2026c. doi: 10.1109/TMM.2026.3716759.

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.

Tongtong Feng, Qing Li, Xin Wang, Mingzi Wang, Guangyao Li, and Wenwu Zhu. Multi-weather cross-view geo-localization using denoising diffusion models. In *Proceedings of the 2nd Workshop on UAVs in Multimedia: Capturing the World from a New Perspective*, UAVM '24, pp. 35–39, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400712067. doi: 10.1145/3689095.3689103. URL https://doi.org/10.1145/3689095.3689103.

Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(6):1397–1409, 2013. doi: 10.1109/TPAMI.2012.213.

Xiu-Wen Huang, Qing-Ling Shu, Kai-Xuan Jiang, Wei Lu, Si-Bao Chen, Jin Tang, and Bin Luo. Foggy-uavcd and dsrf-net: Benchmarking robust satellite–uav change detection under fog degradation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 241:213–230, 2026.

Ke Li, Di Wang, Xiaowei Wang, Zhihong Wu, Yiming Zhang, Yifeng Wang, and Quan Wang. Robust drone-view geo-localization via content-viewpoint disentanglement, 2025. URL https://arxiv.org/abs/2505.11822.

Lin Li, Ru Lei, Kun Zhang, Lingchen Sun, and Rustam Stolkin. Neural network-based framework for wide visibility dehazing with synthetic benchmarks. *Pattern Recognition*, 175:113056, 2026a. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2026.113056. URL https://www.sciencedirect.com/science/article/pii/S0031320326000191.

Yanying Li, Jinyang Li, Shengfeng He, Yangyang Xu, Junyu Dong, and Yong Du. Nimbusgs: Unified 3d scene reconstruction under hybrid weather, 2026b. URL https://arxiv.org/abs/2603.27228.

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.

Tianyi Lyu, Mingye Ju, and Kai-Kuang Ma. Disentanglement-wise image dehazing through cross-domain manifold consensus. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22733–22743, 2026.

Li Mi, Chang Xu, Javiera Castillo-Navarro, Syrielle Montariol, Wen Yang, Antoine Bosselut, and Devis Tuia. Congeo: Robust cross-view geo-localization across ground view variations. In *European Conference on Computer Vision*, pp. 214–230. Springer, 2024.

Mohamed Ahmed Mohamed and Xiaowei Huang. A data efficiency study of synthetic fog for object detection using the clear2fog pipeline. *Neurocomputing*, 703:134798, 2026. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2026.134798. URL https://www.sciencedirect.com/science/article/pii/S092523122602196X.

Christos Sakaridis, Dengxin Dai, Simon Hecker, and Luc Van Gool. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In *Proceedings of the european conference on computer vision (ECCV)*, pp. 687–704, 2018a.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9):973–992, 2018b.

Maxim Shugaev, Ilya Semenov, Kyle Ashley, Michael Klaczynski, Naresh Cuntoor, Mun Wai Lee, and Nathan Jacobs. Arcgeo: Localizing limited field-of-view images using cross-view matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 209–218, 2024.

Yuda Song, Zhuqing He, Hui Qian, and Xin Du. Vision transformers for single image dehazing. *Trans. Img. Proc.*, 32:1927–1941, January 2023. ISSN 1057-7149. doi: 10.1109/TIP.2023.3256763. URL https://doi.org/10.1109/TIP.2023.3256763.

Jie Sun, Heng Liu, Yongzhen Wang, Xiao-Ping Zhang, and Mingqiang Wei. Wdmamba: When wavelet degradation prior meets vision mamba for image dehazing. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(3):2894–2907, 2026. doi: 10.1109/TCSVT.2025.3614173.

Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, and Chia-Wen Lin. Phatnet: A physics-guided haze transfer network for domain-adaptive real-world image dehazing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5591–5600, 2025.

ZhiYuan Wang, Hao Weng, RongXin Tang, Can Su, HongJie Wang, Jie Tan, GuoHong Dai, Wei Tu, LiXin Gan, and WeiChao Yan. Image dehazing with self-supervised physical consistency via atmospheric scattering model-informed feature extraction network. *Optics and Lasers in Engineering*, 200:109613, 2026.

Jiahao Wen, Hang Yu, and Zhedong Zheng. Weatherprompt: Multi-modality representation learning for all-weather drone visual geo-localization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=RjZUldI6Qk.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision – ECCV 2018*, pp. 3–19, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01234-2.

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 Fortieth AAAI Conference on Artificial Intelligence and Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence and Sixteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'26/IAAI'26/EAAI'26. AAAI Press, 2026. ISBN 978-1-57735-906-7. doi: 10.1609/aaai.v40i13.38043. URL https://doi.org/10.1609/aaai.v40i13.38043.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

Songtianhao Xu, Zhongwei Chen, Zhao-Xu Yang, and Weifeng Wang. A unified benchmark and modality-adaptive network for day-and-night drone-view geo-localization, 2026. URL https://arxiv.org/abs/2607.25778.

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

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024.

