ISPRS Journal of Photogrammetry and Remote Sensing 230 (2025) 616–629

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Multi-source geo-localization in urban built environments for crowd-sourced images by contrastive learning

Qianbao Hou [a,b], Ce Hou [d,e], Fan Zhang [e], Qihao Weng [a,b,c,*]

[a] JC STEM Lab of Earth Observations, Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China
[b] Research Centre for Artificial Intelligence in Geomatics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China
[c] Research Institute for Land and Space, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China
[d] Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong, China
[e] Institute of Remote Sensing and Geographical Information System, School of Earth and Space Sciences, Peking University, Beijing, China

ARTICLE INFO

Keywords:
Image geo-localization
Crowd-sourced images
Multi-source data fusion
High-resolution satellite images
Street-view images
Urban spatial analytics

ABSTRACT

Crowd-sourced images (CSIs) offer an unprecedented opportunity for gaining deeper insights into urban built environments. However, the lack of precise geographic information limits their effectiveness in various urban applications. Traditional geo-localization methods, which rely on matching CSIs with geo-tagged street-view images (SVIs), face significant challenges due to sparse coverage and temporal misalignment of reference data, especially in developing countries. To overcome these limitations, this paper proposes a novel contrastive learning framework that integrates SVIs and satellite images (SIs), utilizing a multi-scale channel attention module and InfoNCE loss to enhance the geo-localization accuracy of CSIs. Additionally, we leverage SIs to generate synthetic SVIs in areas where actual SVIs are unavailable or outdated, ensuring comprehensive coverage across diverse urban environments. A simple yet efficient data preprocessing method is proposed to align multi-view images for enhanced feature fusion. As part of our research efforts, we introduce a Multi-Source Geo-localization Dataset (MSGD) consisting of 500k geo-tagged pairs collected from 12 cities across six continents, encompassing diverse urban typologies from dense skyscraper districts to low-density areas, providing a valuable resource for future research and advancements in geo-localization methods. Our experiments show that the proposed method outperforms state-of-the-art approaches on the challenging MSGD dataset, highlighting the importance of incorporating SIs as a complementary data source for accurate geo-localization. Our code and dataset will be released at https://github.com/RCAIG/CrowdsourcingGeoLocalization.

1. Introduction

Crowd-sourced images (CSIs) refer to images uploaded by citizens and shared on the Internet (Heipke, 2010; Li et al., 2016; Huang et al., 2021; Huang et al., 2023). Compared to street-view images (SVIs) from commercial platforms like Google Street View, CSIs can access remote or disaster-affected areas that street-view vehicles may not reach, effectively filling data gaps in those regions (Hou et al., 2024). Moreover, CSIs provide more detailed ground-level perspectives than satellite images (SIs), facilitating real-time urban monitoring and emergency response applications (Huang et al., 2024). As a novel data source emerging in the big data era, CSIs have been recognized as a valuable resource for understanding urban built environments. However, the potential of CSIs remains largely untapped due to the pervasive absence of accurate geographic metadata, primarily caused by privacy protections and transmission losses, making CSI geo-localization a critical yet challenging task.

Prevailing geo-localization methods for CSIs predominantly rely on ground-to-ground retrieval approaches (Cheng et al., 2018; Warburg et al., 2020; Yan et al., 2021; Xu et al., 2023), where query images are matched against reference databases of geo-tagged SVIs. While effective in well-documented urban areas, these approaches fail entirely in regions lacking comprehensive SVIs coverage—a common situation in developing countries, historical districts with limited vehicle access, and recently developed neighborhoods. Even in areas with SVIs coverage, the temporal mismatch between reference databases and current CSIs can substantially reduce matching accuracy as urban environments undergo constant change through construction, renovation, and

* Corresponding author..
E-mail address: qihao.weng@polyu.edu.hk (Q. Weng).

https://doi.org/10.1016/j.isprsjprs.2025.09.024
Received 23 May 2025; Received in revised form 27 September 2025; Accepted 29 September 2025
Available online 9 October 2025

seasonal variations. To overcome these coverage limitations, our framework integrates information from both ground and aerial perspectives, leveraging the near-universal coverage of SIs to provide comprehensive urban environment information. In scenarios where SVIs are unavailable (either non-existent or captured more than one year apart from the query CSI), we leverage SIs to generate synthetic SVIs through a PanoGAN architecture (Wu et al., 2022), enabling comprehensive geolocation capability even in regions previously inaccessible to street-view vehicles. Subsequently, we integrate this synthetic SVI with the original SI to exploit multi-source data complementarity for enhanced geolocation performance.

Although cross-view geo-localization has emerged as a popular alternative to traditional ground-to-ground approaches by leveraging the global coverage of SIs (Toker et al., 2021; Shi et al., 2022; Deuser et al., 2023; Fervers et al., 2024), existing models encounter significant challenges when applied to CSIs due to fundamental differences in data characteristics. Current cross-view models are predominantly optimized for standardized commercial street-view data featuring consistent horizontal perspectives, uniform camera heights, and professional-grade image quality. In contrast, CSIs present substantial variability: diverse capture perspectives, inconsistent image quality, unknown camera orientations, and limited field of view (FoV) that restricts contextual information availability (Fervers et al., 2024). To address these challenges, we develop a contrastive learning-based framework specifically designed for CSI geo-localization that incorporates polar transformation and spatial feature calibration for cross-view alignment, a distortion-resistant feature extraction method based on a modified ConvNeXt-Base architecture (Liu et al., 2022), and a correlation-based orientation estimation mechanism.

The advancement of CSI geo-localization research is further hampered by the absence of appropriate benchmark datasets. Existing datasets typically incorporate only one type of imagery (usually SVIs), preventing the development and evaluation of methods that leverage complementary information from different image sources. Additionally, most geo-localization datasets focus on single countries or regions, such as CVUSA (Workman et al., 2015), Vo (Vo and Hays, 2016), CVACT (Liu and Li, 2019), and VIGOR (Zhu et al., 2020), severely limiting their applicability to diverse global urban contexts. This geographic narrowness creates significant challenges for developing globally generalizable models. Furthermore, most existing datasets focus exclusively on spatial correspondence between query images and reference databases, while neglecting the temporal alignment. Images captured at the same location but at different times can exhibit substantial visual differences due to environmental changes and construction activities. To support the evaluation and advancement of multi-source geo-localization research, we introduce a Multi-Source Geo-localization Dataset (MSGD), the first globally diverse dataset specifically designed for this task. MSGD contains over 500k geo-tagged image pairs (CSIs, SVIs, and SIs) from 12 cities across 6 continents, encompassing diverse urban typologies from dense skyscraper districts to low-density flat areas.

The primary contributions of this work can be summarized as follows:

(1) We propose a contrastive learning-based framework for multi-source CSI geo-localization that integrates ground-level (SVIs) and aerial (SIs) perspectives as complementary reference sources. In regions with incomplete or outdated commercial SVI coverage, our framework generates synthetic SVIs from SIs to to leverage multi-view visual information, thereby expanding global CSI geo-localization capabilities.

(2) We introduce MSGD, the first globally diverse multi-source geo-localization dataset comprising 500k geo-tagged image pairs (CSIs, SVIs, SIs) from 12 cities across 6 continents. Featuring diverse urban typologies and addressing limitations of existing datasets such as restricted geographic coverage, single image types, and lack of temporal alignment, MSGD establishes a new benchmark for more robust and generalizable multi-source geo-localization research.

(3) Extensive experiments on MSGD demonstrate the significant contribution of SIs to the CSI geo-localization task. This finding underscores the necessity of integrating SIs as a complementary source within multi-source geo-localization frameworks.

2. Related work

Positioning is a crucial issue in urban environments. The city's dense skyscraper and complex environment hinder accurate and reliable positioning methods. Image-based positioning methods provide us with a way to effectively utilize such density and complexity of cities for positioning. In the following, we give an overview of image geo-localization, expanding from three aspects: ground-view geo-localization, cross-view geo-localization, and existing geo-localization datasets.

2.1. Ground-view geo-localization

Ground-view image geo-localization, which employs geo-tagged images as a reference database, can be categorized into two principal methods. The first method conceptualizes the task as an image retrieval problem, wherein the geographical location of a query image is determined based on the spatial coordinates of the visually most similar reference image(s) retrieved from the database through feature extraction and matching algorithms (Cheng et al., 2018; Yan et al., 2021; Xu et al., 2023). The second method formulates the problem as a classification task, whereby the geographic space of interest is systematically partitioned into cells, with the image database organized according to these spatial units, thus enabling models to predict the appropriate geographic cell when presented with a query image (Muller-Budack et al., 2018; Berton et al., 2022; Clark et al., 2023). Notably, the retrieval-based approach demonstrates particular efficacy in urban environments characterized by densely distributed SVIs, while the classification-based methodology proves advantageous for global-scale image retrieval and place recognition tasks.

In this task, the challenges lie in overcoming the constantly changing environment of different seasons, different lighting, moving objects, and the changing camera perspective pose. In addition, the reference database of images typically lacks comprehensive coverage of large areas, and its update process is often slow. Ground-view image geo-localization will fail in regions without reference ground images in the dataset.

2.2. Cross-view geo-localization

Cross-view geo-localization approaches leverage ground-level imagery as query input and employ an exhaustive top-down imagery database, wherein each patch serves as a potential reference, to determine precise geographical coordinates of the query location. Early cross-view matching methods utilize hand-crafted features to address this issue (Viswanathan et al., 2014; Wu et al., 2018; Toker et al., 2021). However, due to the substantial perspective gaps between ground-level and overhead images, hand-crafted feature extraction methods result in significant feature disparities and lower retrieval accuracy. Convolutional Neural Networks (CNNs), known for their powerful feature extraction capabilities, have been widely adopted in the field of cross-view geo-localization (Lin et al., 2015; Hu et al., 2018; Shi et al., 2020). There are also studies that have applied Vision Transformer (ViT) (Dosovitskiy et al., 2020) to this task and achieved good results (Yang et al., 2021a; Dai et al., 2022).

To bridge the domain gap between SVIs and SIs, researchers have proposed several transformative approaches: applying polar transformations to SIs (Shi et al., 2020; Shi et al., 2022), converting street-view panoramas into bird's-eye view representations (Fervers et al., 2022; Ye et al., 2024), employing GANs to synthesize either SVIs or SIs (Regmi and Shah, 2019; Toker et al., 2021), or combining SIs with their semantic segmentation masks (Pro et al., 2024), thereby establishing more coherent cross-domain feature correspondences. While these methods enhance retrieval accuracy, existing models are primarily

optimized for standardized data sources like Google Street View, yet struggle with CSIs that present diverse perspectives, unknown orientations, and limited contextual information (Fervers et al., 2024).

