Cross-view image geo-localization with Panorama-BEV Co-Retrieval Network

Junyan Ye[1,2*], Zhutao Lv[1], Weijia Li[1†], Jinhua Yu[1], Haote Yang[2], Huaping Zhong[3], and Conghui He[2,3†]

[1] Sun Yat-Sen University,[2] Shanghai AI Laboratory,[3] SenseTime Research
{yejy53,lvzht5}@mail2.sysu.edu.cn, liweij29@mail.sysu.edu.cn,
{yanghaote,heconghui}@pjlab.org.cn, zhonghuaping@sensetime.com

Abstract. Cross-view geolocalization identifies the geographic location of street view images by matching them with a georeferenced satellite database. Significant challenges arise due to the drastic appearance and geometry differences between views. In this paper, we propose a new approach for cross-view image geo-localization, i.e., the Panorama-BEV Co-Retrieval Network. Specifically, by utilizing the ground plane assumption and geometric relations, we convert street view panorama images into the BEV view, reducing the gap between street panoramas and satellite imagery. In the existing retrieval of street view panorama images and satellite images, we introduce BEV and satellite image retrieval branches for collaborative retrieval. By retaining the original street view retrieval branch, we overcome the limited perception range issue of BEV representation. Our network enables comprehensive perception of both the global layout and local details around the street view capture locations. Additionally, we introduce CVGlobal, a global cross-view dataset that is closer to real-world scenarios. This dataset adopts a more realistic setup, with street view directions not aligned with satellite images. CVGlobal also includes cross-regional, cross-temporal, and street view to map retrieval tests, enabling a comprehensive evaluation of algorithm performance. Our method excels in multiple tests on common cross-view datasets such as CVUSA, CVACT, VIGOR, and our newly introduced CVGlobal, surpassing the current state-of-the-art approaches. The code and datasets can be found at https://github.com/yejy53/EP-BEV.

Keywords: Remote sensing · Street view images · Geo-localization

1 Introduction

Cross-view retrieval geolocalization involves matching ground images with georeferenced satellite images in a database to identify their geographical locations [5,8,15,6,3,33,14,20], as shown in Fig. 1 (a). Cross-view retrieval faces challenges due to the significant differences between satellite and ground imagery perspectives. For example, buildings appear differently in satellite view from their

* This work was partially done during the internship at Shanghai AI Lab.
† Corresponding authors.

rooftops compared to ground perspectives of their facades. The morphological and textural characteristics of these viewpoints vary significantly. However, some elements, such as roads and crops are observable from both ground and satellite viewpoints, despite their visual differences, representing cross-view shared information [34]. The task focuses on harnessing cross-view information to effectively align content and distributions across both perspectives.

Current cross-view retrieval methods primarily leverage deep learning techniques with CNN [26,5,2,19,24,17] and Transformer [27,32,35] architectures to transform images from different perspectives into feature vectors . These vectors are then matched based on similarity calculations in the feature space. However, aligning the embedded feature vectors in the spatial domain remains challenging due to significant perspective differences. To mitigate this issue, some approaches use polar coordinate transformations to reduce geometric differences [15,22]. Specifically, instead of directly matching satellite view with street view images, satellite view images are first transformed into polar view images before being matched with street view images, as shown in Fig. 1 (b). The polar transformation effectively aligns cross-view shared information, such as road orientations, achieving notable performance improvements. However, transformed polar view images still exhibit significant differences in information distribution compared to ground images. For instance, ground images often include some sky information, while polar coordinate-transformed images contain treetop information, with considerable morphological distortion.

We observe that, in addition to converting satellite images to street view perspectives, it's also feasible to transform street view to satellite viewpoints. We transform street view panoramas into explicit Bird's Eye View (BEV) images using azimuth relationships and ground plane constraints. Compared to polar transformation, transforming street view into satellite viewpoints is more intuitive, resulting in transformed images that are more realistic and highlight cross-view shared local information near the shooting location. On the other hand, since our BEV transformation does not rely on depth and 3D structure estimation, images transformed via BEV can exhibit limited visibility and severe distortion in dense urban scenes where they are obstructed by tall structures like buildings. To tackle this challenge, we designed the Panorama-BEV Co-Retrieval Network, which collaboratively leverages street view panoramas and BEV images for satellite image retrieval. We retain the original street view panorama to satellite retrieval branch to expand the perception range and capture more global layout features, while the BEV to satellite retrieval branch focuses on the details near the street view locations.

