Reliable Social Media Flood Image Localization for Urban Flood Situational Awareness

Zihua Zhu[1], Xianfei Yin[1,*] and Qihua Chen[1]

[1]Department of Architecture and Civil Engineering, City University of Hong Kong, Hong Kong SAR
*Corresponding Author
zihuazhu2-c@my.cityu.edu.hk, xianfyin@cityu.edu.hk, qi-hua.chen@my.cityu.edu.hk

Abstract –

Precise geolocation of social media is vital for real-time urban flood monitoring. However, researchers are frequently hindered by the scarcity of geotags due to privacy protections and the limitations of coarse localization based on posting frequency, which fails to pinpoint street-level flood sites. We present a robust end-to-end framework that addresses these critical gaps by integrating state-of-the-art Visual Place Recognition (VPR) with a visual evidence verification mechanism. Our pipeline leverages vision foundation models to maintain operational resilience against extreme visual occlusion caused by disasters. To ensure reliability, an inlier-based visual evidence layer is introduced to verify the top-ranked prediction and provide an interpretable confidence cue. This approach establishes a dependable, city-scale monitoring system that bridges the gap between raw social media data and the precise demands of municipal emergency response.

Keywords –

Urban flood; Visual geo-localization; Social media; Situational awareness

1 Introduction

Urban flooding is becoming a persistent stress test for cities as climate change intensifies heavy precipitation events and pushes drainage and transport systems closer to failure [1]. The increasing frequency of intense rainfall heightens the likelihood of cascading disruptions across critical infrastructure and services [2]. Furthermore, current risk management frameworks often struggle when hazards exceed previously experienced magnitudes, leaving emergency operations exposed to unprecedented conditions [3]. From an infrastructure and civil engineering perspective, such events are not only disaster-management issues but also operational challenges for urban resilience. Consequently, rapid, city-wide flood situational awareness is essential to the development of sensing systems and smart, sustainable, resilient cities. While fixed monitoring approaches, such as physical gauges, provide precise measurements, they are spatially sparse and costly to deploy and maintain at scale. Similarly, although satellite remote sensing can expand spatial coverage, optical revisit time constraints often limit the ability to map urban inundation in real time during rapidly evolving events [4,5].

These limitations underscore the need for multi-source data fusion, which supplements traditional monitoring with complementary data streams. Social media serves as a high-velocity stream of Volunteered Geographic Information (VGI), embodying the "citizens as sensors" paradigm, in which individuals act as observers [6]. During disasters, crowdsourced content provides near real-time ground-truth verification, filling critical gaps in sparse official sensor networks [7]. For flood management, visual data complements numerical metrics by capturing the physical reality of hazards, enabling engineers to directly evaluate inundation severity and infrastructure functionality [8, 9]. Therefore, crowdsourced imagery represents a form of opportunistic urban sensing that facilitates infrastructure-aware decision workflows.

However, a critical barrier to integrating this visual data into engineering workflows is the lack of precise geolocation [10]. Due to strict platform policies and growing concerns regarding user privacy, geospatial metadata is often removed from public uploads [11]. Without reliable coordinates, even highly informative photographs cannot be anchored to specific infrastructure assets, isolating visual evidence from location-driven operations such as emergency dispatch and pump deployment [12]. Transforming raw visual information into spatially actionable intelligence is, therefore, a prerequisite for operationalizing social media data within professional infrastructure management systems [13].

Advances in computer vision offer a practical route to operational geolocation via Visual Place Recognition (VPR), which infers coordinates by retrieving the most similar views from a georeferenced street-view database and transferring their associated location priors [14]. The field was notably advanced by NetVLAD, which

introduced trainable aggregation layers for compact global descriptors [15], and more recently by CosPlace, which reframes geo-localization as a classification task to achieve city-wide efficiency without expensive mining [16]. However, these benchmarks predominantly focus on cyclic environmental changes, such as illumination or seasons [17]. In contrast, flooding introduces abrupt visual domain shifts, ranging from submerged roads to complex reflections [18]. Furthermore, raw retrieval accuracy alone cannot support engineering adoption; emergency response requires quantifiable reliability metrics to ensure that location predictions are trustworthy enough to guide operations [19].

This research addresses these gaps by presenting an end-to-end framework for flood image geo-localization from social media imagery, with a specific focus on reliable sensing for resilient urban infrastructure operations. Rather than proposing a new backbone architecture, this study asks a practically important question: which existing VPR paradigms remain reliable under severe flood-induced visual corruption, and how can their predictions be verified before use in human decision workflows? To answer this question, we first benchmark state-of-the-art VPR models against extreme flood scenarios to assess their reliability for street-level localization. To support human-in-the-loop management, a geometric verification layer is added to produce verifiable visual evidence alongside location predictions. This work demonstrates how AI-enabled sensing can be effectively harnessed to support resilient city operations.