2.3. Geo-localization datasets

Geo-localization datasets provide essential benchmarks for developing and evaluating location recognition algorithms. These datasets generally fall into two categories: ground-view geo-localization and cross-view geo-localization, each presenting distinct challenges and limitations.

Ground-view geo-localization datasets typically utilize SVIs as reference data to localize other ground-level query images, with both query and reference data commonly sourced from the same platform. The GSV dataset (Zamir and Shah, 2014), derived from Google Street View, contains merely 60,000 images from three US cities, which is not enough to justify the spatial generalization capability of the proposed method. Similarly, Pittsburgh250k (Torii et al., 2015), extracted from Google Street View panoramas in Pittsburgh, utilizes only 10,586 panoramas with limited angular sampling (two yaw directions and twelve pitch directions). This restricted geographical coverage results in significantly limited applicability and transferability of these methods in other urban environments. While MSLS (Warburg et al., 2020) offers greater diversity with images from 30 cities across six continents, it remains confined to street-level imagery within a single perspective paradigm, thereby limiting its application for comprehensive spatial understanding across diverse environments.

Cross-view geo-localization datasets attempt to leverage overhead imagery as reference data to localize SVIs, introducing the fundamental challenge of bridging drastically different viewpoints. CVUSA (Workman et al., 2015) contains 35,532 ground-to-satellite image pairs across the United States, while CVACT (Liu and Li, 2019) provides 35,532 training pairs and 92,802 testing pairs from Australia. However, both datasets employ simplistic one-to-one retrieval paradigms with strictly aligned image pairs, which demonstrates limited utility in real-world applications. The work by Vo and Hays (2016) expanded geographical diversity by collecting paired images from 11 U.S. cities, contributing 1.6 million precisely aligned street-satellite image pairs while maintaining the same one-to-one matching constraints. Most recently, the VIGOR dataset (Zhu et al., 2020) has advanced dataset design by introducing non-centrally aligned imagery across four U.S. cities, allowing multiple street views to correspond with a single satellite area, thus more accurately reflecting the complex spatial relationships encountered in real-world localization tasks.

Despite recent progress, current datasets remain limited by insufficient geographical coverage, platform biases, extreme viewpoint disparities, and neglect of temporal factors—locations change visibly over time due to environmental shifts and construction. Our dataset overcomes these limitations through diverse query sources, complementary multi-perspective reference data, temporal consistency, and global representation of varied urban typologies.

3. Multi-source geo-localization dataset (MSGD)

3.1. Data collection and processing

This study addresses the CSI geo-localization by using two complementary reference sources: SVIs and SIs. Unlike existing datasets that assume perfect one-to-one or one-to-many correspondences between query and reference images, our dataset accounts for realistic scenarios where reference data may be incomplete or temporally misaligned with query images.

The query CSIs are collected from Mapillary, a global platform for crowd-sourced geo-tagged imagery (Zielstra and Hochmair, 2012; Ma et al., 2020). Reference SVIs are obtained through the Google Street View API, while SIs are acquired from the Google Maps Static API at zoom level 20 (0.149 m per pixel). SIs are pre-processed by Google Maps to remove clouds and other artifacts, so no additional cloud removal was performed. To ensure comprehensive spatial sampling, we utilized road network data from OpenStreetMap (OSM) across 12 cities, as shown in Fig. 1.

The data acquisition pipeline is structured as follows: Initially, we collected CSIs across the study area from Mapillary, using road network data provided by OSM. Previous studies have shown that SVIs typically represent the surrounding urban landscape within a 50-meter radius (Kang et al., 2021; Yang et al., 2021b; Fan et al., 2024). Therefore, for each retained CSI location, we acquired the corresponding SI centered on the CSI coordinates and clipped to a 50 m $\times$ 50 m area. We then retrieved SVIs within a 50-meter radius from four cardinal directions — 0, 90, 180, and 270 degrees, classifying reference conditions as either "available" (SVIs captured within one year and 50 m of the query CSI) or "unavailable" (no SVIs within 50 m, or SVIs captured more than one year apart from the query CSI). In scenarios where SVIs are unavailable, we train city-specific PanoGAN models (Wu et al., 2022) using available SVI-SI pairs from each city. The training process is conducted independently for each city to better capture local architectural and urban characteristics. These trained models then generate synthetic street-view perspectives from SIs in regions where only SIs are available. The proposed cross-view synthesis module transforms overhead imagery into ground-level viewpoints, providing essential reference data in underrepresented regions and maintaining operational effectiveness even in areas with limited or outdated street-view coverage. Based on the above pipeline, our approach is built on two main assumptions: the scene does not undergo substantial changes within a few months and leveraging SI to generate synthetic SVI does not introduce notable artifacts or mismatches (as validated in Section 5.4.1 and Appendix A). Examples from the proposed dataset are illustrated in Fig. 2.

3.2. Dataset comparison

Based on the comparative analysis with existing datasets (see Table 1), the advantages of the proposed MSGD can be consolidated into three principal aspects:

(1). MSGD provides global coverage including developing countries, with temporally consistent query-reference image pairs. This extensive coverage across cities with diverse architectural and environmental characteristics enables robust evaluation of geo-localization algorithms under varied real-world conditions, addressing a significant limitation in prior benchmarks.
(2). MSGD innovatively combines both SVIs and SIs as reference data, and employs a PanoGAN model (Wu et al., 2022) to generate synthetic SVIs in areas where SVIs are unavailable or outdated. This approach effectively addresses the inherent limitations of existing multi-source fusion methods that fail in areas without SVIs, effectively eliminating geographical blind spots in algorithm evaluation.
(3). MSGD accommodates real-world visual challenges including unfixed orientation, limited FoV, and viewpoint diversity. These variable imaging conditions create a more authentic testing environment that better reflects the complexities of practical geo-localization applications.

3.3. Evaluation protocol

Our evaluation framework is designed to assess both the generalization within familiar geographic contexts and the transferability to entirely new urban environments. We establish two distinct evaluation protocols:

(1). Same-area: For each city in our dataset, we partition the data into training, validation, and test sets in a 3:1:1 ratio. This protocol

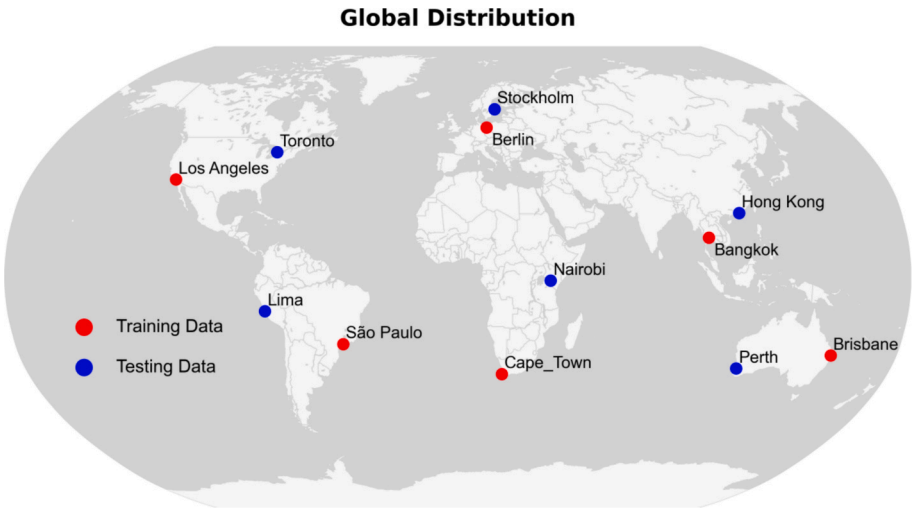

Fig. 1. Study area overview. MSGD encompasses imagery from 12 cities globally; red dots indicate training cities, while blue ones represent testing cities. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

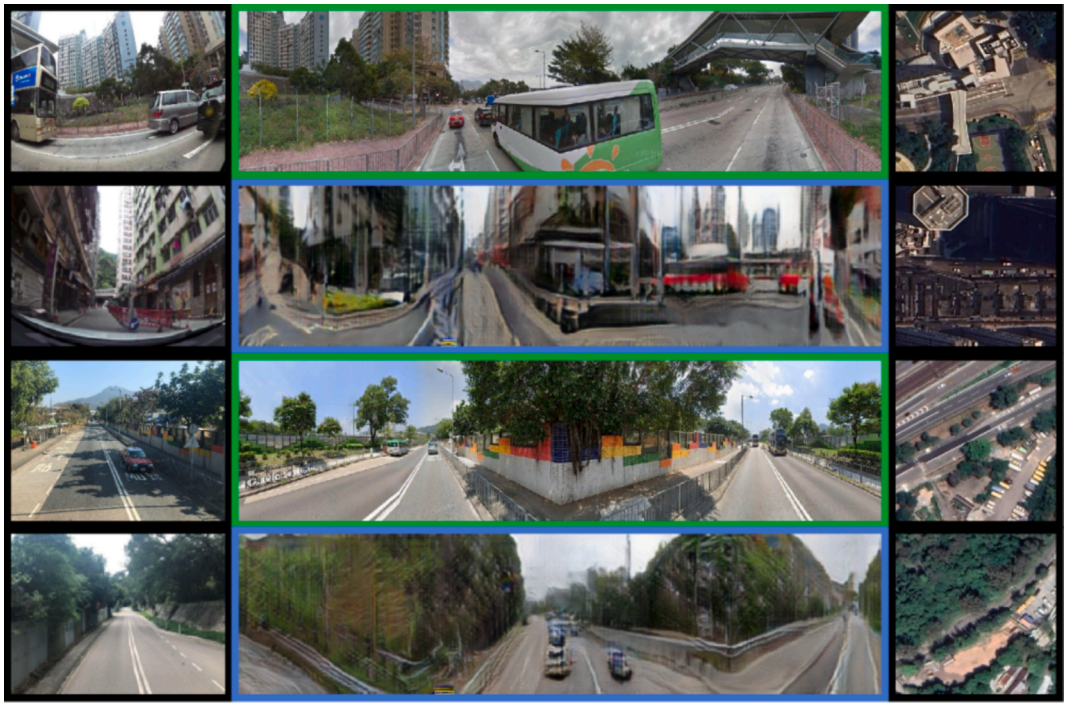

Fig. 2. Examples of the dataset. The query crowd-sourced images (CSIs) are on the left, with their corresponding street-view images (SVIs) in the center and satellite images (SIs) on the right. Green and blue boxes indicate real and synthetic SVIs, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Comparison between the proposed MSGD with existing datasets.