Current cross-view retrieval research primarily utilizes datasets like CVUSA[26] , CVACT [8], and VIGOR[33], with CVUSA achieving top-1 recall rates over 98%, demonstrating the effectiveness of cross-view methods. However, a gap still exists between these datasets and real-world applications. Firstly, existing datasets mainly focus on a single country, limiting evaluations across diverse global scenes. Secondly, challenges with street views lacking metadata extend beyond unknown locations to include uncertain camera orientations and

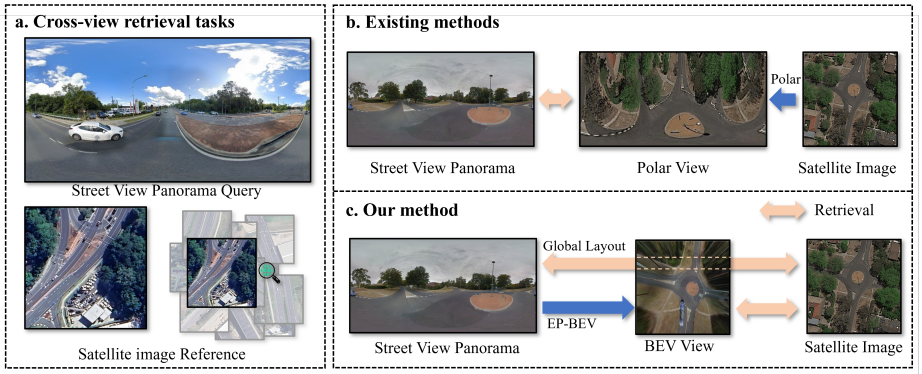

Fig. 1. The goal of cross-view retrieval is to identify the georeferenced satellite image most visually similar to the street view panorama query (a). Existing methods use polar transformation to convert satellite views into Polar views and then proceed with retrieval (b). Our method employs Explicit Panoramic BEV transformation to convert street view images into the BEV perspective consistent with satellite views, while preserving the previous street view panorama to satellite retrieval path (c).

shooting times. Currently, there are few datasets with street views that have uncertain orientations; most approaches simulate random captures by rotating fixed-orientation street view images [17,32]. Additionally, there is a lack of cross-temporal retrieval task evaluations, raising questions about whether current satellite imagery can accurately locate street views captured at unknown times. Furthermore, there have been few attempts to use map data instead of satellite data in the current datasets. Map data has advantages over satellite data, such as easier accessibility and storage. To address these challenges, we introduce a global cross-view retrieval dataset named CVGlobal. This dataset includes cross-regional, cross-temporal, and street view to map retrieval tests, aiming for a comprehensive assessment of algorithmic performance.

Our main contributions can be summarized as follows:

- We propose a novel transformation approach for cross-view retrieval tasks, explicitly converting street view panoramas into BEV views, effectively bridging the gap between street and satellite perspectives. By designing the Panorama-BEV Co-Retrieval Network, we facilitate collaborative satellite retrieval with street view panoramas and BEV images, surpassing BEV's perceptual limits to fully perceive global layouts and local details.
- We introduce CVGlobal, a global cross-view retrieval dataset that is closer to real-world application scenarios. The dataset features indeterminate street view orientations and supports evaluations of cross-regional, cross-temporal, and street view to map retrieval tasks.
- Our method has been extensively evaluated across multiple datasets and outperforms the current state-of-the-art approaches. In challenging cross-regional

tasks such as VIGOR-cross and from CVUSA to CVACT, our method improves the top-1 recall rate, demonstrating the generalization capability.