2 Methodology

The proposed end-to-end framework for flood situational awareness is illustrated in Figure. 1. This pipeline is designed to transform unstructured social media imagery into actionable engineering intelligence through three interconnected stages. Firstly, the process begins with the data acquisition phase (left panel), where real-time crowdsourced images are captured and referenced against a citywide street-view database. Secondly, the core flood localization module (middle panel) utilizes VPR algorithms to extract query embeddings and perform vector similarity searches, identifying the most probable geographic coordinates from the reference dataset. Lastly, the workflow culminates in the situational awareness stage (right panel), where retrieved locations undergo visual verification to map the flood extent and support municipal decision-making. The specific implementation details of each module are described in the Section 3.1

2.1 Dataset Acquisition and Processing

The city-scale geo-localization framework uses two datasets: a georeferenced database of street-view imagery and a query set of external test images whose locations must be inferred. This study utilizes the San Francisco eXtra Large (SF-XL) dataset as the reference database and evaluates retrieval performance across three separate query datasets whose statistics are summarized in Table 1. Among these query sets, SF-Flood is a newly curated dataset in this study.

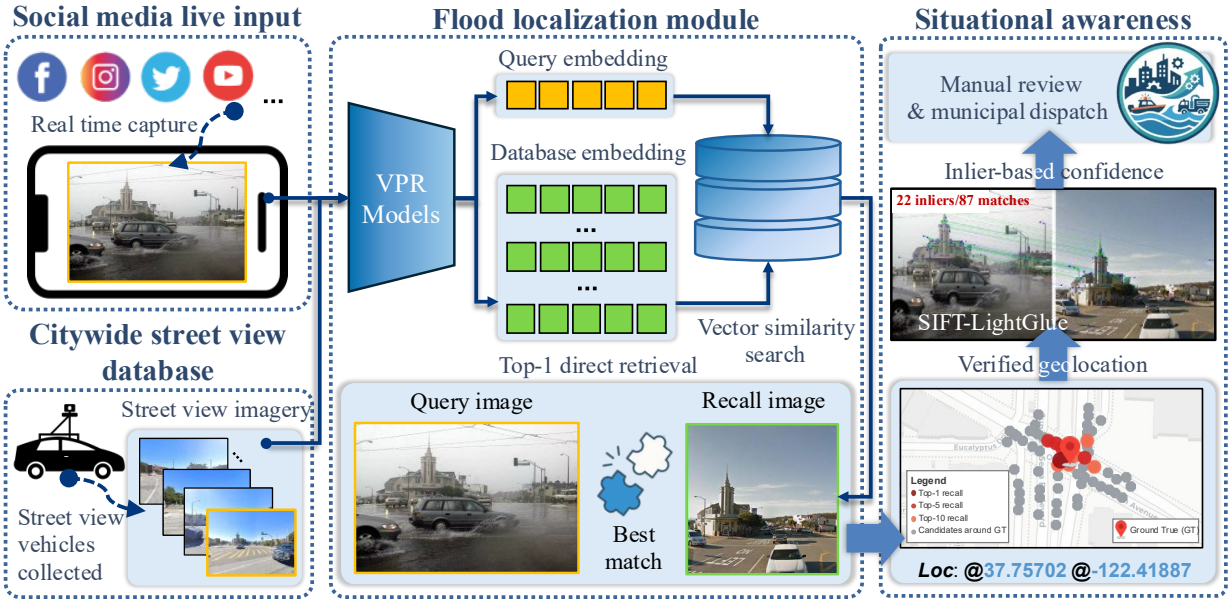

Figure 1. The schematic of the proposed end-to-end framework for urban flood situational awareness. A flood query image is matched against a geo-referenced street-view database using VPR-based descriptor retrieval. The top-1 recalled image provides an initial geolocation estimate, then verified by SIFT-LightGlue matching.

SF-XL, originally introduced by CosPlace [16], is a large-scale open-source street-view benchmark covering San Francisco. The full SF-XL release contains 41.2 million Google Street View panoramas collected across multiple years. In this study, we employ a 2.8 million-image subset as the query set for large-scale retrieval. This choice preserves city-wide spatial coverage and temporal variation, while keeping large-scale vector search computationally viable.

Table 1. Statistics for each subset used in the study.