	GSV	Pittsburgh250k	MSLS	CVUSA	Vo	CVACT	VIGOR	MSGD
Query	CSI	SVI	CSI	SVI	SVI	SVI	SVI	CSI
Reference	SVI	SVI	CSI	SI	SI	SI	SI	SVI + SI
Global scale	×	×	✓	×	×	×	×	✓
Include developing country	×	×	✓	×	×	×	×	✓
Unfixed orientation	✓	✓	✓	×	✓	×	×	✓
Limited FOV	✓	✓	✓	×	✓	×	×	✓
Temporal consistency	×	×	×	×	×	×	×	✓
Viewpoint diversity	✓	×	✓	×	×	×	×	✓
Overcome SVI limits	×	×	✓	×	×	×	×	✓

assesses how effectively models can leverage learned urban patterns and architectural styles to localize novel views within familiar city environments.

(2). Cross-area: As shown in Fig. 1, the 6 red-marked cities (Los Angeles, São Paulo, Berlin, Cape Town, Bangkok, and Brisbane) are used for training and validation, while the 6 blue-marked cities (Toronto, Lima, Stockholm, Nairobi, Hong Kong, and Perth) are used for testing. This protocol assesses the generalizability of the model across varying urban landscapes, architectural styles, and geographic locations, thereby evaluating its robustness and adaptability in real-world applications.

4. Methodology

Here, we propose a two-step method as illustrated in Fig. 3. The first step focuses on aligning SIs with ground-level views. Reference SIs undergo polar transformation, converting them into polar-transformed SIs to facilitate cross-view alignment. Following this, spatial feature calibration is applied to the transformed SIs to ensure consistency in feature representation, preparing them for subsequent feature fusion process. In the second step, the weight-shared ConvNeXt-Base variant is used to extract features from CSIs, SVIs, and aligned SIs. Street-view and satellite features are fused to create comprehensive multi-view representations, and the correlation between query and fused features is calculated to identify the best match and determine the GPS location. In the subsequent subsections, we provide detailed explanations for each process step.

4.1. Problem formalization

Given a query CSI with unknown location information, our objective is to determine its geographic coordinates (latitude and longitude) by matching it against a database of geo-tagged reference images. During the training phase, each query CSI is paired with its corresponding SI and SVI from the same location to form a positive sample, while all other SI-SVI pairs from different locations are treated as negative samples. The goal is to learn a robust feature representation that can effectively distinguish positive pairs from negative pairs by maximizing the similarity between features of positive pairs and minimizing the similarity for negative pairs. The inputs to our training model are a query CSI and a candidate SI-SVI pair. The output is the probability that the query image and the candidate SI-SVI pair are spatially close to each other. During the testing phase, the input is a query CSI, and the output is the estimated GPS location, which is determined by matching the query image against the reference dataset.

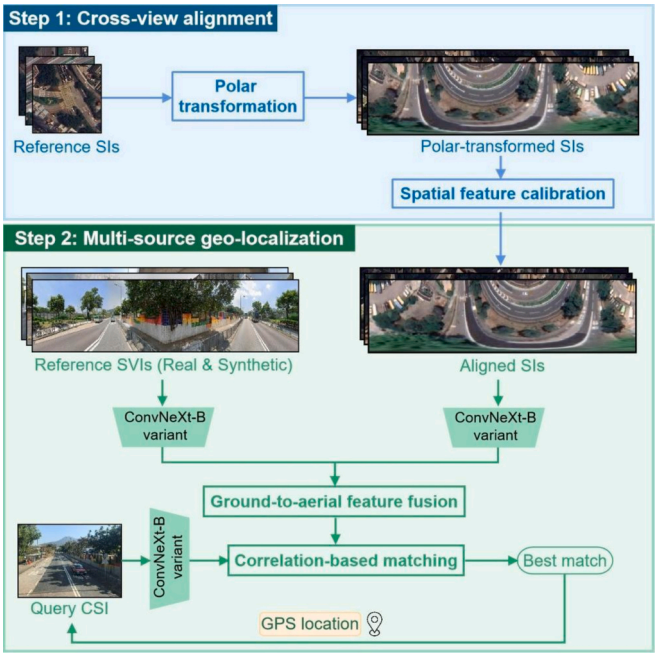

Fig. 3. Flowchart of the study. Firstly, reference SIs undergo a polar transformation and spatial feature calibration for feature alignment with SVIs. Then, the aligned SIs, SVIs, and CSIs are input into the multi-source geo-localization model to find the best match for the query CSI, providing its GPS location data.

4.2. Cross-view alignment

To bridge the gap between SIs and SVIs, we first perform polar transformation on SIs, which is an established technique in cross-view localization. However, due to the unknown northward direction in SIs and SVIs, directly fusing the polar-transformed SIs with SVIs may degrade localization accuracy. To address this, we propose a novel feature calibration method that calculates the offset between SIs and SVIs, and then crops and stitches the SIs based on this offset. This process facilitates effective feature fusion between the two image types. The complete workflow, including both the standard polar transformation and our proposed feature calibration, is visually summarized in Fig. 3 (Step 1) for clarity. For further intuitive understanding, Fig. 4 (a)-(c) provides visual examples comparing SIs before polar transformation, after transformation, and after calibration, clearly illustrating the effects of each step.

4.2.1. Polar transformation

We utilize the polar transformation to convert SIs from Cartesian coordinates to polar coordinates. This transformation helps to bridge the gap between perspective and spatial representation, providing a more accurate basis for comparison and matching in subsequent feature extraction and alignment steps. When the scene is planar, the horizontal lines that span the entire 360° FoV in the street-view panorama appear as circles in the SI, while the vertical lines correspond to rays starting from the center of it. The application of polar transformation can roughly align the space layout of cross-view images.

Specifically, the center of the SI is selected as the polar origin for the polar transformation. In geographic coordinates, the north direction aligned with the positive y-axis of the SVI is set to 0°. The polar transformed image is resized to match the SVI. In the transformed image, each column corresponds to the same angle as the SVI. For a more homogeneous representation, a radial line sampling strategy was applied to the SIs. This ensures that the outermost and innermost circles of the SI are mapped to the top and bottom lines of the transformed image, respectively. The polar transformation between the original points (x_i^s, y_i^s) and polar-transformed ones (x_i^t, y_i^t) is then defined as:

$$x_i^s = \frac{S_a}{2} - \frac{S_a}{2}\frac{(H_g - x_i^t)}{H_g}\cos\cos\left(\frac{2\pi}{W_g}y_i^t\right), \tag{1}$$

where S_a is the size of the original SI; H_g and W_g denote the height and width of the target polar-transformed image, respectively. For each pixel in the target polar-transformed image, we use these forward mapping equations to identify its corresponding sampling location in the original SI. The pixel value at this location is then obtained via bilinear interpolation and assigned to the target pixel. The application of a polar transformation enables the bridge of the domain gap between the geometric features of cross-view images, as illustrated in Fig. 4 (a) and (b).

4.2.2. Spatial feature calibration

The polar transformation alone does not fully resolve the orientation misalignment between SVIs and polar-transformed SIs, primarily because their relative orientations remain uncalibrated. Directly fusing the polar-transformed SIs with the corresponding SVIs without proper alignment can significantly degrade localization accuracy. To address this, we adopt an ORB-based feature matching approach for orientation

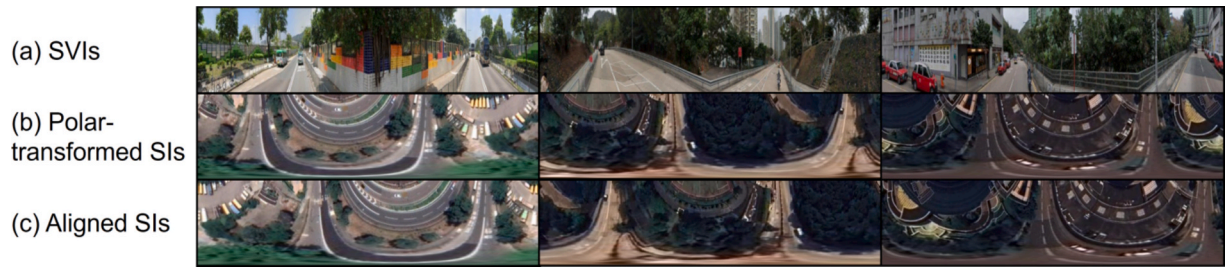

Fig. 4. Examples of ground-to-aerial images. (a) SVIs. (b) Polar-transformed SIs. (c) Polar-transformed SIs after calibration.

calibration, as ORB demonstrates superior efficiency and accuracy compared to alternative methods. Further justification for this choice is provided in Appendix B. First, keypoints and corresponding binary descriptors are extracted from the SVI and the polar-transformed SI using the ORB feature detector (Rublee et al., 2011). Then, a brute-force matcher with Hamming distance (Norouzi et al., 2012) is used to establish correspondences between the descriptors. Subsequently, the average horizontal offset of the matched keypoints in the polar-transformed domain is computed, indicating the relative rotational difference between the two images. Finally, the polar-transformed SI is circularly shifted horizontally by the computed offset, with wrap-around such that pixels moving beyond one boundary reappear at the opposite edge. This single operation aligns the orientation of the SI with that of the SVI, as illustrated in Fig. 4 (b) and (c).

By comparing Fig. 4 (b), which represents the images before calibration, with Fig. 4 (c), which shows the results after the proposed feature matching-based calibration, it is evident that this step effectively addresses the geometric disparities between two views. In Fig. 4 (b) and (a), there is a noticeable miscalibration between the corresponding features and structures in the two images. In contrast, Fig. 4 (c) demonstrates a significantly better calibration, where the key visual cues are well-matched between the street-view and polar-transformed satellite perspectives. In our implementation, both polar transformation and spatial feature calibration are used as pre-processing steps to reduce computational time in training and testing steps.