2 Related Work

2.1 Cross-view retrieval

Cross-view image retrieval methods use ground images as queries and all patches in a satellite image database as references for geolocation. Early retrieval efforts relied on manual features to match images across the two domains [1,22]. With the advent of deep learning algorithms, methods have evolved to embed images into global feature descriptors for retrieval [7,26,23,21,29]. Deuser et al. [3] employed the infoNCE loss combined with global hard negative mining, achieving state-of-the-art results. To mitigate the significant differences between satellite and ground images, many studies have improved retrieval accuracy through polar coordinate transformation algorithms [15,18,27,32]. Polar coordinate transformation, which relies on orientation relationships for direct conversion, introduces certain distortions when converting satellite images to street views. Toker et al. [22] leveraged GANs [4] to learn to eliminate these distortions.

Current algorithms perform well in top-5 and top-10 recall rates but struggle with low top-1 recall due to the challenge of distinguishing similar images in dense scenes when embedding street and satellite images directly. Our method enhances distinguishability by employing transformed BEV images for retrieval, incorporating more features near the shooting location.

2.2 BEV transformation

Transforming ground images into Bird's Eye View (BEV) representations is a key method for tasks such as autonomous driving and localization [11,12,10,28]. However, current methods based on BEV have a high demand for camera intrinsic and extrinsic parameters. OrienterNet [13] achieves precise localization with known approximate GPS positions by estimating camera parameters and scene depth for BEV feature mapping. Boosting [16] explores BEV feature-level projection based on geometric methods, yet converting tens of thousands of satellite images in a database to BEV feature representations, instead of efficient vector representations, remains a significant cost issue for retrieval tasks. Wang et al.[25] conducted explicit image transformations for cross-view localization tasks, achieving good results. However, this requires knowing the corresponding satellite image for a ground image, which is a subsequent task after retrieval.

Our Explicit Panoramic BEV Transformation utilizes geometric relationships and the ground plane assumption, starting from a predefined BEV plane to inversely calculate the panorama's indices, achieving explicit BEV transformation without the need for intrinsic or depth estimation. Unlike many localization approaches mapping street views to BEV feature representations with higher computational costs, our method converts BEV into an image representation, enabling direct feature embedding and efficient searching.

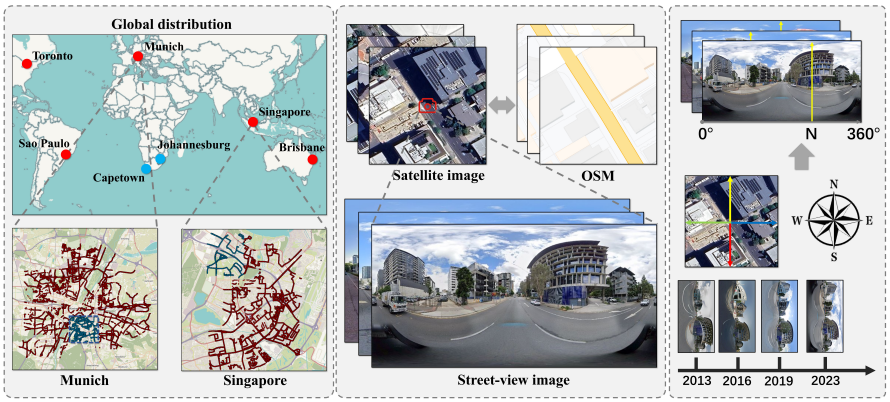

Fig. 2. The cross-view retrieval dataset CVGlobal encompasses data from various distinct style cities around the world, with red sample points representing training data and blue points indicating regional testing data (a). Since street views are captured by car-mounted cameras, they are usually centered on the road, and the north direction is not fixed (b). Additionally, CVGlobal introduces new tasks such as cross-temporal evaluation (c) and street view to map evaluation (d).