Jinze Yu, Yiqun Wang, Aiheng Jiang, Zhengda Lu, Jianwei Guo, Yong Li, Hongxing Qin, and Xiaopeng Zhang. Dehazegs: seeing through fog with 3d gaussian splatting. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence and Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence and Sixteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'26/IAAI'26/EAAI'26. AAAI Press, 2026. ISBN 978-1-57735-906-7. doi: 10.1609/aaai.v40i14.38205. URL https://doi.org/10.1609/aaai.v40i14.38205.

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.

Qianxi Zhang, Tao Gao, Ting Chen, Yuanbo Wen, Ziqi Li, and Shan Liang. A wavelet-guided and physics-aware network for remote sensing image dehazing. *Expert Systems with Applications*, pp. 132456, 2026a.

Weidong Zhang, Baiqiang Yu, Wenyi Zhao, Zheng Liang, Peixian Zhuang, and Keran Zhu. Underwater image enhancement via intelligent optimized multi-exposure image fusion. *IEEE Transactions on Image Processing*, 35:6488–6501, 2026b. doi: 10.1109/TIP.2026.3702354.

Xiaohan Zhang, Xingyu Li, Waqas Sultani, Yi Zhou, and Safwan Wshah. Cross-view geo-localization via learning disentangled geometric layout correspondence. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 3480–3488, 2023.

Hu Zhao, Keyan Ren, Tianyi Yue, Chun Zhang, and Shuai Yuan. Transfg: A cross-view geo-localization of satellite and uavs imagery pipeline using transformer-based feature aggregation and gradient guidance. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–12, 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.

APPENDIX

A GROUND-VIEW IMAGE SCREENING

A.1 PURPOSE AND SCOPE

The fog synthesis pipeline relies on sky segmentation, relative depth, atmospheric light, and pixel-wise transmittance. Exposure, image-integrity, and panorama-completeness issues can affect these estimates and introduce synthesis artifacts. We therefore screen the source ground-view images of CVUSA and CVACT to obtain subsets suitable for controlled fog synthesis.

A.2 EXCLUSION CRITERIA

A ground-view image is excluded if it satisfies at least one of the following conditions and the corresponding artifact affects a substantial portion of the panorama or prevents the dominant scene layout from being reliably observed. As shown in Fig. 1 of the main paper.

Table 5: Dataset statistics after screening and repartitioning. Train and Val. denote the newly constructed splits of the retained pairs.

Datasets	Original	Excluded	Train	Val.	Retained	Ratio
CVUSA-Fog	44,415	13,930	24,385	6,100	30,485	68.6%
CVACT-Fog	44,415	13,552	24,763	6,100	30,863	69.5%

Severe exposure or illumination failure. This category includes extensive overexposure or underexposure, strong lens flare, widespread saturation or blackout, and abrupt nonphysical exposure discontinuities across the panorama. Such failures remove recoverable scene information and may result in unstable sky masks or atmospheric-light estimates.

Image corruption. Images containing decoding errors, abnormal color or noise patterns, severe defocus or motion blur, or substantial panorama-stitching artifacts are excluded. These artifacts can produce unreliable local structures and discontinuous relative-depth predictions.

Incomplete panorama content. Images with large blank or black regions, missing panorama sectors, severe cropping, or extensive camera-body obstruction are excluded when the remaining visual content is insufficient to recover the scene layout required by the synthesis pipeline.

A.3 Screening Procedure

We first screen the complete source pool used to construct each dataset, following the criteria in Section A.2. Excluding a ground-view panorama also excludes its associated ground–satellite pair. The original satellite counterpart of each retained sample are preserved.

The retained pairs are then repartitioned into training and validation sets, with each ground-view image and its satellite counterpart assigned to the same split. These new split assignments are shared by the clear condition and all synthetic fog intensities $\beta \in \{1.5, 2.0, 2.5, 3.0, 3.5\}$, ensuring that comparisons use the same training and validation locations.

A.4 Dataset Statistics

Table 5 summarizes the screening and repartitioning statistics. Of the 44,415 source pairs in each dataset, 30,485 are retained for CVUSA-Fog and 30,863 for CVACT-Fog. The new splits contain 24,385 training and 6,100 validation pairs for CVUSA-Fog, and 24,763 training and 6,100 validation pairs for CVACT-Fog.

A.5 Limitations of Manual Screening

Manual screening involves subjective judgments, particularly for borderline samples, so an independent screening process may yield a different retained subset.

B WASA Insertion and Implementation Details

B.1 Insertion and Backbone Adaptation

For all evaluated backbones, a single WASA block is inserted after the first stage of the ground-view encoder, immediately before the subsequent downsampling operation. Let $E_g^{(1)}$ denote the ground-view encoder up to the first-stage output and $E_g^{(2:L)}$ denote the remaining stages. The ground-view descriptor is computed as

$$F_g = E_g^{(1)}(g), \qquad \widehat{F}_g = \mathrm{WASA}(F_g), \qquad z_g = E_g^{(2:L)}(\widehat{F}_g). \tag{14}$$

We use the first-stage output to retain spatial detail for local wavelet and gradient analysis. Later stages have lower spatial resolution and more abstract representations, which may limit fine-grained

structural calibration. Early insertion also allows the subsequent backbone stages to integrate the calibrated responses into the final retrieval representation.