4.3. Multi-source geo-localization

Applying a translation offset to a polar-transformed image along its x-axis equates to rotating the SI. Therefore, learning the translational equivariant features has replaced the challenge of learning the rotational equivariant features. As shown in Fig. 5, weight-shared ConvNeXt-Base variant was used to extract features from CSIs, SVIs, and aligned SIs. Street-view and aligned satellite features were fused into new features, and the correlation between crowd-sourced and new features was calculated along the azimuth angle axis. The fusion features corresponding to the FoV of the CSIs were cropped at the location with the highest similarity score. Finally, InfoNCE loss (Oord et al., 2018) was applied to optimize the model, ensuring that the features are effectively learned, and similar images are closer in feature space.

4.3.1. Spatial feature extraction

After evaluating a range of feature extraction methods and conducting a comprehensive comparative analysis (see Appendix C), we adopted a weight-shared ConvNeXt-Base variant as our backbone due to its superior accuracy compared to alternative architectures. We utilized the first three stages of ConvNeXt-Base to extract features, striking a balance between semantic richness and computational efficiency. To address potential distortions introduced by the polar transformation, we introduced a series of convolutional layers that progressively reduce the height of the feature maps while preserving their width. These layers employ asymmetric kernels and strides, making the extracted features more robust against vertical distortions and retaining horizontal spatial information. At the same time, these layers decrease the channel dimension of the features to ensure that the fused features F_f match the channel count of the crowd-sourced features F_{CSI}, ultimately reducing the channel number of crowd-sourced features F_{CSI} to 16 (producing a feature with a size of $4 \times W \times 16$), and reducing the channel number of the street-view features F_{SVI} and polar-transformed features F_{SI} to 8 (producing a feature with a size of $4 \times 16 \times 8$). The width of the crowd-sourced features is W instead of 16 because the CSIs constrain the visible area to a more limited, focused FoV rather than a full 360-degree

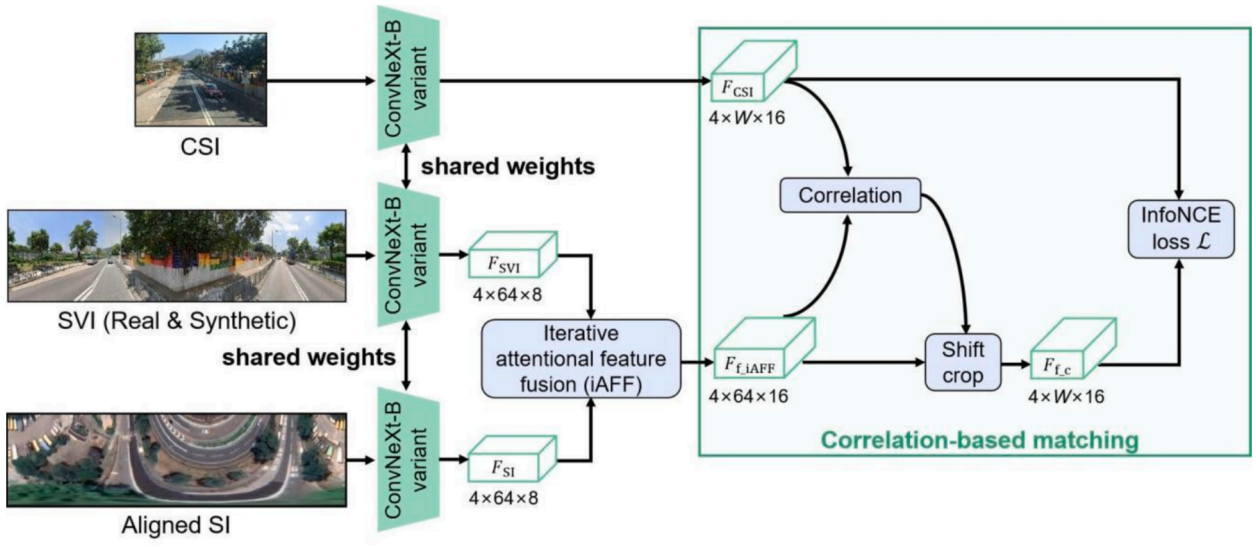

Fig. 5. The network architecture of the proposed multi-source geo-localization model.

perspective. In contrast, the polar-transformed SIs and SVIs are both panoramic images.

By integrating these techniques, the model is better suited for applications requiring accurate feature extraction from images with potential distortions, ultimately improving the performance in image matching scenarios.

4.3.2. Ground-to-aerial feature fusion

Upon finishing the feature extraction process, we fused the corresponding two features to better utilize spatial information from different sources. Conventional feature fusion methods such as concatenation or addition do not consider the applicability of fusion for specific objects. To guarantee scale consistency during the feature fusion process, we adopted iterative attentional feature fusion (iAFF), which enhances the attentional feature fusion (AFF) framework by incorporating iterative mechanisms (Dai et al., 2021). The iAFF introduces a multi-scale channel attention module (MS-CAM) block, which captures channel-wise dependencies and attends to features at different scales. In this context, "multi-scale" does not mean that multi-scale feature maps are explicitly extracted from different network layers or with different receptive fields. Instead, the MS-CAM module aggregates both global context (via global average pooling) and local context (via point-wise convolution) from the same feature map, enabling the attention mechanism to adaptively focus on information at different spatial extents within the channel dimension. By leveraging channel-wise attention, MS-CAM dynamically adjusts the importance of different channels to focus on more informative features for the task. A visual representation of MS-CAM, AFF, and iAFF can be found in Fig. 6. Based on the MS-CAM block M, iAFF can be formulated as

$$F_f = M(F_{SI} \uplus F_{SVI}) \otimes F_{si} + (1 - M(F_{SI} \uplus F_{SVI})) \otimes F_{SVI} \tag{2}$$

where $\otimes$ indicates the element-wise multiplication, $\uplus$ denotes the initial feature integration, and $M(F_{SI} \uplus F_{SVI})$ denotes the attentional weights generated by MS-CAM. In AFF, the initial integration $\uplus$ is element-wise summation. In iAFF, $F_{SI} \uplus F_{SVI}$ in Eq. (3) can be reformulated as

$$F_{SI} \uplus F_{SVI} = M(F_{SI} + F_{SVI}) \otimes F_{SI} + (1 - M(F_{SI} + F_{SVI})) \otimes F_{SVI} \tag{3}$$

4.3.3. Correlation-based matching

The orientation of CSIs is not always available, and the limited FoV of these images further increases the difficulty of geo-localization. When humans use maps to locate their own positions, they match the observed environment with the information on the map to jointly determine their position and orientation. To simulate this process in the network, we calculated the correlation between the crowd-sourced and fused features in the azimuthal dimension. More specifically, we calculated the inner product between crowd-sourced and fused features across all possible orientations by using crowd-sourced features as sliding windows.

Let $F_f \in R^{H \times W_f \times C}$ and $F_{CSI} \in R^{H \times W_{CSI} \times C}$ represent the fused and crowd-sourced features respectively, where C and H denote the channel and height numbers, W_f and W_{CSI} denote the width of the fused and crowd-sourced features. The relationship between F_f and F_{CSI} can be

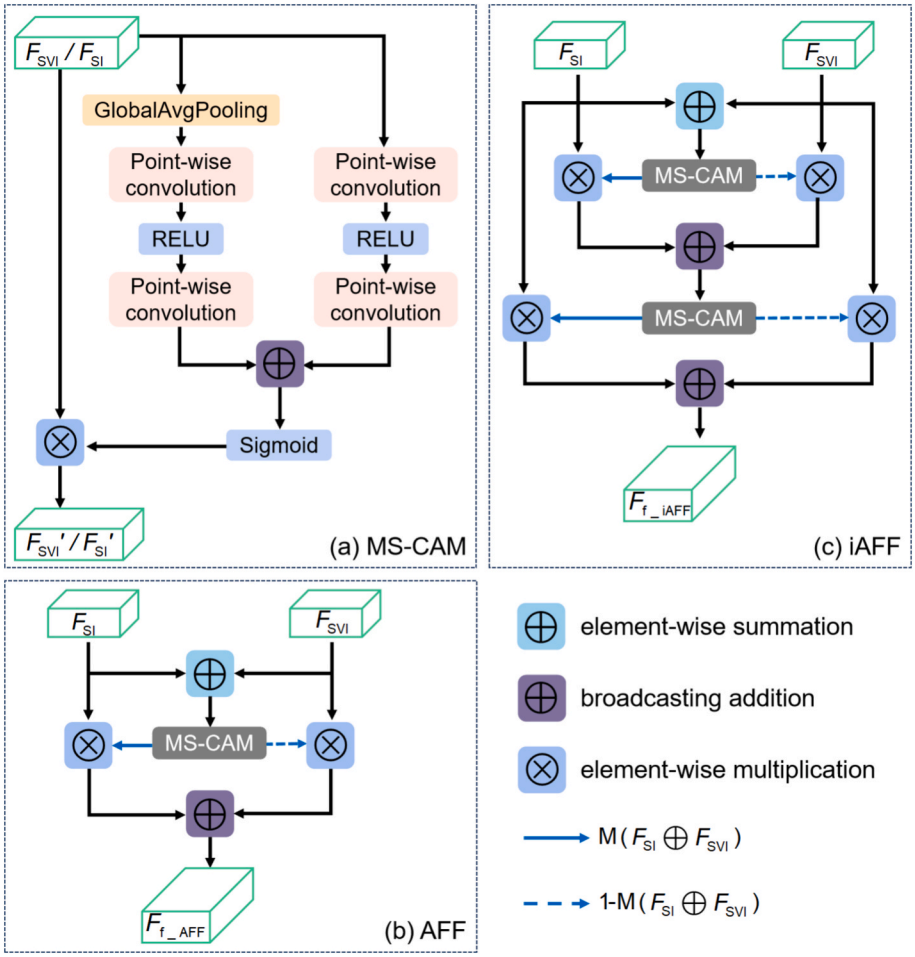

Fig. 6. Structure diagram of MS-CAM, AFF, and iAFF. (a) MS-CAM: Using point-wise convolution to aggregate global and local feature contexts. (b) AFF: Using MS-CAM to fuse satellite features (F_{SI}) and street-view features (F_{SVI}). (c) IAFF: Building upon AFF by iteratively feeding the fused output (F_f) back into the process to further refine the feature integration.

represented as

$$[F_f * F_{CSI}](i) = \sum_{c=1}^{C} \sum_{h=1}^{H} \sum_{\omega=1}^{W_{CSI}} F_f(h, mod(i+\omega, W_{CSI}), c) F_{CSI}(h, \omega, c) \quad (4)$$

where operator mod represents the modulo operation, and the $F(h, \omega, c)$ is feature response at index (h, ω, c). With the correlation calculated, fused features aligned with the FoV of the CSIs were cropped at the location of the highest similarity score. Subsequently, the cropped fused features F_{f_crop} were re-normalized. By calculating the L_2 distance between the crowd-sourced and fused features, the similarity score for matching was determined. If multiple maximum similarity scores are found, a random selection is made, indicating that the SIs contain indistinguishable symmetries. To facilitate end-to-end training, in our implementation, the hard maximum selection is approximated by a differentiable Gumbel-Softmax operator (Jang et al., 2016), which enables gradient flow through the alignment step.