2.3 Cross-view Datasets

Several cross-view geolocation datasets have been introduced, including CVUSA [26], CVACT [8], Vo[23], Universities-1652 [31], and VIGOR [33]. The CVUSA dataset features 355,332 ground-to-satellite image pairs from the United States, while CVACT has a similar training/validation volume and a larger test set, CVACT-test. CVUSA and CVACT are the most commonly used cross-view retrieval datasets, employing a one-to-one retrieval setup. The VIGOR dataset includes data from multiple cities and it evaluates the model's transferability across different geographic regions. In this dataset, street panoramas and satellite images are not centrally aligned. Multiple street-view images cover the same satellite image area, with overlapping regions between different satellite images. Vo's research collected paired images from 11 U.S. cities, combining street views with Google Maps satellite photos. Universities-1652 extends the dataset by incorporating drone imagery in addition to street and satellite images.

Existing cross-view datasets provide comprehensive evaluations of cross-view retrieval algorithms across multiple dimensions and tasks. However, the existing datasets still fall short of real-world application scenarios. These include the need for data from more diverse cities worldwide, varying street view orientations, considering street view image retrieval at different times, or using map data instead of satellite images for retrieval. To tackle these issues, we introduce CVGlobal, a global cross-view retrieval dataset. It features street views with non-fixed orientations and includes cross-regional, cross-temporal, and street-to-map retrieval tests, aiming for a comprehensive evaluation of algorithm performance.

Table 1. Comparison of the proposed CVGlobal dataset with existing datasets in cross-view retrieval.

	Vo [23]	Uni.-1652[1] [31]	CVUSA [26]	CVACT [8]	VIGOR [33]	CVGlobal
Satellite images	∼ 450,000	50,218	44,416	128,334	90,618	134,233
Query images	∼ 450,000	41,135	44,416	128,334	105,214	134,233
Full panorama	✗	✗	✓	✓	✓	✓
Center aligned	✓	✗	✓	✓	✗	✓
Global scale	✗	✓	✗	✗	✗	✓
Unfixed Orientation	✓	✓	✗	✗	✗	✓
Multiple time	✗	✗	✗	✗	✗	✓
Map supplement	✗	✗	✗	✗	✗	✓

[1]Uni.-1652 refers to the Universities-1652 dataset.

3 Dataset

3.1 Dataset collection

We downloaded 134,233 street view images from seven cities globally in 2023 using Google Street View Download 360[1], including Munich, Toronto, Singapore, São Paulo, Brisbane, Cape Town, and Johannesburg, with an average distance of 50 m between images. Additionally, street views from Brisbane for 2013, 2016, and 2019 were collected to evaluate the algorithm's cross-temporal retrieval capabilities. Using the Google Maps Static API[2], we acquired corresponding satellite images and map data based on the latitude and longitude of the street views. The satellite images were at a size of 512×512, covering a spatial range of about 70m $\times$ 70$textm$. Map data and satellite imagery share the same coverage area and resolution.

3.2 Dataset comparison

Table 1 showcases a comparison between our dataset and previous benchmarks, illustrating that our dataset is closer to real-world scenarios with more potential application. Covering cities with a wide range of styles allows for an effective assessment of the algorithm's robustness across various scenarios. Additionally, the orientation of street views is not fixed. The dataset also includes street view data from Brisbane over multiple years, allowing for the evaluation of cross-temporal retrieval tasks. We use street view data from past years as queries and current satellite imagery as references. Retrieving images from different time periods represents a novel endeavor. Moreover, we've gathered map data slices aligned with satellite imagery, setting up street view to map slice retrieval tasks to probe their utility in cross-view retrieval. We use street view images as queries and rasterized map data slices as references. Compared to the high capture and storage costs of high-resolution satellite images, map data is easier to acquire and store. However, map data lacks the texture information present in satellite views, retaining only partial shape information. Particularly in underdeveloped

[1] https://svd360.istreetview.com/

[2] https://developers.google.com/maps/documentation/maps-static/

areas, where statistical data is sparse and updates are slow, map data contains very little useful information, posing challenges to the task.