	SF-XL (Original)	SF-XL test v1 (Test)	SF-XL test v2 (Test)	SF-Flood (Test)
Database	41.2M	2.8M	2.8M	2.8M
Queries		1000	598	400

These sets are designed to benchmark capabilities ranging from general environmental adaptability to extreme disaster resilience:

- **SF-XL Test v1 & v2 :** These serve as baselines for routine localization. v1 contains 1,000 Flickr images representing unconstrained photography [16], while v2 comprises 598 queries from the San Francisco Landmark Dataset [20] with accurate 6-DoF ground truth labels.
- **SF-Flood Dataset:** This specialized query set was curated in this study to evaluate localization under flood-induced domain shifts. Images were crawled from social media and strictly filtered to ensure they contain both visible flood evidence and identifiable geographic anchors (e.g., building façades or road layouts). To ensure the reliability of the 50 m evaluation criterion, each image was manually georeferenced through image-by-image matching in Google Street View. By explicitly incorporating extreme conditions such as water glare, reflections, and nighttime scenes, as illustrated in Figure 2, SF-Flood provides a rigorous stress test for VPR models under severe flood-related appearance changes.

We categorize the query images by their dominant localization challenge into four groups: Rich texture, Low texture, Viewpoint change, and Illumination change. Each image is assigned to the most prominent difficulty source for reporting, as details are shown in Table 2.

Table 2. The index system for crawling and extracting social media images related to flood.

Categories	Number	Definition
Rich texture	146	Repeated or visually similar urban structure; easy to confuse with look-alike locations.
Low texture	108	Low-density or peripheral scene with fewer strong localization anchors.
Viewpoint change	92	Major camera angle or framing change relative to street-view perspective.
Illumination change	54	Appearance dominated by lighting, night scenes, rain haze, or strong reflections.
Total	400	

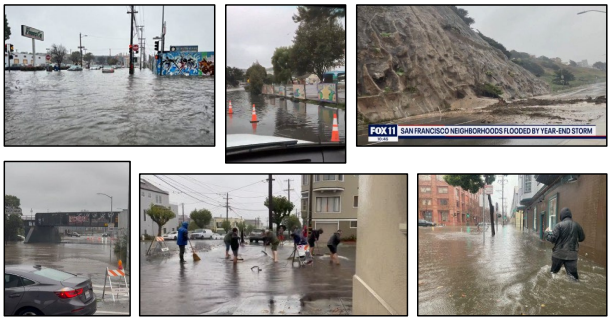

Figure 2. Examples of images of SF-Flood dataset.

2.2 VPR Mechanism and Evaluated Models

VPR) is formulated as a large-scale image retrieval task. Given a query image I_q and a geo-referenced gallery image I_i, a VPR model $f(\cdot)$ maps each image to a d-dimensional normalized global descriptor:

$$z = f(I), \qquad z \in \mathbb{R}^{\mathbb{d}}, \qquad |z|_2 = 1 \tag{1}$$

Localization is then performed by nearest-neighbor search in the descriptor space using cosine similarity. The best-matched gallery image is obtained by

$$i^* = \arg\max_i \mathrm{z_q^\top z_i} \tag{2}$$

Where z_q and z_i denote the descriptors of the query image and the i-th gallery image, respectively. The location associated with the best-matched gallery image is taken as the predicted geolocation of the query.

This study evaluates a representative set of global-descriptor VPR methods spanning classic aggregation-based models, scalable classification-based training strategies, modern feature-mixing architectures, and foundation-model backbones. All methods are evaluated using the authors' released architectures and publicly available pretrained weights.

2.3 Reliability Verification and Visualization

While global-descriptor retrieval provides candidate locations, raw similarity alone can be unreliable in flood scenarios because severe appearance changes may lead to visually plausible but spatially incorrect matches.

Therefore, we introduce a post-retrieval verification stage for the top-ranked prediction.

Specifically, local correspondences between the query image and the retrieved image are established through a hybrid matching pipeline. First, Scale-Invariant Feature Transform (SIFT) is used to detect and describe local keypoints [21]. Second, the extracted features are matched using LightGlue [22], a neural feature matcher designed to improve correspondence quality and suppress outliers under challenging visual conditions. This choice is motivated by recent VPR literature, which shows that SIFT-LightGlue is a strong-performing option for post-retrieval verification and that inlier support is an effective indicator of retrieval confidence [23].

In this framework, inlier counts provide an interpretable reliability metric. Verified correspondences are visualized as direct connections, offering a verifiable link that supports manual audits of AI outputs. This stage acts as a confidence-aware module, bridging the gap between automated retrieval and accountable human-in-the-loop decision-making.