The same relative insertion rule is applied to GeoDTR, Sample4Geo, ConGeo, and SinGeo. Only the channel-dependent projection layers inside WASA are adjusted to match the feature dimension of each backbone. No WASA block is added to the satellite-view branch, whose original architecture is retained.

B.2 Wavelet and Structural Branch Details

The single-level Haar DWT is applied independently to each feature channel, producing four subbands at half the input spatial resolution. In Eq. 8 of the main paper, F_{LL} describes slowly varying content, whereas F_{LH}, F_{HL}, and F_{HH} encode directional details. The channel-wise subband energies are aggregated to obtain the low-frequency concentration ρ_h.

In Eq. 9 of the main paper, ϕ_l is a 1×1 pointwise convolution and $\phi_{\deg}$ is a 3×3 convolution that produces a single-channel response. The upsampling operator restores the response to the spatial resolution of F_g, and σ denotes Sigmoid. Consequently, $P_h, A_h \in [0,1]^{1\times H\times W}$.

We interpret P_h as a relative degradation prompt rather than an absolute fog-probability estimate. The Softplus parameterization guarantees $\zeta_r > 0$, while the exponential mapping bounds the dynamic range of A_h.

The horizontal and vertical Sobel operators are fixed and applied channel-wise. For clarity, the compact structural formulation in Eq. 10 of the main paper can be expanded as

$$\begin{aligned} G_g &= [\mathcal{D}_x(F_g), \mathcal{D}_y(F_g)], \\ H_g &= \mathrm{DWConv}_{3\times3}(G_g), \\ S_g &= \mathrm{PWConv}_{1\times1}\left(\mathrm{ReLU}\left(\mathrm{GN}(H_g)\right)\right). \end{aligned} \tag{15}$$

The depthwise convolution aggregates local gradient patterns, group normalization standardizes their responses, and the pointwise convolution maps them to $S_g \in \mathbb{R}^{C\times H\times W}$. Although the Sobel filters provide generic boundary priors, the subsequent learnable layers are optimized through cross-view retrieval supervision and therefore learn task-relevant structural responses.

Structural saliency is computed by channel aggregation and sample-wise spatial standardization:

$$\begin{aligned} \overline{S}_g &= \mathrm{Mean}_c(|S_g|), \\ \mu_s &= \mathrm{Mean}_{h,w}(\overline{S}_g), \qquad \sigma_s = \mathrm{Std}_{h,w}(\overline{S}_g), \\ \widetilde{S}_g &= \sigma\left(\frac{\overline{S}_g - \mu_s}{\sigma_s + \epsilon}\right). \end{aligned} \tag{16}$$

Here, $\epsilon > 0$ ensures numerical stability. Thus, $\widetilde{S}_g$ represents the relative structural saliency within each image rather than an absolute response magnitude. Both $\widetilde{S}_g$ and the resulting calibration gate M_g have shape $1 \times H \times W$. The gate is broadcast along the channel dimension when applied to S_g.

The learnable residual scale γ in Eq. 12 of the main paper is initialized to 10^{-2}. This initialization keeps WASA close to an identity mapping at the beginning of training while preserving gradient propagation through both branches.

B.3 Retrieval Objective

WASA introduces no auxiliary degradation or structural supervision. Each backbone retains the retrieval objective used by its corresponding baseline, and WASA is optimized jointly with the retrieval model. For the Sample4Geo-based configuration, the ground-to-satellite contrastive term for a mini-batch of B matched pairs is

$$\mathcal{L}_{g\to s} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\left(\mathrm{sim}(z_g^i, z_s^i)/\tau\right)}{\sum_{j=1}^{B} \exp\left(\mathrm{sim}(z_g^i, z_s^j)/\tau\right)}, \tag{17}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity and $\tau = 0.07$. The matched descriptors form the positive pair, whereas the remaining satellite descriptors in the mini-batch serve as negatives. When the

original backbone objective contains a symmetric satellite-to-ground term or other retrieval terms, these components are retained without modification.

B.4 Placement of Comparison Modules

For the generic-module comparison, CBAM and Conv Adapter are inserted at the same ground-view feature location as WASA, ensuring that all three modules receive features with identical spatial resolution and channel width. DehazeFormer instead processes the foggy ground-view RGB image before backbone feature extraction and therefore represents an image-restoration baseline rather than a feature-space adapter. The remaining retrieval pipeline is left unchanged.