For the geo-localization task, triplet loss is primarily employed (Schroff et al., 2015), along with several variants such as soft-margin triplet loss (Vo and Hays, 2016) and weighted soft-margin triplet loss (Hu et al., 2018). In the triplet loss framework, each anchor is compared to one positive example and one negative example, which is very prone to model collapsing. While soft-margin triplet loss improves upon the standard triplet loss by smoothing the optimization process and reducing the likelihood of model collapse, it still fundamentally depends on identifying hard negatives. This dependency can be limiting in large-scale, cross-view geo-localization tasks, where the diversity of negatives—rather than their hardness—plays a more critical role in bridging the domain gap between ground-level and aerial imagery. Recent cross-view geo-localization studies (Deuser et al., 2023; Ye et al., 2024) have therefore moved to InfoNCE loss (Oord et al., 2018), which leverages all negatives in a batch. The temperature-scaled softmax further stabilizes gradients and enhances the contribution of relevant negatives, yielding richer negative supervision than margin-based triplet losses. The key advantage of InfoNCE loss is its ability to include all negatives in the batch during optimization, which allows the model to learn from a broader range of examples. This not only enhances generalization across diverse urban contexts but also mitigates the instability issues often associated with hard-negative mining in triplet loss. Consequently, to leverage these advantages and effectively utilize the rich negative information within a batch, we adopt the InfoNCE loss in this work. Specifically, for each anchor (i.e., a query CSI), its positive sample is defined as the SI and SVI pair from the same location. All other SI-SVI pairs in the batch that originate from different locations serve as negative samples. The training loss is defined as follows:

$$\mathscr{L}(q, R) = -log\left(\frac{exp(q \bullet r_+/\tau)}{\sum_{i=0}^{R} exp(q \bullet r_i/\tau)}\right) \quad (5)$$

where q denotes the crowd-sourced query, and R is a set of references. The InfoNCE loss is low when q and the positive match r_+ are similar, while it increases when they are dissimilar. The temperature parameter τ can either be a learnable value or set to a fixed static value. To demonstrate the superiority of InfoNCE loss for this task, we conducted a comparative experiment with triplet loss and its variants in Appendix D.

5. Results

5.1. Experimental details

All experiments were conducted on one Nvidia GeForce RTX 3080 GPU. The code was implemented in Python (version 3.7.0) and utilized TensorFlow (version 2.7.0) with GPU support as the framework for contrastive learning. Before being fed into the network, street-view panoramas and polar-transformed SIs were resized to 128×512. The Adam optimizer (Kingma and Ba, 2014) was employed with a learning rate of 5×10^{-5}.

5.2. Evaluation metrics

In this work, we evaluate geo-localization performance using recall@N with a 25 m threshold, following widely adopted protocols in previous work (Liu et al., 2018; Ge et al., 2020; Warburg et al., 2020; Berton et al., 2022). Recall@N is defined as the percentage of queries for which at least one of the top-N retrieved candidates is within a specified distance threshold from the ground truth location. Specifically, a retrieval is considered correct if at least one of the top-N predictions falls within 25 m of the ground truth.

5.3. Comparison with other methods

5.3.1. Comparison with other fusion-based geo-localization methods

To verify the superiority of the proposed method in image geo-localization, we compared our method with several state-of-the-art fusion-based approaches on the proposed MSGD. Regmi and Shah (2019) generated synthetic SIs from ground-level query images, subsequently integrating features from both sources. Shi et al. (2022) applied polar and projective transforms to SIs, extracting and concatenating features from both. Pro et al. (2024) proposed a Semantic Align Net (SAN) and Semantic Channel Net (SCN) to combine SIs with their semantic segmentation masks, as additional information. Differently from SAN, the concatenation in SCN is done at the image level and not at the features level.

For a fair comparison, all query images in these methods are CSIs. Table 2 shows that in the same area setting, our method achieves the best performance across all metrics, illustrating its effectiveness in leveraging urban features and architectural cues within familiar environments. In the more challenging cross area setting, which assesses the generalization ability to unseen cities, our method also outperforms all baselines, despite the performance degradation typically caused by domain shifts in urban structure and appearance. The performance improvements in both settings can be mainly attributed to the fusion of detailed ground-level features and broad aerial perspectives, which together provide a more comprehensive and robust representation of the urban environment.

Fig. 7 qualitatively compares the Top-1 retrieval results of the four fusion-based methods and our method under real-world conditions. Evidently, our method consistently retrieves the most visually coherent match, even under limited FoV or viewpoint variations, whereas the competing methods frequently yield noticeable misalignments. These observations corroborate the quantitative superiority reported in Table 2.

5.3.2. Comparison with other feature fusion methods

This section focuses on analyzing the performance of our model using different feature fusion methods: (1) concatenation; (2) fully connected layer (FCL); (3) attention-weighted average (AWA); (4) AFF and (5) iAFF. According to Table 3, iAFF exhibits the best performance. Compared to AFF, it incorporates a more complex fusion module to better utilize the information from different features, resulting in the highest geo-localization accuracy. The performance of AFF is second only to iAFF. Concat outperforms FCL and AWA, which may be attributed to its simple and direct feature connection method, allowing it to better preserve the original information of multi-source features. In contrast, FCL performs complex nonlinear transformations on features, which increases feature confusion and leads to the loss of effective information from the original features. As Tolias et al. (2016) suggest, the impact of FCL on image retrieval tasks is usually negative.

Table 2
Comparison with other fusion-based methods.

Methods		Same area				Cross area			
		R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
Regmi and Shah (2019)		28.34 %	49.27 %	62.13 %	79.02 %	13.19 %	28.70 %	41.55 %	53.41 %
Shi et al., (2022)		44.51 %	68.22 %	79.76 %	87.16 %	23.67 %	33.29 %	52.44 %	60.58 %
Pro et al., (2024)	SCN	38.12 %	61.48 %	75.34 %	84.21 %	19.44 %	30.37 %	46.28 %	57.25 %
	SAN	47.38 %	71.06 %	81.81 %	88.37 %	25.83 %	36.55 %	55.09 %	62.70 %
Our method		**54.27 %**	**75.31 %**	**84.64 %**	**92.03 %**	**32.12 %**	**46.82 %**	**63.51 %**	**74.44 %**

Fig. 7. Qualitative comparison of Top-1 retrieval results. Each column presents one query CSI (top row) followed by the corresponding Top-1 retrieval SVI-SI pairs returned by Regmi and Shah (2019), Shi et al. (2022), SCN, SAN, and our method. Green and red boxes indicate correct and incorrect retrievals, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Comparison among different feature fusion methods.

Methods	Same area				Cross area			
	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
Concatenation	48.47 %	68.00 %	77.67 %	88.98 %	28.67 %	41.14 %	55.10 %	70.27 %
FCL	44.04 %	62.57 %	71.53 %	85.33 %	26.03 %	37.82 %	50.72 %	67.25 %
AWA	48.36 %	67.85 %	77.41 %	88.87 %	28.62 %	41.01 %	54.92 %	70.12 %
AFF	48.93 %	68.74 %	78.26 %	89.32 %	28.92 %	41.44 %	55.63 %	70.75 %
iAFF	**54.27 %**	**75.31 %**	**84.64 %**	**92.03 %**	**32.12 %**	**46.82 %**	**63.51 %**	**74.44 %**

5.4. Ablation analysis

5.4.1. Impact of synthetic SVIs on geo-localization performance and coverage

In practical geolocation tasks, a significant portion of CSI samples lack corresponding SVIs, either due to incomplete datasets (e.g., missing SVIs from certain regions) or substantial temporal mismatches (e.g., outdated SVIs relative to recent CSIs). To address these gaps, we utilize SIs to generate synthetic SVIs using the PanoGAN architecture (Wu et al., 2022). Synthetic SVIs provide contextual information (e.g., street-level visual details and perspectives) that complements the overhead spatial structures captured by SI. By fusing the two, we can leverage their complementary strengths to enhance geolocation performance. Since the synthesis process may introduce artifacts, such as visual distortions, unrealistic textures, or geometric inconsistencies, it is important to evaluate how these artifacts affect localization performance. We begin by quantifying the proportion of CSI samples in our study that do not have corresponding SVI. To analyze the effectiveness and limitations of using synthetic SVI, we conduct comparative experiments under the following settings:

- Real SVIs (Strict): Localization is performed only on CSI samples with available SVI, representing the upper bound of performance when all necessary data is available.
- Real SVIs (Hybrid): For CSI samples with available SVI, fusion with SI is used. For those without SVI, only SI is used for localization, demonstrating the limitations of incomplete SVI coverage.
- Nearest SVIs: Each CSI sample is paired with the nearest available SVI and SI, without filtering for temporal or spatial mismatch, as in previous studies.
- Our method: For CSI samples lacking available SVI, synthetic SVI is generated from SI and fused for localization. For samples with available SVI, the real SVI-SI pair is used, achieving a balance between coverage and accuracy.

As shown in Table 4, fusing real SVI and SI (Strict SVIs) achieves the highest localization accuracy but is limited to only 62.45 % of the dataset due to incomplete SVI coverage. In contrast, methods that ensure

Table 4
Quantitative model performance with and without synthetic SVI.