3 Experiment and Result

3.1 Experiment setup

To evaluate localization performance, we use Recall@K (R@K) with a success threshold of 50 m. A query is counted as successfully localized at rank K if at least one of the top-K retrieved gallery images lies within 50 m of the ground-truth location. The 50 m threshold is used to represent street-level localization accuracy. In practical terms, this means that the predicted location falls within the correct street or flood vicinity, which is sufficiently precise to help municipal responders perform on-site confirmation. Under this definition, R@1 is the proportion of cases in which the system places the query within 50 m of the true flood location on the first retrieval, while R@10 indicates whether the correct location is included within a short candidate list for follow-up review.

$$R@K = \frac{1}{|Q|}\sum_{q\in Q} 1!\left(\min_{1\le j\le K} d(q, I_j) \le 50\right) \tag{3}$$

Where Q is the query set, I_j is the j-th retrieved gallery image, $d(\cdot)$ is the geodetic distance in meter.

All experiments were conducted in a Linux environment (Ubuntu 22.04, x86-64) using Python 3.12.6, PyTorch 2.5.1, CUDA 12.4, and a single NVIDIA RTX 4090 GPU (24 GB). For global retrieval, images were resized to 224 × 224. For geometric verification, images were resized to 512 × 512 to preserve local structural details for keypoint matching. The number of verified inliers is used as an indicator of prediction reliability. Unless otherwise stated, default hyperparameter settings are used in the geometric verification stage.

Although the present study focuses on retrieval-based evaluation, the ranked candidate list may also support downstream spatial aggregation or map-based clustering analysis for practical deployment. We leave such application-oriented extensions to future work.

Table 3. Recall metrics on general-purpose datasets: SF-XL test v1 and v2. Detail Rcall@K on flood image dataset: SF-Flood. Best overall results on each dataset are in **bold**, second-best results <u>underlined</u>.

Methods	Year	Backbone	Descriptor Dimension	SF-XL test v1		SF-XL test v2		SF-Flood			
				R@1	R@10	R@1	R@10	R@1	R@5	R@10	R@20
NetVLAD [15]	2016	VGG16	4096	40.1	57.7	76.9	91.1	4.2	8.2	12.2	16.2
AP-GeM [24]	2017	ResNet101	2048	37.9	54.1	66.4	88.6	5.5	11.0	12.8	16.5
SFRS [25]	2020	VGG16	4096	51.2	66.6	83.1	93.5	7.8	15.0	17.2	21.0
Cosplace [16]	2022	ResNet50	2048	76.6	85.5	88.8	96.8	16.5	26.8	32.5	38.5
EigenPlaces [26]	2023	ResNet50	2048	84.0	90.7	90.8	96.7	20.8	31.0	36.2	42.4
Conv-AP [27]	2022	ResNet50	4096	47.5	63.8	74.4	89.0	20.5	30.8	36.5	42.2
MixVPR [28]	2023	ResNet50	4096	72.5	80.9	88.6	95.0	35.2	47.2	52.2	56.0
ProGEO [29]	2024	ResNet50	1024	84.7	90.3	93.0	96.7	53.8	64.0	70.5	73.8
SALAD [30]	2024	DINOv2	8448	88.7	94.4	94.6	98.2	56.0	71.8	75.5	79.5
CliqueMining[31]	2024	DINOv2	8448	85.5	92.6	94.5	98.3	55.5	68.2	74.0	76.8
DINO-Mix [32]	2024	DINOv2	4096	81.7	90.7	<u>95.2</u>	<u>98.7</u>	53.2	63.7	71.0	76.2
BoQ [33]	2024	DINOv2	12288	<u>91.9</u>	<u>96.6</u>	**95.2**	**98.7**	<u>70.5</u>	<u>79.5</u>	<u>81.8</u>	<u>84.5</u>
Megaloc [34]	2025	DINOv2	8448	**95.3**	**98.0**	94.8	98.5	**76.8**	**82.8**	**86.5**	**89.0**

3.2 Benchmarking on Flood Scenarios

Table 3 compares VPR performance on two general-purpose query sets (SF-XL test v1/v2) and the flood-specific query set (SF-Flood). On the standard sets, most recent methods achieve high recall, indicating that general urban localization is already relatively mature. For example, Megaloc reaches 95.3% R@1 on SF-XL test v1, while BoQ and DINO-Mix both reach 95.2% R@1 on SF-XL test v2. Even several CNN-based methods remain competitive under these routine conditions.