C Computation of the Luminance Term L and the Gray-White Atmospheric Light Constraint

We estimate A_{sky} as the mean RGB value of the brightest 5% of sky pixels selected by luminance. To reduce the influence of sky-color variations, we further construct a gray-white atmospheric light component with consistent luminance. Let

$$A_{\mathrm{sky}} = \left[A^R_{\mathrm{sky}}, A^G_{\mathrm{sky}}, A^B_{\mathrm{sky}}\right], \tag{18}$$

where A^R_{sky}, A^G_{sky}, and A^B_{sky} denote the values of the initial atmospheric light in the three RGB channels, respectively. Its perceived luminance is first calculated using the standard luminance transformation:

$$L_0 = 0.299A^R_{\mathrm{sky}} + 0.587A^G_{\mathrm{sky}} + 0.114A^B_{\mathrm{sky}}. \tag{19}$$

The channel weights reflect the different contributions of the RGB components to perceived luminance. Since the original sky color may be affected by a blue cast under clear skies, local overexposure, or segmentation errors, directly using L_0 may result in atmospheric light that is excessively dark or bright. We therefore constrain it to a predefined luminance range:

$$L = \mathrm{clip}\left(L_0, L_{\min}, L_{\max}\right), \tag{20}$$

where $L_{\min}$ and $L_{\max}$ denote the minimum and maximum luminance allowed for the gray-white atmospheric light and are set to 0.82 and 0.90, respectively. Based on the clipped luminance L, the neutral gray-white atmospheric light is constructed as

$$A_{\mathrm{gray}} = [L, L, L]. \tag{21}$$

Assigning identical values to the three RGB channels produces a gray-white color without an evident chromatic bias. Consequently, A_{gray} suppresses abnormal color casts caused by variations in sky color while preserving the overall luminance level of the current image. Finally, A_{gray} serves as a gray-white constraint on A_{sky}, producing a more stable atmospheric light estimate:

$$A = \alpha A_{\mathrm{sky}} + (1-\alpha)A_{\mathrm{gray}}, \tag{22}$$

where $\alpha = 0.7$ controls the amount of original sky-color information retained. This constraint allows the synthesized foggy images to maintain consistent illumination while reducing bluish or otherwise color-biased fog effects.

D Determination of $d_{\min}$

We determine $d_{\min}$ according to the desired minimum scattering contribution in near-field regions. Let η denote the minimum fog contribution in the near field. We then have

$$1 - \exp\left(-\beta_{\mathrm{ref}} d_{\min}\right) = \eta, \tag{23}$$

where β_{ref} denotes the reference fog attenuation coefficient. Accordingly, $d_{\min}$ can be derived as

$$d_{\min} = \frac{-\ln(1-\eta)}{\beta_{\mathrm{ref}}}. \tag{24}$$

During dataset construction, we aim to preserve substantial scene structure in near-field regions under the reference fog intensity while introducing a moderate atmospheric scattering effect. We therefore set the minimum near-field scattering contribution to approximately 14% and adopt $\beta_{\mathrm{ref}} = 1.5$, yielding

$$d_{\min} = \frac{-\ln(1 - 0.14)}{1.5} \approx 0.10. \tag{25}$$

Accordingly, we set $d_{\min} = 0.10$.

E Determination of the Fog Attenuation Coefficient β

In the atmospheric scattering model, transmittance is jointly determined by scene depth and the fog attenuation coefficient:

$$t(x) = \exp\big(-\beta d(x)\big). \tag{26}$$

The coefficient β controls the rate at which transmittance decreases with depth. In general, a larger β produces lower transmittance and therefore stronger fog degradation. It should be noted that the monocular depth estimation model used in this work produces relative depth, which is further normalized to the interval $[0, 1]$. Consequently, β does not correspond to a strict physical visibility measure defined in metric depth space. Instead, it serves as a fog attenuation coefficient in the normalized depth space and controls the degree of visibility degradation in the synthesized images.

For a normalized depth map, a far-field region approximately satisfies

$$d(x) \approx 1. \tag{27}$$

Its transmittance can therefore be expressed as

$$t_{\mathrm{far}} = \exp(-\beta). \tag{28}$$

Accordingly, given a desired far-field transmittance t_{far}, the corresponding attenuation coefficient can be derived as

$$\beta = -\ln\left(t_{\mathrm{far}}\right). \tag{29}$$

Based on this relationship, we determine the fog attenuation coefficients by specifying different far-field transmittance levels. A larger far-field transmittance corresponds to weaker fog degradation, whereas a smaller value represents more severe visibility attenuation in distant regions. For example, when $\beta = 3.5$, $t_{\mathrm{far}} = \exp(-3.5) \approx 0.030$, indicating that the appearance of far-field regions is predominantly governed by the atmospheric light component. We therefore adopt the following coefficients to cover visibility degradation ranging from relatively mild to severe: $\beta \in \{1.5, 2.0, 2.5, 3.0, 3.5\}$.

In addition, a minimum depth offset is introduced before calculating the transmittance:

$$d'(x) = d_{\min} + (1 - d_{\min})\, d(x). \tag{30}$$

The corrected depth is then used to calculate the transmittance:

$$t(x) = \exp\big(-\beta d'(x)\big). \tag{31}$$

This design assigns a nonzero effective propagation distance to near-field regions, preventing them from completely retaining their original appearance during fog synthesis. Since $d_{\min} = 0.10$, the transmittance at the minimum corrected depth is $t_{\mathrm{near}} = \exp\big(-\beta d_{\min}\big)$.

When $\beta = 1.5$, $t_{\mathrm{near}} = \exp(-1.5 \times 0.10) \approx 0.861$. Thus, near-field regions retain approximately 86.1% of their original image content while receiving a moderate baseline scattering effect. When $\beta = 3.5$, $t_{\mathrm{near}} = \exp(-3.5 \times 0.10) \approx 0.705$.