3.3 Evaluation schemes

We selected street view satellite data from selected areas of Munich, Toronto, Singapore, São Paulo, and Brisbane from the year 2023 as our training set. Similar to CVUSA, we used data randomly divided from the same regions as the training set for our validation set. To address real-world application scenarios, we designed multiple experimental evaluation schemes:

Cross-regional retrieval. Our cross-regional tests are of two types: one within different areas of the training cities, as illustrated by the blue areas in Fig. 2, and the second using Cape Town and Johannesburg in Africa as test sets, increasing the task's difficulty. During testing, the satellite image database corresponding to the query images only includes the regional test set.

Cross-temporal retrieval. As mentioned earlier, our model includes Brisbane's training data from 2023, then tests its cross-temporal performance. We use street view images from Brisbane from the years 2013, 2016, and 2019 as queries, and satellite images from the corresponding locations in 2023 as the database to investigate whether the algorithm performs well across different years. We also combined data from these three years as input to investigate changes in the algorithm's performance.

Street view to map retrieval. We replaced the corresponding satellite images with map data for retraining and testing. We employed evaluations consistent with satellite imagery to explore the application potential of map data.

4 Methods

4.1 Overview

In cross-view retrieval task, the goal is to identify the most similar satellite imagery in a database based on the visual features of an input street view panorama query, thereby achieving geolocation of street view data. The primary challenge of this task lies in the significant perspective difference between street and satellite imagery. We address this by employing an explicit panoramic BEV transformation to bridge the gap between the two domains, highlighting cross-view information. Furthermore, to overcome the limited observation range of the BEV's fidelity mapping, we additionally utilize street view panorama branches to access a broader range of global observations.

As illustrated in Fig. 3, this paper introduces a novel cross-view retrieval method, Panorama-BEV Co-Retrieval Network. In the BEV branch, street view images are transformed into the satellite perspective through EP-BEV transformation for retrieval (see Section 4.2). Meanwhile, the street view panorama branch directly uses panoramas to search for satellite images. We achieve collaborative retrieval by simultaneously utilizing street view panoramas and BEV

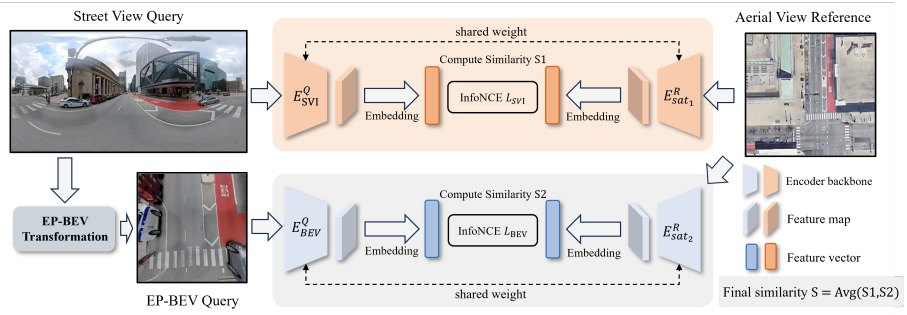

Fig. 3. Schematic of the Panorama-BEV Co-Retrieval Network. Street view images and their images transformed via EP-BEV serve as query inputs, with satellite images as reference inputs. The network comprises a street view branch focusing on matching global layout information with satellite images and a BEV branch emphasizing detailed feature matching between the nearby street view area and the satellite perspective.

images (see Section 4.3). We trained two models for the street view and BEV branches using the same contrastive image retrieval objective. During testing, the network will simultaneously apply both branches to retrieve the matching images for a given street view query, and the final decision will be made by combining the similarity scores from both branches.

4.2 Explicit panoramic BEV transformation

Traditional Bird's Eye View (BEV) transformation processes rely on accurate estimation of depth information and camera parameters. In contrast, our proposed method utilizes a geometric back-projection process based on the ground plane assumption, which directly calculates the corresponding positions of points on the BEV plane in the panorama (as shown in Fig. 4).