The picture changes substantially on SF-Flood. Under flood-induced appearance change, conventional global-descriptor methods degrade sharply. NetVLAD drops to 4.2% R@1, while CosPlace and EigenPlaces reach only 16.5% and 20.8%, respectively. MixVPR improves to 35.2%, but still remains far below the best-performing models. These results show that methods that perform well under normal urban conditions may fail when the scene is heavily altered by water coverage, reflections, occlusions, and adverse illumination.

In contrast, methods built on DINOv2 features remain much more robust. On SF-Flood, Megaloc achieves the best performance with 76.8% R@1, 82.8% R@5, 86.5% R@10, and 89.0% R@20, followed by BoQ with 70.5% R@1 and 84.5% R@20. This margin over the strongest CNN-based competitor (53.8% R@1 for ProGEO) suggests that foundation-model descriptors are better able to preserve stable structural cues under severe flood-related appearance change.

From an operational perspective, the result of 76.8% R@1 means that the system can place the query within 50 m of the true flood location in about three out of four cases at the first prediction. In addition, the strong R@10 and R@20 results indicate that the correct location is often contained within a short candidate list, which is useful for human review when the first prediction is uncertain

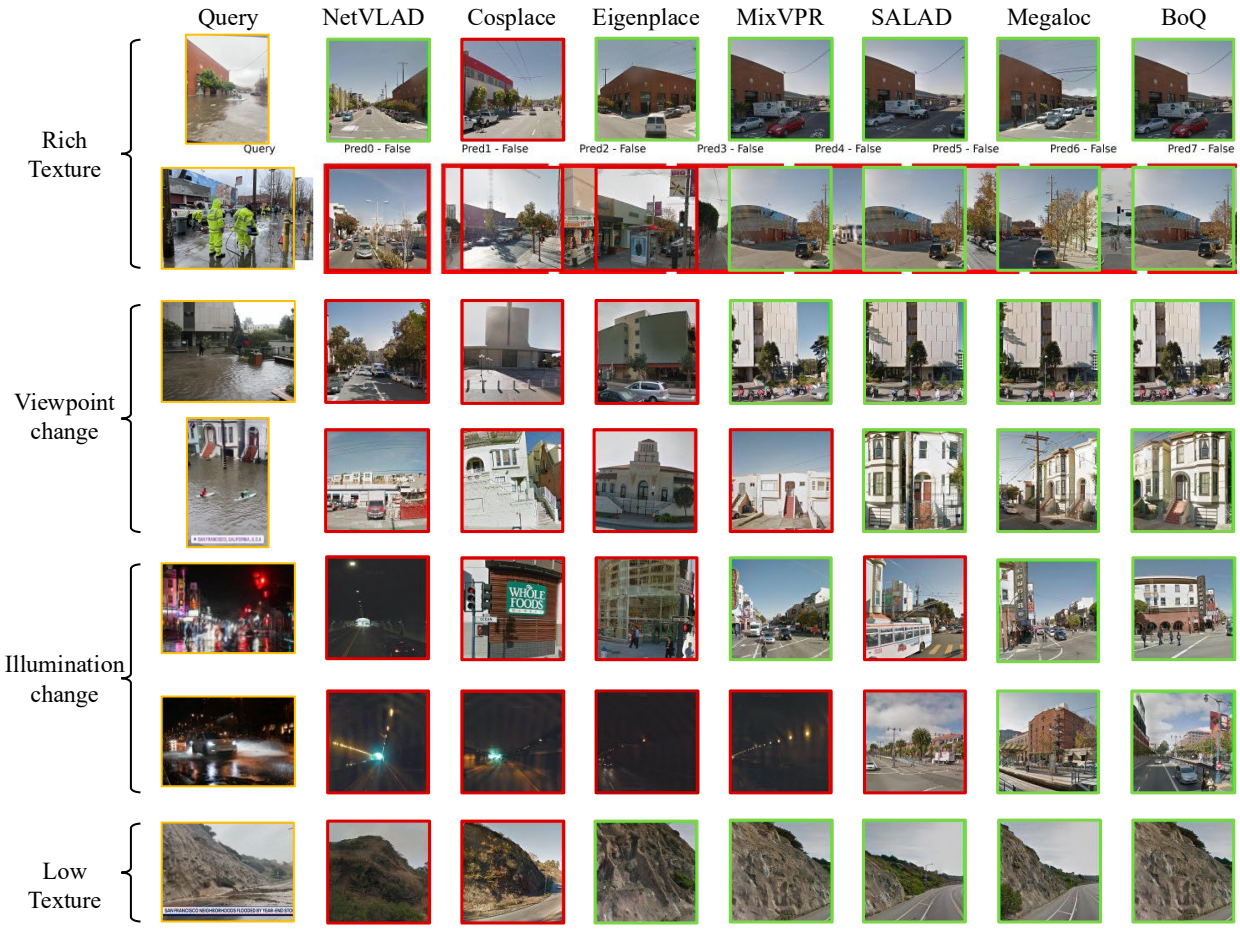

Figure 3. Qualitative retrieval comparison on the SF-Flood dataset under challenging conditions. Green borders indicate correct localizations within 50 meters of the ground truth, while red borders denote failures, illustrating representative mislocalized cases and highlighting failure modes under severe visual ambiguity.