Even under severe fog degradation, near-field regions retain approximately 70.5% of their original structural information. This setting strengthens weather-induced degradation in distant regions while preventing geographically discriminative cues, such as roads, vehicles, and nearby buildings, from being excessively obscured.

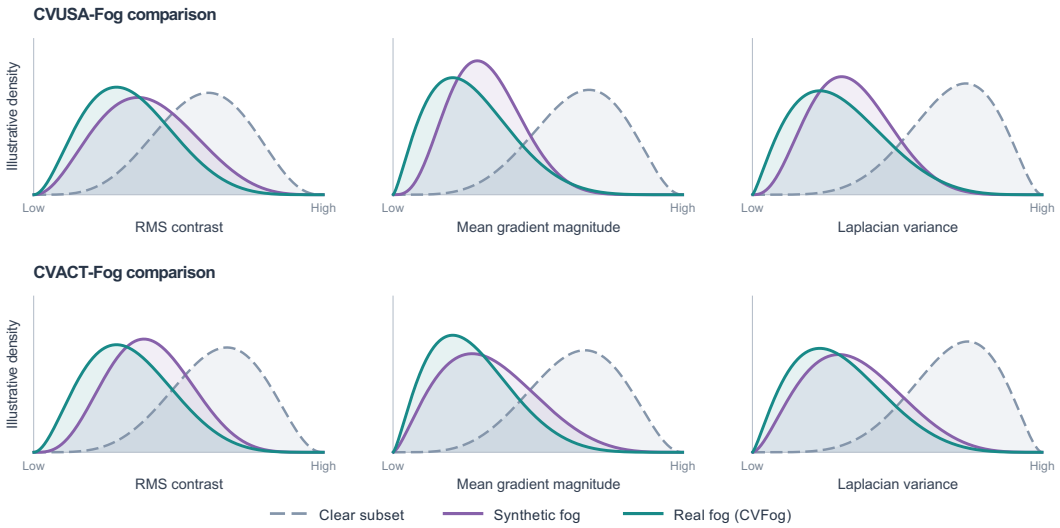


Figure 10: Illustration of image-statistic comparisons between clear, synthetic-fog, and real-fog images.

In summary, the values of β used in this work are not directly adopted from physical visibility parameters defined for metric depth. Instead, they are determined according to far-field transmittance in the normalized depth space. This strategy maintains the controllability of the fog synthesis process while covering multiple levels of visibility degradation, thereby providing a data foundation for analyzing the robustness of cross-view geo-localization models.

Illustration of fog-related image statistics. Fig. 10 illustrates a comparison of clear, synthetic-fog, and real-fog images using three complementary statistics. RMS contrast measures intensity variation, mean gradient magnitude characterizes local intensity transitions, and Laplacian variance serves as a sharpness proxy. The illustrated shift toward lower values represents contrast attenuation and reduced local detail. Partial overlap, together with differences in peak location and spread, illustrates how shared appearance characteristics and remaining distribution differences can be examined when comparing synthetic fog with CVFog.

F Additional Details of Semi-Supervised Real-Fog Adaptation

F.1 Real-Fog Data Protocol

The curated CVFog dataset contains 1,441 ground–satellite pairs collected from Mapillary. It is divided into fixed, mutually exclusive training, validation, and test splits with approximate ratios of 70%, 10%, and 20%, respectively, with each ground image and its paired satellite image assigned to the same split. Before partitioning, we remove consecutive and near-duplicate ground images as well as pairs with substantially overlapping satellite coverage. This prevents overly similar ground or satellite content from being distributed across different splits.

Let

$$\mathcal{D}_S = \{(g_i^{\mathrm{syn}}, s_i^{\mathrm{syn}})\}_{i=1}^{N_S} \tag{32}$$

denote the fully paired synthetic-fog training set, and let

$$\mathcal{D}_R^{\mathrm{tr}} = \{(g_i^{r}, s_i^{r})\}_{i=1}^{N_R^{\mathrm{tr}}} \tag{33}$$

denote the real-fog training set. For each label budget $r \in \{5\%, 10\%, 20\%\}$, we define

$$\begin{aligned} \mathcal{D}_R^L(r) &= \{(g_i^r, s_i^r) \mid i \in \mathcal{I}_l\}, \\ \mathcal{D}_R^U(r) &= \{g_j^r \mid j \in \mathcal{I}_u\}, \\ r &= \frac{|\mathcal{I}_l|}{N_R^{\mathrm{tr}}}, \qquad \mathcal{I}_l \cap \mathcal{I}_u = \varnothing, \\ \mathcal{I}_l \cup \mathcal{I}_u &= \{1, \ldots, N_R^{\mathrm{tr}}\}. \end{aligned} \tag{34}$$

For each random sampling seed, the labeled subsets are nested:

$$\mathcal{I}_l^{5\%} \subset \mathcal{I}_l^{10\%} \subset \mathcal{I}_l^{20\%}. \tag{35}$$

This construction ensures that differences between successive budgets result from additional labeled pairs rather than unrelated subset compositions. Each partial-budget experiment is repeated using three labeled-subset samplings, while the validation and test splits remain fixed.