Methods	Same area				Cross area				Coverage
	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%	
Strict SVIs	**60.62 %**	**80.96 %**	**90.51 %**	**96.25 %**	**37.90 %**	**51.50 %**	**68.37 %**	**79.59 %**	62.45 %
Hybrid SVIs	46.34 %	64.78 %	76.90 %	85.32 %	26.44 %	39.95 %	57.18 %	67.10 %	**100 %**
Nearest SVIs	48.18 %	68.05 %	78.23 %	87.40 %	28.57 %	42.29 %	58.92 %	70.16 %	**100 %**
Our method	54.27 %	75.31 %	84.64 %	92.03 %	32.12 %	46.82 %	63.51 %	74.44 %	**100 %**

full coverage, such as SI alone (Hybrid SVIs) or unfiltered nearest SVI pairing, suffer from significantly reduced accuracy (e.g., R@1 decreases to 46.34 % and 48.18 % in the Same Area scenario, respectively). Our proposed method, which generates synthetic SVIs for missing references and fuses them with SI, achieves a balance between accuracy and coverage. Specifically, while the accuracy (R@1) decreases slightly from 60.62 % (Strict) to 54.27 % (Our method) in the Same Area scenario, our method achieves 100 % coverage and significantly outperforms other full-coverage methods (e.g., Hybrid: 46.34 %; Nearest: 48.18 %). This demonstrates that synthetic SVIs effectively complement SI and provide substantial localization performance improvements compared to alternative approaches.

5.4.2. *Effectiveness of spatial feature calibration*

Even though our method performs better than using only a single data source for geo-localization, incorporating the spatial feature calibration block can further enhance the effectiveness of ground-to-aerial feature fusion and the positing result. To investigate the contribution of the spatial feature calibration block (dubbed as SFC in Table 5), we conducted an ablation experiment by removing this process from our pre-processing steps. Table 5 shows that introducing the spatial feature calibration block obviously improves the experimental results. This spatial explicit correction between SVI and RS images can contribute to the effective cross-view feature fusion.

5.4.3. *Advantages of multi-view feature fusion*

To showcase the effectiveness of feature fusion in improving CSI geo-localization, we compared our method with using only a single data source (SVIs or SIs). The results in Table 6 demonstrate that our fusion method outperforms the single-data-source approaches ("Only SVI" and "Only SI") across all evaluation metrics in both same-area and cross-area scenarios, highlighting the benefits of data fusion in image geo-localization.

Table 6 further shows that using SIs alone outperforms using SVIs alone across various evaluation metrics, indicating that SIs make an obvious contribution to the fusion-based geo-localization method. This contradicts the conventional assumption that SVIs, sharing a similar perspective with CSIs, should yield better localization results. This may be attributed to the two key limitations of SVIs: (1) occlusions that restrict comprehensive environmental perception, and (2) artifacts by the synthesis of SVIs. In contrast, SIs provide broader coverage and more stable natural features, making them more reliable for geo-localization tasks. This finding highlights the crucial role of SIs in multi-view geo-localization systems.

The activation maps of satellite, street-view, and crowd-sourced features are visualized in Fig. 8, which are computed using gradient-weighted class activation mapping (Grad-CAM) (Selvaraju et al., 2017). It can be seen that satellite features focus more on road features and street-view features place greater emphasis on architectural features. Crowd-sourced features are oriented towards both architectural and road features and are more inclined towards architectural features. The fusion of satellite and street-view features is consistent with the focus on crowd-sourced features, indicating that by fusing multi-view features, more useful information can be retained, thereby enhancing geographic positioning results. It can also be observed that satellite features provide more information, thereby enhancing the accuracy of geographic positioning.

5.4.4. *Effectiveness of correlation-based matching*

The correlation-based matching module is designed to address the challenge of unknown camera orientation in CSIs, where no prior information such as compass angle, gravity direction, or camera intrinsics is available. This module aims to automatically infer the optimal orientation of the query image by maximizing feature similarity with the reference panorama, thus enabling effective feature alignment and robust geo-localization.

However, as discussed in Section 4.3.3, ground-truth orientation labels for CSIs are generally unavailable in real-world datasets such as MSGD. Therefore, it is not feasible to directly evaluate the orientation estimation accuracy of this module. To indirectly assess its effectiveness, we perform ablation experiments comparing the overall geo-localization performance with and without the correlation-based matching module. Specifically, we remove the correlation-based matching component (dubbed as CBM in Table 7) from the pipeline and compare the resulting recall@N metrics with those obtained using the complete framework. As shown in Table 7, the absence of the correlation-based matching module leads to a significant drop in localization accuracy across all evaluation metrics. These results provide strong evidence that the correlation-based matching module is essential for robust geo-localization in realistic crowd-sourced scenarios, as it effectively addresses the challenge of unknown camera orientation and leads to significant improvements in localization accuracy.

6. Discussion

6.1. *Practical applications and broader implications*

The proposed multi-source geo-localization framework, which fuses satellite and street-view imagery, is well-suited for urban monitoring, emergency response, and urban planning. This approach enables robust detection of urban infrastructure changes and addresses data sparsity by localizing crowdsourced images in regions with limited or outdated street-view coverage, particularly in rapidly developing or disaster-affected areas (Li et al., 2020; Eyre et al., 2020).

The integration of synthetic SVIs generated by PanoGAN further broadens the framework's applicability by enabling geo-localization where real street-view data is unavailable. This is especially valuable for emergency response, as it allows rapid and accurate localization of

Table 5
Quantitative model performance with and without SFC.

Methods	Same area				Cross area			
	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
Our method without SFC	50.06 %	70.09 %	78.79 %	88.97 %	29.67 %	43.19 %	59.12 %	71.06 %
Our method with SFC	**54.27 %**	**75.31 %**	**84.64 %**	**92.03 %**	**32.12 %**	**46.82 %**	**63.51 %**	**74.44 %**

Table 6
Comparison with using a single data source.

Methods	Same area				Cross area			
	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
Only SVI	40.54 %	60.12 %	70.33 %	85.25 %	22.41 %	34.92 %	49.13 %	62.37 %
Only SI	44.47 %	66.12 %	76.18 %	88.24 %	26.29 %	40.08 %	54.38 %	68.13 %
Our method	**54.27 %**	**75.31 %**	**84.64 %**	**92.03 %**	**32.12 %**	**46.82 %**	**63.51 %**	**74.44 %**

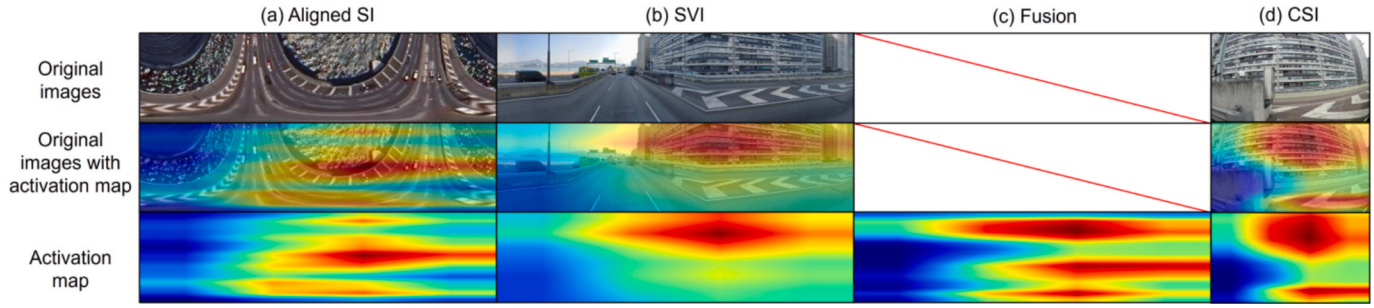

Fig. 8. Activation map visualization. (a) Polar-transformed SI and activation map. (b) SVI and activation map. (c) Fusion activation map. (d) CSI and activation map.

Table 7
Quantitative model performance with and without SFC.

Methods	Same area				Cross area			
	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
Our method without CBM	36.32 %	57.21 %	61.67 %	70.13 %	22.54 %	35.33 %	52.62 %	64.25 %
Our method with CBM	**54.27 %**	**75.31 %**	**84.64 %**	**92.03 %**	**32.12 %**	**46.82 %**	**63.51 %**	**74.44 %**

citizen-contributed images in inaccessible regions, thereby supporting effective disaster management (Li et al., 2024).

Beyond these domains, our method also supports geographic analysis of social media imagery, enabling applications such as tourism analytics, event detection, and urban behavior studies (Shao et al., 2017). Importantly, the release of the MSGD dataset provides a comprehensive benchmark for advancing geo-localization methods and fostering innovation in urban spatial analytics.

6.2. Limitations and future research directions

Although substantial progress has been made in this research, several limitations persist and warrant further investigation. The first limitation is that the ORB-based matching method may limit the robustness of alignment in scenarios with significant viewpoint variations or complex scene semantics. Secondly, while our spatial feature calibration module successfully aligns the orientations of satellite and street-view features, facilitating subsequent feature fusion, it does not address the misalignment between the center of the SI and the camera location of the SVI. Thirdly, the use of polar transformation assumes a planar ground scene, which oversimplifies the complex 3D geometry of urban environments. This assumption may introduce distortions or fail to account for occlusions caused by tall vertical structures in dense urban areas. While the spatial feature calibration module and distortion-resistant feature extraction method mitigate some of these challenges, future work could focus on developing 3D-aware transformations that explicitly account for vertical structures, such as incorporating depth information or height maps.

Advancements in generative models and large language models (LLMs) offer promising research directions that can be explored to further enhance the proposed framework. Firstly, generative models are promising for improving synthetic SVI quality. Leveraging advanced image generation models such as diffusion models, we could reduce artifacts and enhance realism in synthetic SVIs, thereby improving localization performance applicability in regions with limited street-view data. Secondly, the use of multimodal LLMs in conjunction with visual models is an emerging trend in the field of geo-localization. Multimodal LLMs can not only understand the complex contextual information in the images, but also possess the reference abilities to enhance the robustness of geo-localization (Wang et al., 2025; Yin et al., 2025). LLM-enhanced geo-localization will be a promising direction to solve the geo-localization problem better.

7. Conclusions

This study presents a comprehensive solution for CSI geo-localization through three key contributions. Firstly, we introduced a novel multi-source data fusion framework that significantly improves the geo-localization accuracy of CSIs through contrastive learning, thus advancing the field beyond traditional single-source approaches. This methodological innovation demonstrates how integrating different perspectives of urban imaging can effectively enhance geo-localization accuracy. Secondly, our approach addresses the limitation of existing methods by leveraging synthetic SVIs, enabling CSI geo-localization in areas where SVIs are non-existent or outdated. Thirdly, we introduced the MSGD, a dataset comprising over 500k precisely geo-tagged pairs of CSIs, SVIs, and SIs from 12 cities across 6 continents. This dataset provides researchers with a comprehensive testbed for evaluating CSI geo-localization methods and multi-source data fusion approaches. It fills a crucial gap in the literature by offering aligned data from multiple perspectives, enabling future research in areas such as urban measurements (Suel et al., 2021), urban village identification and classification (Chen et al., 2022; Fan et al., 2022), and local climate zone mapping (Cao et al., 2023).