Figure 3 provides qualitative examples under several difficult conditions, including viewpoint changes, nighttime scenes, rich-texture settings, and low-texture settings. In many of these cases, conventional methods retrieve visually similar but incorrect locations, whereas the stronger foundation-model-based methods more often recognize the right place or keep it within the top-ranked list. At the same time, the red-box examples in Figure 3 show that failure cases still remain, especially when visual ambiguity is high. Overall, the results indicate that flood localization is harder than standard urban geo-localization and that foundation-model-based descriptors provide the most reliable retrieval backbone among the methods tested.

3.3 Visual Evidence Validation

Recall metrics summarize retrieval performance at the dataset level, but practical use also requires evidence for whether a prediction is trustworthy. We therefore apply a post-retrieval geometric verification step using SIFT-LightGlue to assess the top-ranked match.

Representative verification examples are shown in Figure 4. Even under floodwater, reflections, nighttime conditions, and partial occlusion, the verified correspondences are concentrated on stable structural elements, such as façades, poles, and traffic infrastructure, rather than transient flood textures. This provides visual evidence for why some heavily degraded queries can still be localized correctly.

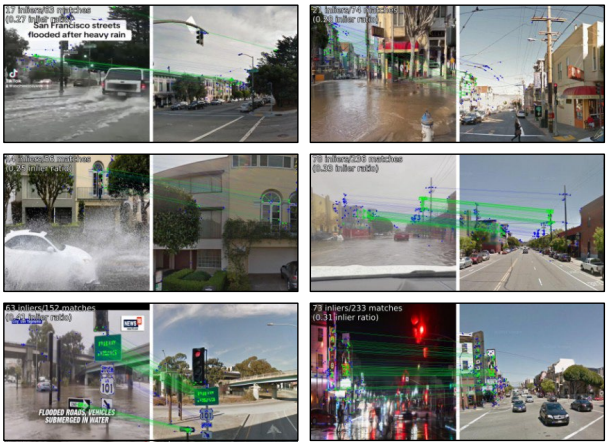

Figure 4. Qualitative examples of visual evidence validation.

Figure 5 further quantifies the role of visual evidence extraction. Figure 5(a) shows the localization accuracy within the 50 m threshold across different inlier ranges. All groups maintain relatively high accuracy, indicating that visual evidence is informative for match reliability. Also, the relationship is not monotonic, especially in a sparsely scattered range, which suggests that inlier count should not be interpreted as a universal hard threshold.

Figure 5(b) shows that successful predictions are not only below the 50 meter threshold, but are often concentrated at very small error distances (average 2.4 m), indicating that correct matches are typically spatially close to the ground truth. By contrast, the failed cases form a smaller group (23 cases) above the 50 m threshold and are more common when inlier support is limited. Still, because some correct cases also appear at relatively low inlier counts, the plot does not support a single clean cutoff. Instead, it supports using inlier count as a practical confidence indicator to prioritize stronger matches and flag uncertain ones for review.

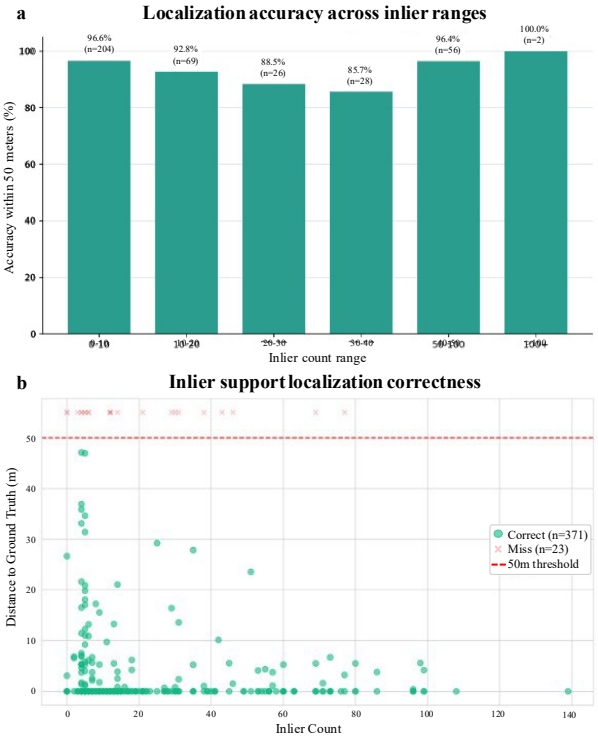

Figure 5. Quantitative analysis of inlier-support visual evidence validation.