For each pair-unlabeled query, its satellite image remains in the shuffled training gallery, but its GPS coordinate, and pairing metadata are withheld during pseudo-pair generation and student optimization. The complete correspondences are used only for post-hoc evaluation of pseudo-pair quality. Neither geographic-group metadata nor location information is used during training or retrieval.

F.2 Training Stages

The complete procedure contains three stages.

Synthetic-fog pretraining. Sample4Geo equipped with WASA is trained on $\mathcal{D}_S$, generated with $\beta = 2.5$, to obtain a common source initialization M_S shared by all real-fog label-budget experiments.

Teacher adaptation. For each budget r, M_S is adapted using $\mathcal{D}_S$ and $\mathcal{D}_R^L(r)$:

$$\mathcal{L}_T = \mathcal{L}_{\mathrm{syn}} + \mathcal{L}_{\mathrm{real}}^L. \tag{36}$$

Here, $\mathcal{L}_{\mathrm{syn}}$ and $\mathcal{L}_{\mathrm{real}}^L$ are the standard retrieval losses on the synthetic-fog and exposed real-fog pairs, respectively. The satellite encoder is frozen, whereas the ground-view encoder, ground-view projection layers, and WASA are updated, producing the fixed teacher $T_r = (T_g, T_s)$.

Student adaptation. After pseudo-pair generation, both Real-Sup and RealFog-SSL-Hard are initialized from the same T_r. Real-Sup continues training with only $\mathcal{D}_S$ and $\mathcal{D}_R^L(r)$, whereas RealFog-SSL-Hard additionally uses the retained pseudo-pairs. The two settings use identical supervised data, initialization, optimization schedules, training iterations, and model-selection criteria. Thus, their only difference is access to pair-unlabeled real-fog images.

F.3 Pseudo-Pair Generation

Let $\mathcal{G}_R^{\mathrm{tr}} = \{s_k^r\}_{k=1}^{N_R^{\mathrm{tr}}}$ denote the complete shuffled satellite gallery of the real-fog training split. For each pair-unlabeled query g_j^r, the fixed teacher independently processes two weakly augmented views $A_w^{(1)}(g_j^r)$ and $A_w^{(2)}(g_j^r)$. The similarity between view v and candidate k is

$$a_{j,v}^T(k) = \mathrm{sim}\Big(\mathrm{Norm}\Big(T_g\Big(A_w^{(v)}(g_j^r)\Big)\Big), \mathrm{Norm}(T_s(s_k^r))\Big), \quad v \in \{1, 2\}. \tag{37}$$

For each view, the Top-1 and Top-2 indices are

$$\begin{aligned} \hat{k}_j^{(v)} &= \arg\max_k a_{j,v}^T(k), \\ \widetilde{k}_j^{(v)} &= \arg\max_{k \neq \hat{k}_j^{(v)}} a_{j,v}^T(k), \end{aligned} \tag{38}$$

and their similarity margin is

$$m_j^{(v)} = a_{j,v}^T\Big(\hat{k}_j^{(v)}\Big) - a_{j,v}^T\Big(\widetilde{k}_j^{(v)}\Big). \tag{39}$$

Following Eq. 13 of the main paper, we set $c_j = 1$ if both weak views select the same Top-1 satellite and $\min_{v\in\{1,2\}} m_j^{(v)} \geq \delta$, and $c_j = 0$ otherwise. When $c_j = 1$, let $\hat{k}_j = \hat{k}_j^{(1)} = \hat{k}_j^{(2)}$ and $\hat{s}_j^r = s_{\hat{k}_j}^r$. The retained set is

$$\widehat{\mathcal{D}}_R^U(r) = \left\{ \left(g_j^r, \hat{s}_j^r\right) \mid j \in \mathcal{I}_u,\ c_j = 1 \right\}. \tag{40}$$

We calibrate δ once on the CVFog validation split by pooling predictions from nine teachers, corresponding to three labeled-subset samplings at each of the 5%, 10%, and 20% pairing-label budgets. We evaluate 101 candidate thresholds in [0, 0.20] with a step size of 0.002, using the full acceptance rule combining weak-view Top-1 consistency and the minimum-margin criterion.

For each threshold, precision P is the fraction of accepted predictions that are correct, and coverage C is the fraction of all pooled predictions that are accepted. Both are computed from pooled counts across the nine teachers. We select the threshold maximizing the precision–coverage harmonic mean $2PC/(P + C)$. Ties are resolved by higher precision, then higher coverage, and finally a smaller threshold. The selected value, $\delta = 0.006$, is shared across all label budgets, sampling seeds, and filtering variants. The test split is used for neither threshold calibration nor model selection. The teacher scores the complete training gallery without querying the withheld correspondence or inserting the ground-truth satellite to restrict the candidate set. It remains frozen during student training, and all pseudo-pairs are generated once and cached. We use neither an exponential-moving-average teacher nor iterative pseudo-label refinement.