Overall, this work offers several key insights: (1) the effectiveness of contrastive learning in handling multi-source data; (2) the potential of multi-source data fusion in enhancing geo-localization accuracy; and (3) CSI geo-localization is still challenging in regions with varying

architectural styles and geographic characteristics. Additionally, the improved geo-localization accuracy achieved through our framework can significantly enhance location-based services, urban navigation systems, and emergency response applications. For instance, accurately locating CSIs can assist in urban emergency management by precisely identifying incident locations and supporting urban planning by better understanding citizen-reported issues.

Declaration of AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used ChatGPT to increase the readability of the text. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

CRediT authorship contribution statement

Qianbao Hou: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis. **Ce Hou:** Writing – original draft, Data curation. **Fan Zhang:** Data curation, Conceptualization. **Qihao Weng:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research has received funding from Global STEM Professorship, Hong Kong SAR Government (P0039329) and Hong Kong Polytechnic University (P0046482 and P0038446).

Appendix A. Comparison with other feature detectors for SFC

To justify our choice of feature detector in the SFC module, we conducted a controlled experiment comparing ORB, SIFT (Lowe, 1999), and SuperPoint (DeTone et al., 2018) under identical pipeline settings. As summarized in Table A1, ORB consistently outperforms both SIFT and SuperPoint in terms of recall while also providing higher computational efficiency. Although SIFT and SuperPoint typically detect more keypoints, this does not yield better calibration results in our scenario. Our task reduces to estimating a single azimuthal shift on polar-transformed panoramas, where the stability of the dominant horizontal offset is more important than the quantity of keypoints or sub-pixel geometric precision. In summary, ORB achieves the best balance of accuracy and efficiency for SFC in our pipeline, and is therefore adopted as the default choice.

Table A1
Comparison with other feature detectors for SFC.

Methods	Same area				Cross area				Time (ms)
	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%	
SIFT	37.92 %	52.79 %	59.23 %	64.45 %	22.41 %	32.76 %	44.48 %	52.12 %	45
SuperPoint	44.08 %	61.14 %	68.72 %	74.76 %	26.05 %	38.03 %	51.58 %	60.47 %	14
ORB	**54.27 %**	**75.31 %**	**84.64 %**	**92.03 %**	**32.12 %**	**46.82 %**	**63.51 %**	**74.44 %**	**3**

Appendix B. Quantitative evaluation of the cross-view panorama image generation

To further investigate the quality of synthetic SVIs, we conduct a comprehensive quantitative comparison between PanoGAN model and several representative baseline methods, including Pix2pix (Isola et al., 2016), X-Fork (Regmi and Borji, 2018), X-Seq (Regmi and Borji, 2018), and SelectionGAN (Tang et al., 2020). For this evaluation, we follow the standard metrics used in previous works: Inception Score, which reflects both the diversity and realism of generated images based on a pre-trained classifier (higher is better); Prediction Accuracy, which measures the semantic consistency between generated and ground-truth images (higher is better); KL Score (Kullback-Leibler divergence), which quantifies the feature distribution difference between generated and real images (lower is better); SSIM (Structural Similarity Index), which evaluates the structural similarity at the pixel level (higher is better); PSNR (Peak Signal-to-Noise Ratio), which assesses the pixel-level fidelity with higher values indicating less distortion; and SD (Sharpness Difference), which measures the similarity in image sharpness between generated and real images (higher is better).

Table B1 presents the quantitative results for all methods. The findings demonstrate that PanoGAN achieves competitive or superior performance compared to other baseline models across most evaluation metrics, indicating its effectiveness in generating realistic and semantically meaningful panorama images. However, there remains a measurable difference between generated and real images, especially in terms of fine structural details and sharpness. These results provide a comprehensive assessment of the strengths and limitations of current cross-view panorama generation models, and offer a clearer understanding of the performance gap between synthetic and real-world SVIs.

Table B1
Comparison between generated and real SVIs. (*) Inception Score for real SVIs is 4.9498, 3.3464, and 5.0254 for all, top-1, and top-5 setups, respectively.

Methods	Inception score* ↑			Accuracy ↑		KL ↓	SSIM ↑	PSNR ↑	SD ↑
	All	Top-1	Top-5	Top-1	Top-5				
Pix2pix	3.55	2.42	3.85	22.85	43.21	10.51 ± 1.78	0.3811	19.8378	18.8799
X-Fork	3.86	2.82	4.16	31.68	64.22	6.26 ± 1.45	0.4526	20.8987	19.6318
X-Seq	3.35	2.65	3.65	27.05	60.34	7.05 ± 1.65	0.4429	20.7854	19.5185
SelectionGAN	3.68	2.72	3.98	29.87	62.05	6.14 ± 1.42	**0.4547**	21.7536	**20.4571**
PanoGAN	**3.95**	**2.94**	**4.25**	**38.45**	**73.52**	**4.30 ± 1.27**	0.4535	**22.2789**	20.3734

Appendix C. Comparison with other feature extraction backbones

In order to determine a proper selection for our approach, we compared R@1 and the time required to locate a crowd-sourced query by replacing the weight-shared ConvNeXt-Base variant with other deep learning models. Both the ViT and the classical CNN models (i.e., VGG (Simonyan and Zisserman, 2014), ResNet (He et al., 2015), and EfficientNet (Tan and Le, 2019)) were considered. Previous approaches propose to use separate encoders for multi-source images. Therefore we also tested ConvNeXt without shared weights, but we achieved better results when using the same encoder for multi-source images. Sharing weights across all three image types (CSI, SVI, SI) is attractive for its potential to improve model robustness, especially if the CSI provides a diverse range of viewing angles, resolutions, and conditions. This can improve generalization across various image acquisition scenarios. It also reduces the total number of parameters, thereby reducing the geo-localization time. Table C1 suggests that the weight-shared ConvNext-Base variant is preferable for tasks requiring high accuracy and efficiency in localization.

Table C1
Comparison with other deep learning models.

Method	Weights shared with	R@1	Geo-localization time
VGG16	SVI, SI, CSI	37.48 %	55.1 ms
ResNet-50	SVI, SI, CSI	41.93 %	75.4 ms
EfficientNet-B0	SVI, SI, CSI	40.98 %	60.9 ms
ViT	SVI, SI, CSI	43.66 %	104.4 ms
ConvNext-Base	SVI, SI, CSI	45.04 %	118.9 ms
ConvNext-Base variant (ours)	None	48.37 %	211.7 ms
	SVI, SI	51.62 %	159.5 ms
	SVI, SI, CSI	**54.27 %**	107.3 ms

Appendix D. Comparison with other loss functions

In this Appendix, we compare the proposed InfoNCE loss with the commonly used triplet loss and its variant, soft-margin triplet loss, to demonstrate the advantages of InfoNCE loss in image geo-localization tasks. As shown in Table D1, InfoNCE loss significantly outperforms both alternatives, with triplet loss exhibiting the lowest performance across all metrics. Specifically, InfoNCE achieves a 14.93 % improvement in R@1 (Same Area) and a 21.67 % improvement in R@1 (Cross Area) compared to triplet loss, while also outperforming soft-margin triplet loss by a notable margin. These improvements demonstrate the effectiveness of InfoNCE loss in leveraging the diversity of negatives within a batch, which helps the model learn more robust and discriminative representations. Notably, the performance gap is more pronounced in the Cross Area setting, highlighting the superior ability of InfoNCE loss to generalize across varying urban landscapes, architectural styles, and geographic locations. This makes it particularly suitable for real-world geo-localization applications, where robustness and adaptability are critical.

Table D1
Comparison with other loss functions.

Methods	Same area				Cross area			
	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
triplet loss	39.34 %	47.14 %	50.96 %	55.55 %	11.45 %	16.42 %	21.98 %	28.05 %
soft-margin triplet loss	51.08 %	72.56 %	82.62 %	90.67 %	20.15 %	33.89 %	49.42 %	61.36 %
InfoNCE loss	**54.27 %**	**75.31 %**	**84.64 %**	**92.03 %**	**32.12 %**	**46.82 %**	**63.51 %**	**74.44 %**

References

Berton, G.M., Masone, C., Caputo, B., 2022. Rethinking visual geo-localization for large-scale applications. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4868–4878.

Cao, R., Liao, C., Li, Q., Tu, W., Zhu, R., Luo, N., Qiu, G., Shi, W., 2023. Integrating satellite and street-level images for local climate zone mapping. Int. J. Appl. Earth Obs. Geoinf. 119, 103323.

Chen, B., Feng, Q., Niu, B., Yan, F., Gao, B., Yang, J., Gong, J., Liu, J., 2022. Multi-modal fusion of satellite and street-view images for urban village classification based on a dual-branch deep neural network. Int. J. Appl. Earth Obs. Geoinf. 109, 102794.

Cheng, L., Yuan, Y., Xia, N., Chen, S., Chen, Y., Yang, K., Ma, L., Li, M., 2018. Crowd-sourced pictures geo-localization method based on street view images and 3D reconstruction. ISPRS J. Photogramm. Remote Sens. 141, 72–85.

Clark, B., Kerrigan, A., Kulkarni, P.P., Cepeda, V.V., Shah, M., 2023. Where we are and what we're looking at: Query based worldwide image geo-localization using hierarchies and scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 23182–23190.

Dai, M., Hu, J., Zhuang, J., Zheng, E., 2022. A transformer-based feature segmentation and region alignment method for UAV-view geo-localization. IEEE Trans. Circuits Syst. Video Technol. 32 (7), 4376–4389.

Dai, Y., Gieseke, F., Oehmcke, S., Wu, Y., Barnard, K., 2021. Attentional feature fusion. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 3559–3568.

DeTone, D., Malisiewicz, T., Rabinovich, A., 2018. SuperPoint: Self-supervised interest point detection and description. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 337–33712.

Deuser, F., Habel, K., Oswald, N., 2023. Sample4Geo: Hard negative sampling for cross-view geo-localisation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 16801–16810.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2020. An image is worth 16×16 words: Transformers for image recognition at scale. arXiv: 2010.11929.

Eyre, R.W., De Luca, F., Simini, F., 2020. Social media usage reveals recovery of small businesses after natural hazard events. Nat. Commun. 11 (1), 1629.