This result is consistent with the qualitative examples in Figure 4. Visual evidence verification does not eliminate all failures, but it provides a practical way to distinguish stronger from weaker predictions. In deployment, this makes the retrieved location more interpretable and supports human-in-the-loop review when flood imagery is ambiguous.

4 Discussion and Conclusion

This study presents an end-to-end framework for geo-localizing flood-related social media images, with a focus on reliable sensing for resilient urban infrastructure

operations. Rather than proposing a new VPR backbone, the study addresses a practical question: which existing VPR paradigms remain reliable under severe flood-induced visual disruption, and how can their predictions be verified before entering human decision workflows? The results show that flood scenarios introduce a much harder setting than standard urban benchmarks, and that foundation-model-based descriptors currently provide the most reliable retrieval backbone among the methods tested. More importantly, the added verification stage makes the retrieved result interpretable by attaching explicit visual evidence to the predicted location.

The proposed framework shows how crowdsourced visual data can be transformed into infrastructure-oriented situational awareness, supporting tasks such as rapid identification of affected urban areas and awareness of access routes. In this sense, the study extends social media imagery from online content into an opportunistic sensing layer for urban infrastructure management.

There are several limitations in this study. First, the current evaluation is limited to a single-city case study in San Francisco, and broader validation across cities with different streetscapes, infrastructure conditions, and development contexts remains necessary. Second, the SF-Flood dataset contains 400 query images and, while deliberately designed to include difficult cases such as nighttime scenes, viewpoint changes, and visually degraded flood conditions, it still cannot fully capture the diversity of urban flood imagery. Finally, although the current system is suitable for time-sensitive analysis, lighter verification pipelines and more comprehensive runtime evaluation are still needed for future deployments.

Future work should therefore move in three directions. The first is broader validation, including multi-city testing and more systematic analysis of visually ambiguous urban forms. The second is deployment-oriented refinement, such as lighter matching models and latency tests. The third is multi-modal extension, where visual localization can be combined with textual, temporal, or other contextual signals from social media to improve robustness when image evidence alone is insufficient.

Acknowledgement

The authors gratefully acknowledge the financial support provided by the National Natural Science Foundation of China (Grant No. 72404233) and the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2025A1515010190).

References

[1] A. F. Prein, R. M. Rasmussen, K. Ikeda, C. Liu, M. P. Clark, and G. J. Holland. The future intensification of hourly precipitation extremes. *Nature Climate Change*, 7(1):48–52, 2017.

[2] IPCC. AR6 WGII, Chapter 6: Cities, settlements and key infrastructure. IPCC. On-line: https://www.ipcc.ch/report/ar6/wg2/downloads/report/IPCC_AR6_WGII_Chapter06.pdf, Accessed: 15/12/2025.

[3] H. Kreibich, A. F. Van Loon, K. Schröter, P. J. Ward, M. Mazzoleni, N. Sairam, et al. The challenge of unprecedented floods and droughts in risk management. *Nature*, 608(7921):80–86, 2022.

[4] C. Mydlarz, P. Sai Venkat Challagonda, B. Steers, J. Rucker, T. Brain, B. Branco, et al. FloodNet: Low-cost ultrasonic sensors for real-time measurement of hyperlocal, street-level floods in New York City. *Water Resources Research*, 60(5):e2023WR036806, 2024.

[5] E. Portalés-Julià, G. Mateo-García, C. Purcell, and L. Gómez-Chova. Global flood extent segmentation in optical satellite images. *Scientific Reports*, 13(1):20316, 2023.

[6] M. F. Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221, 2007.

[7] Y. Chen, M. Hu, X. Chen, F. Wang, B. Liu, and Z. Huo. An approach of using social media data to detect the real time spatio-temporal variations of urban waterlogging. *Journal of Hydrology*, 625:130128, 2023.

[8] J. Li, R. Cai, Y. Tan, H. Zhou, A. M. Sadick, W. Shou, and X. Wang. Automatic detection of actual water depth of urban floods from social media images. *Measurement*, 216:112891, 2023.