F.4 Optimization Details

The student $S = (S_g, S_s)$ is initialized from T_r. Its satellite encoder remains frozen, whereas the ground-view encoder, ground-view projection layers, and WASA are optimized. For a nonempty retained set, the pseudo-pair loss is

$$\mathcal{L}_{\text{pseudo}} = \frac{1}{|\widehat{\mathcal{D}}_R^U(r)|} \sum_{\left(g_j^r, \hat{s}_j^r\right) \in \widehat{\mathcal{D}}_R^U(r)} \mathcal{L}_{\text{ret}}\left(A_s(g_j^r), \hat{s}_j^r\right), \tag{41}$$

where A_s is a stronger augmentation that preserves geographic identity; we set $\mathcal{L}_{\text{pseudo}} = 0$ if no pseudo-pair is retained. The student objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{syn}} + \mathcal{L}_{\text{real}}^L + \mathcal{L}_{\text{pseudo}}. \tag{42}$$

We use AdamW with learning rates of 1×10^{-5} for the ground-view backbone and 1×10^{-4} for the trainable projection and adaptation layers, weight decay 1×10^{-4}, linear warm-up, and cosine decay. Weak transformations comprise resizing, normalization, and mild photometric perturbations. The stronger student transformation additionally uses color jitter, blur, image compression, and random erasing; transformations that invalidate geographic identity are excluded.

At each student-training iteration, one labeled synthetic-fog batch and one labeled real-fog batch are sampled independently. RealFog-SSL-Hard additionally samples one pseudo-paired real-fog batch. Small labeled real-fog sets are sampled with replacement. The three losses are averaged independently before combination, and Real-Sup and RealFog-SSL-Hard use the same number of iterations and the same supervised-sample exposure.

F.5 Pseudo-Pair Quality Analysis

Although the withheld correspondences are unavailable during pseudo-pair generation, they permit post-hoc reliability analysis. Let k_j^* denote the withheld ground-truth satellite index of query j. We define

$$\text{Precision} = \frac{\sum_{j\in\mathcal{I}_u} c_j \mathbb{I}\left[\hat{k}_j = k_j^*\right]}{\sum_{j\in\mathcal{I}_u} c_j + \epsilon}, \tag{43}$$

and

$$\text{Coverage} = \frac{\sum_{j\in\mathcal{I}_u} c_j}{|\mathcal{I}_u|}. \tag{44}$$

Table 6: Quality of retained pseudo-pairs under different real-fog pairing-label budgets. The candidate-pool percentages are measured before filtering, and withheld identities are used only for post-hoc evaluation. Results are mean ± standard deviation over three labeled-subset samplings. R@1 is evaluated for the final RealFog-SSL-Hard student on the fixed CVFog test split.

Labeled Real	Candidate Pool	Precision (%)	Coverage (%)	RealFog R@1 (%)
5%	95%	21.37±0.25	78.38±0.16	31.28±0.53
10%	90%	21.68±0.74	76.36±0.53	34.75±0.81
20%	80%	22.82±0.06	79.13±0.71	42.87±0.64

Table 7: Ablation of confidence-filtered pseudo-pairing using 10% labeled real-fog pairs. Results are mean ± standard deviation over three labeled-subset samplings.

Variant	Top-1	Margin	Consistency	Precision (%)	Coverage (%)	R@1 (%)
Real-Sup	✗	✗	✗	–	–	31.83±0.72
Unfiltered Top-1	✓	✗	✗	19.16±0.95	100.00	31.44±0.56
+ Margin	✓	✓	✗	22.12±0.17	81.84±0.51	30.67±0.23
Full	✓	✓	✓	21.68±0.74	76.36±0.53	34.75±0.81

Here, $\epsilon > 0$ is a numerical stabilizer. Precision measures the correctness of retained pseudo-pairs, whereas coverage is the retained share of the pair-unlabeled candidate pool.

Table 6 examines pseudo-pair quality under different real-fog pairing-label budgets. Increasing the labeled budget from 5% to 20% improves RealFog R@1 from 31.28% to 42.87%, an absolute gain of 11.59%. The improvement is larger from 10% to 20% (+8.12%) than from 5% to 10% (+3.47%), showing greater gains at higher labeled budgets in this setting. Pseudo-pair precision increases modestly from 21.37% to 22.82%, whereas coverage remains relatively stable between 76.36% and 79.13% despite the changing candidate-pool size. However, precision remains below 23%, indicating substantial residual pairing noise after filtering. The small standard deviations across the three subset samplings indicate limited variation across the sampled labeled subsets. The observed retrieval gains are obtained by using these noisy pseudo-pairs as supplementary training signals alongside labeled synthetic and real-fog pairs.

F.6 Ablation of Pseudo-Pair Filtering

We evaluate the filtering components using 10% labeled real-fog pairs. Unfiltered Top-1 assigns the highest-ranked satellite to every pair-unlabeled query. The margin variant retains only predictions whose Top-1–Top-2 gap exceeds δ, while the complete variant additionally requires the Top-1 prediction to remain unchanged across two weakly augmented views.