Fan, R., Li, J., Li, F., Han, W., Wang, L., 2022. Multilevel spatial-channel feature fusion network for urban village classification by fusing satellite and streetview images. IEEE Trans. Geosci. Remote Sens. 60, 1–13.

Fan, Z., Feng, C., Biljecki, F., 2024. Coverage and bias of street view imagery in mapping the urban environment. Comput. Environ. Urban Syst. 117, 102253.

Fervers, F., Bullinger, S., Bodensteiner, C., Arens, M., Stiefelhagen, R., 2022. Uncertainty-aware vision-based metric cross-view geolocalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 21621–21631.

Fervers, F., Bullinger, S., Bodensteiner, C., Arens, M., Stiefelhagen, R., 2024. Statewide visual geolocalization in the wild. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 438–455.

Ge, Y., Wang, H., Zhu, F., Zhao, R., Li, H., 2020. Self-supervising fine-grained region similarities for large-scale image localization. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 369–386.
He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778.
Heipke, C., 2010. Crowdsourcing geospatial data. ISPRS J. Photogramm. Remote Sens. 65 (6), 550–557.
Hou, C., Li, Y., Zhang, F., 2024. Sensing Urban Physical Environment with GeoAI and Street-Level Imagery. In: Handbook of Geospatial Approaches to Sustainable Cities. CRC Press, pp. 3–30.
Hu, S., Feng, M., Nguyen, R.M.H., Lee, G.H., 2018. CVM-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7258–7267.
Huang, H., Yao, X.A., Krisp, J.M., Jiang, B., 2021. Analytics of location-based big data for smart cities: Opportunities, challenges, and future directions. Comput. Environ. Urban Syst. 90, 101712.
Huang, X., Li, X., Yang, D., Zou, L., 2023. Chapter 6 - crowdsourced geospatial data in human and earth observations: opportunities and challenges. In: Stathopoulos, N., Tsatsaris, A., Kalogeropoulos, K. (Eds.), Earth Obs., Geoinformatics Geosci. Elsevier, pp. 109–129.
Huang, X., Wang, S., Lu, T., Liu, Y., Serrano-Estrada, L., 2024. Crowdsourced geospatial data is reshaping urban sciences. Int. J. Appl. Earth Obs. Geoinf. 127, 103687.
Isola, P., Zhu, J., Zhou, T., Efros, A.A., 2016. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5967–5976.
Jang, E., Gu, S.S., Poole, B., 2016. Categorical reparameterization with Gumbel-Softmax. arXiv:1611.01144.
Kang, Y., Zhang, F., Peng, W., Gao, S., Rao, J., Duarte, F., Ratti, C., 2021. Understanding house price appreciation using multi-source big geo-data and machine learning. Land Use Policy 111, 104919.
Kingma, D. P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv: 1412.6980.
Li, H., Deuser, F., Yina, W., Luo, X., Walther, P., Mai, G., Huang, W., Werner, M., 2024. Cross-view geolocalization and disaster mapping with street-view and VHR satellite imagery: a case study of hurricane IAN. ISPRS J. Photogramm. Remote Sens. 220, 841–854.
Li, H., Herfort, B., Huang, W., Zia, M., Zipf, A., 2020. Exploration of OpenStreetMap missing built-up areas using twitter hierarchical clustering and deep learning in Mozambique. ISPRS J. Photogramm. Remote Sens. 166, 41–51.
Li, S., Dragicevic, S., Castro, F.A., Sester, M., Winter, S., Coltekin, A., Pettit, C., Jiang, B., Haworth, J., Stein, A., Cheng, T., 2016. Geospatial big data handling theory and methods: a review and research challenges. ISPRS J. Photogramm. Remote Sens. 115, 119–133.
Lin, T.-Y., Cui, Y., Belongie, S., Hays, J., 2015. Learning deep representations for ground-to-aerial geolocalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5007–5015.
Liu, L., Li, H., 2019. Lending orientation to neural networks for cross-view geo-localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5617–5626.
Liu, L., Li, H., Dai, Y., 2018. Stochastic attraction-repulsion embedding for large scale image localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2570–2579.
Liu, Z., Mao, H., Wu, C., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A convNet for the 2020s. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11966–11976.
Lowe, D.G., 1999. Object recognition from local scale-invariant features. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1150–1157.
Ma, D., Fan, H., Li, W., Ding, X., 2020. The state of mapillary: an exploratory analysis. ISPRS Int. J. Geo Inf. 9 (1), 10.
Muller-Budack, E., Pustu-Iren, K., Ewerth, R., 2018. Geolocation estimation of photos using a hierarchical model and scene classification. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 563–579.
Norouzi, M., Fleet, D.J., Salakhutdinov, R.R., 2012. Hamming distance metric learning. In: Proceedings of the Neural Information Processing Systems (NeurIPS), pp. 1061–1069.
Oord, A.V., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. arXiv:1807.03748.
Pro, F., Dionelis, N., Maiano, L., Le Saux, B., Amerini, I., 2024. A semantic segmentation-guided approach for ground-to-aerial image matching. arXiv:2404.11302.
Regmi, K., Borji, A., 2018. Cross-view image synthesis using conditional GANs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3501–3510.
Regmi, K., Shah, M., 2019. Bridging the domain gap for ground-to-aerial image matching. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 470–479.
Rublee, E., Rabaud, V., Konolige, K., Bradski, G., 2011. ORB: an efficient alternative to SIFT or SURF. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2564–2571.
Schroff, F., Kalenichenko, D., Philbin, J., 2015. FaceNet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 815–823.
Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D., 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 618–626.
Shao, H., Zhang, Y., Li, W., 2017. Extraction and analysis of city's tourism districts based on social media data. Comput. Environ. Urban Syst. 65, 66–78.
Shi, Y., Yu, X., Campbell, D., Li, H., 2020. Where am I looking at? Joint location and orientation estimation by cross-view matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4064–4072.
Shi, Y., Yu, X., Liu, L., Campbell, D., Koniusz, P., Li, H., 2022. Accurate 3-DoF camera geo-localization via ground-to-satellite image matching. IEEE Trans. Pattern Anal. Mach. Intell. 45, 2682–2697.
Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.
Suel, E., Bhatt, S., Brauer, M., Flaxman, S., Ezzati, M., 2021. Multimodal deep learning from satellite and street-level imagery for measuring income, overcrowding, and environmental deprivation in urban areas. Remote Sens. Environ. 257.
Tan, M., & Le, Q.V., 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. arXiv:1905.11946.
Tang, H., Xu, D., Yan, Y., Corso, J.J., Torr, P.H., Sebe, N., 2020. Multi-channel attention selection GANs for guided image-to-image translation. IEEE Trans. Pattern Anal. Mach. Intell. 45, 6055–6071.
Toker, A., Zhou, Q., Maximov, M., Leal-Taixé, L., 2021. Coming down to earth: Satellite-to-street view synthesis for geo-localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6484–6493.
Tolias, G., Sicre, R., Jégou, H., 2016. Particular object retrieval with integral max-pooling of CNN activations. In: Proceedings of the International Conference on Learning Representations (ICLR), pp. 1–12.
Torii, A., Sivic, J., Okutomi, M., Pajdla, T., 2015. Visual place recognition with repetitive structures. IEEE Trans. Pattern Anal. Mach. Intell. 37, 2346–2359.
Viswanathan, A., Pires, B.R., Huber, D., 2014. Vision based robot localization by ground to satellite matching in GPS-denied situations. In: Proceedings of the International Conference on Intelligent Robots and Systems (IROS), pp. 192–198.
Vo, N.N., Hays, J., 2016. Localizing and orienting street views using overhead imagery. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 494–509.
Wang, C., Pan, X., Pan, Z., Wang, H., Song, Y., 2025. GRE suite: Geo-localization inference via fine-tuned vision-language models and enhanced reasoning chains. arXiv:2505.18700.
Warburg, F., Hauberg, S., López-Antequera, M., Gargallo, P., Kuang, Y., Civera, J., 2020. Mapillary street-level sequences: a dataset for lifelong place recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2623–2632.
Workman, S., Souvenir, R., Jacobs, N., 2015. Wide-area image geolocalization with aerial reference imagery. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 3961–3969.
Wu, B., Xie, L., Hu, H., Zhu, Q., Yau, E., 2018. Integration of aerial oblique imagery and terrestrial imagery for optimized 3D modeling in urban areas. ISPRS J. Photogramm. Remote Sens. 139, 119–132.
Wu, S., Tang, H., Jing, X., Zhao, H., Qian, J., Sebe, N., Yan, Y., 2022. Cross-view panorama image synthesis. IEEE Trans. Multimedia 25, 3546–3559.
Xu, Y., Shamsolmoali, P., Granger, E., Nicodeme, C., Gardes, L., Yang, J., 2023. TransVLAD: Multi-scale attention-based global descriptors for visual geo-localization. In: Proceedings of the IEEE International Winter Conference on Applications of Computer Vision (WACV), pp. 2839–2848.
Yan, L., Cui, Y., Chen, Y., Liu, D., 2021. Hierarchical attention fusion for geo-localization. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 2220–2224.
Yang, H., Lu, X., Zhu, Y., 2021a. Cross-view geo-localization with evolving transformer. arXiv:2107.00842.
Yang, J., Rong, H., Kang, Y., Zhang, F., Chegut, A., 2021b. The financial impact of street-level greenery on New York commercial buildings. Landsc. Urban Plan. 214, 104162.
Ye, J., Lv, Z., Li, W., Yu, J., Yang, H., Zhong, H., He, C., 2024. Cross-view image geo-localization with panorama-BEV co-retrieval network. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 74–90.
Yin, W., Xue, Y., Liu, Z., Li, H., Werner, M., 2025. LLM-enhanced disaster geolocalization using implicit geoinformation from multimodal data: a case study of Hurricane Harvey. Int. J. Appl. Earth Obs. Geoinformation 137, 104423.
Zamir, A.R., Shah, M., 2014. Image geo-localization based on multiple nearest neighbor feature matching using generalized graphs. IEEE Trans. Pattern Anal. Mach. Intell. 36 (8), 1546–1558.
Zhu, S., Yang, T., Chen, C., 2020. VIGOR: Cross-view image geo-localization beyond one-to-one retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5316–5325.
Zielstra, D., Hochmair, H.H., 2012. Using free and proprietary data to compare shortest-path lengths for effective pedestrian routing in street networks. Transp. Res. Rec. 2299 (1), 41–47.