[9] J. K. Kavota, J. R. K. Kamdjoug, and S. F. Wamba. Social media and disaster management: Case of the north and south Kivu regions in the Democratic Republic of the Congo. *International Journal of Information Management*, 52:102068, 2020.

[10] X. X. Zhu, Y. Wang, M. Kochupillai, M. Werner, M. Häberle, E. J. Hoffmann, et al. Geoinformation harvesting from social media data: A community remote sensing approach. *IEEE Geoscience and Remote Sensing Magazine*, 10(4):150–180, 2023.

[11] S. Ying, Y. Huang, L. Qian, and J. Song. Privacy paradox for location tracking in mobile social networking apps: The perspectives of behavioral reasoning and regulatory focus. *Technological Forecasting and Social Change*, 190:122412, 2023.

[12] Y. Feng, X. Huang, and M. Sester. Extraction and analysis of natural disaster-related VGI from social media: review, opportunities and challenges. *International Journal of Geographical Information*

Science, 36(7):1275–1316, 2022.

[13] S. Xiao, H. Gu, D. Shen, Z. Niu, J. Xiao, and F. Yu. A systematic review of social media-enabled flood disaster informatics: Method, technology and application. *International Journal of Applied Earth Observation and Geoinformation*, 144:104935, 2025.

[14] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford. Visual place recognition: A survey. *IEEE Transactions on Robotics*, 32(1):1–19, 2015.

[15] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5297–5307, 2016.

[16] G. Berton, C. Masone, and B. Caputo. Rethinking visual geo-localization for large-scale applications. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4878–4888, 2022.

[17] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla. 24/7 place recognition by view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1808–1817, 2015.

[18] M. Rahnemoonfar, T. Chowdhury, A. Sarkar, D. Varshney, M. Yari, and R. R. Murphy. Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access*, 9:89644–89654, 2021.

[19] M. Zaffar, L. Nan, and J. F. Kooij. On the estimation of image-matching uncertainty in visual place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 17743–17753, 2024.

[20] D. M. Chen, G. Baatz, K. Koser, S. S. Tsai, R. Vedantham, ¨ T. Pylvan¨ ainen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, ¨ B. Girod, and R. Grzeszczuk. City-scale landmark identification on mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 737–744, 2011.

[21] D. G. Lowe. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2):91–110, 2004.

[22] P. Lindenberger, P. E. Sarlin, and M. Pollefeys. Lightglue: Local feature matching at light speed. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 17627–17638, 2023.

[23] D. Sferrazza, G. Berton, G. Trivigno, and C. Masone. To match or not to match: Revisiting image matching for reliable visual place recognition. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 2849–2860, 2025.

[24] A. Gordo, J. Almazan, J. Revaud, and D. Larlus. End-to-end learning of deep visual representations for image retrieval. International Journal of Computer Vision, 124(2):237–254, 2017.

[25] Y. Ge, H. Wang, F. Zhu, R. Zhao, and H. Li. Self-supervising fine-grained region similarities for large-scale image localization. In Proceedings of the European Conference on Computer Vision, pages 369–386, 2020.

[26] G. Berton, G. Trivigno, B. Caputo, and C. Masone. Eigenplaces: Training viewpoint robust models for visual place recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 11080–11090, 2023.

[27] A. Ali-bey, B. Chaib-Draa, and P. Giguere. Gsv-cities: Toward appropriate supervised visual place recognition. Neurocomputing, 513:194–203, 2022.

[28] A. Ali-Bey, B. Chaib-Draa, and P. Giguere. Mixvpr: Feature mixing for visual place recognition. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 2998–3007, 2023.

[29] J. Hu, C. Mao, C. Tan, H. Li, H. Liu, and M. Zheng. Progeo: Generating prompts through image-text contrastive learning for visual geo-localization. In Proceedings of the International Conference on Artificial Neural Networks, pages 448–462, 2024.

[30] S. Izquierdo and J. Civera. Optimal transport aggregation for visual place recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 17658–17668, 2024.

[31] S. Izquierdo and J. Civera. Close, but not there: Boosting geographic distance sensitivity in visual place recognition. In Proceedings of the European Conference on Computer Vision, pages 240–257, 2024.

[32] G. Huang, Y. Zhou, X. Hu, C. Zhang, L. Zhao, and W. Gan. DINO-Mix enhancing visual place recognition with foundational vision model and feature mixing. Scientific Reports, 14(1):22100, 2024.

[33] A. Ali-Bey, B. Chaib-draa, and P. Giguere. BoQ: A place is worth a bag of learnable queries. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 17794–17803, 2024.

[34] G. Berton and C. Masone. Megaloc: One retrieval to place them all. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 2861–2867, 2025.