Table 7 analyzes the individual contributions of the confidence-filtering criteria under a 10% labeled real-fog budget. Using only the Top-1 prediction retains all candidates, but yields low pseudo-pair precision (19.16%) and provides no improvement over Real-Sup (31.44% vs. 31.83% R@1). Adding the margin criterion substantially improves precision to 22.12% and reduces coverage to 81.84%, confirming that confidence margins effectively remove uncertain candidates. However, its R@1 decreases slightly to 30.67%, indicating that precision alone does not guarantee the most useful training signal for retrieval. Incorporating weak-view consistency further reduces coverage to 76.36% while maintaining comparable precision (21.68%) and raises R@1 to 34.75%, exceeding Real-Sup by 2.92%. These results suggest that consistency filtering complements confidence-based selection by suppressing unstable pseudo-pairs and retaining more cross-view-consistent supervision. The small standard deviations across three samplings further demonstrate the stability of the complete filtering strategy.

Table 8: Cross-domain generalization results between CVUSA-Fog and CVACT-Fog. Models are trained on the source domain and directly evaluated on the target domain.

Methods	Setting		CVUSA-Fog → CVACT-Fog		CVACT-Fog → CVUSA-Fog	
	Training	Testing	R@1	R@1%	R@1	R@1%
Sample4Geo	Clear	Syn. Fog	11.44	61.49	19.26	70.33
	Syn. Fog	Syn. Fog	15.89	63.28	22.34	74.54
	Syn. Fog + WASA	Syn. Fog	19.31	70.10	27.32	80.03
ConGeo	Clear	Syn. Fog	12.71	64.13	23.98	77.67
	Syn. Fog	Syn. Fog	16.87	73.20	26.26	82.43
	Syn. Fog + WASA	Syn. Fog	19.11	77.65	28.49	84.39

Table 9: Comparison of fixed Sobel and learnable local filters on CVUSA-Fog at $\beta = 2.5$ using Sample4Geo. All recall values are reported as percentages.

Variant	R@1	R@5	R@10	R@1%
WASA w/ learnable filters	91.59	97.83	98.89	99.71
WASA w/ fixed Sobel	97.93	99.44	99.59	99.87

G Additional Experiments

Cross-domain generalization. Table 8 reports cross-domain generalization performance between CVUSA-Fog and CVACT-Fog. Compared with training exclusively on clear data, synthetic fog training consistently improves both models in both transfer directions, indicating that fog-domain supervision helps models learn cross-dataset representations that are more robust to weather degradation. Incorporating WASA further improves performance. For example, the R@1 of Sample4Geo in the CVACT-Fog → CVUSA-Fog direction increases from 22.34% under the Fog setting to 27.32%, while the R@1% of ConGeo in the CVUSA-Fog → CVACT-Fog direction increases from 73.20% to 77.65%. These results demonstrate that WASA provides consistent improvements under cross-domain conditions jointly characterized by dataset discrepancies and weather degradation, while exhibiting stable adaptation capability across different backbone models. Nevertheless, the two transfer directions show noticeable performance differences, and cross-domain results remain lower than their within-domain counterparts in Table 1 of the main paper. This observation suggests that the distribution shift between CVUSA-Fog and CVACT-Fog is asymmetric and that cross-dataset generalization under foggy conditions remains challenging.

Learnable local filtering. We examine whether the fixed gradient prior contributes beyond learnable local filtering. We replace the horizontal and vertical Sobel operators in WASA with two independently initialized, learnable depthwise 3×3 convolutions. Both convolutions are bias-free and preserve the original stride and padding. Their outputs are concatenated to retain the same $2C$-channel representation, where C denotes the input channel dimension. This replacement introduces $18C$ trainable parameters. All subsequent structural processing, the degradation branch, and the calibration gate remain unchanged.

Both variants are trained using Sample4Geo on CVUSA-Fog at $\beta = 2.5$, with the same backbone initialization, data splits, and training settings. The new filters use the same optimization settings as the other convolutions in the structural branch.

Table 9 shows that fixed Sobel operators improve R@1 from 91.59% to 97.93%, an absolute gain of 6.34% over learnable filters. This result supports the value of an explicit gradient prior: fixed Sobel operators provide directional gradient responses, while subsequent learnable layers adapt them to cross-view matching. This suggests that preserving an explicit gradient prior helps the model exploit remaining structural cues for accurate matching under fog.

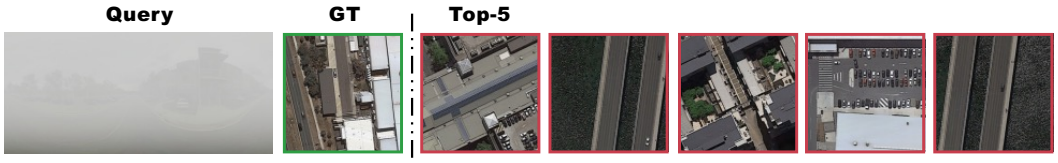


Figure 11: Failure case under severe fog degradation.

H FAILURE CASE

Fig. 11 presents a representative failure case of Sample4Geo+WASA on CVACT-Fog under severe fog degradation. The ground-view query is affected by severe fog degradation, resulting in substantially reduced overall visibility and the obscuration of distant structures and local texture information. Consequently, the discriminative cues available for cross-view matching are significantly diminished, preventing the ground-truth satellite image from being correctly retrieved. This case demonstrates that CVGL remains highly challenging when severe fog causes substantial loss of informative visual content.