Given our objective to transform street view images into BEV views spatially aligned with satellite imagery, we first define a predetermined BEV plane aligned with the satellite perspective (as shown in Fig. 4 (a)), assuming the camera is located at the center of this BEV plane. Next, utilizing the grid relationship of the plane, i, j, we can determine the coordinates of the required mapping point $P(x, y, z = 0)$ (see Eq. 1). By setting the camera height to H and positioning the camera at $Cam(0, 0, H)$, we establish the spatial coordinate system (as illustrated in Fig. 4 (b)). Using geometric relationships, we can calculate the corresponding pitch angle θ and azimuth angle φ (see Eq. 2). With the equirectangular cylindrical projection characteristic of panoramic images, we can use θ and φ to calculate the respective row and column numbers v, u (see Eq. 3). By mapping the index relationship between i, j and v, u, we achieve the image transformation from the street view perspective to the BEV perspective. More details will be found in the supplementary material.

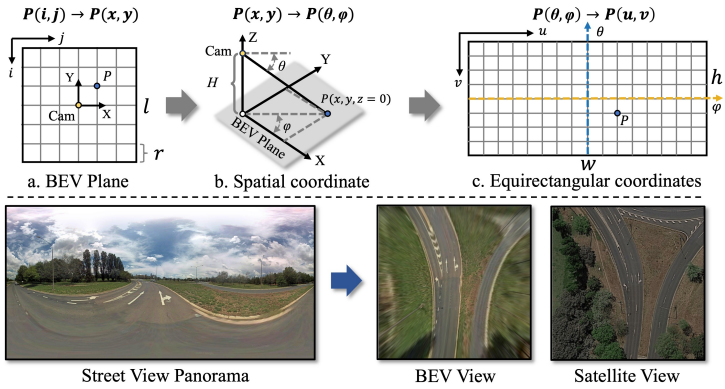

Fig. 4. Schematic of the Explicit Panoramic BEV Transformation, with the upper part showing specific transformation details and the lower part displaying the results of the BEV conversion.

$$x = \left(j - \frac{l}{2}\right) \times r, \qquad y = \left(\frac{l}{2} - i\right) \times r \tag{1}$$

$$\theta = -\arctan\left(\frac{H}{\sqrt{x^2 + y^2}}\right), \qquad \varphi = \arctan 2(y, x) \tag{2}$$

$$v = \left(\frac{\pi}{2} - \theta\right) \times \frac{h}{\pi}, \qquad u = \left(\frac{\varphi + \pi}{2\pi}\right) \times w \tag{3}$$

In the formulas, y and x represent the coordinates in three-dimensional space, which are calculated through the row number i and column number j on the BEV plane, the edge length of the BEV plane is l, and its resolution is r. The camera height is set to H, the pitch angle θ refers to the angle between the line connecting the camera to point P and the camera plane and the azimuth angle φ is the angle with respect to the positive direction of the x-axis. v and u are the row and column numbers of the panoramic image, respectively, while h and w are the height and width of the panoramic image.

Through explicit panoramic BEV transformation, we project street view images into a bird's-eye view without needing depth estimation or camera parameters. Although transformed images lose information on tall objects, like building facades, this information is unique to the ground perspective and not visible from the satellite view. Explicit panoramic BEV transformation effectively minimizes the correspondence gap between the two domains, highlighting information on objects observable in both satellite and street views.

4.3 Dual-branch cross-view image retrieval

Our method employs a dual-branch structure to accomplish the collaborative retrieval (co-retrieval) task of street view panoramas and BEV. For the street

view retrieval branch, we embed street and satellite images as feature vectors through encoders and optimize using InfoNCE loss L_{Pan}. This branch directly utilizes the original street view inputs, covering a wider observation range and aligning with satellite images in a more global layout distribution. In the BEV retrieval branch, following the transformation described in Section 4.2, we first convert street views into BEV views to perform detailed feature alignment under the satellite perspective, optimizing with InfoNCE loss L_{BEV}. This branch, using the transformed EP-BEV inputs, emphasizes cross-view information near the street view, aiding in distinguishing between similar satellite images. Since the optimization directions of the different branches are not the same, the two branches are trained separately to obtain the final models.

$$L = -\log\left(\frac{\exp(Q \cdot R^{+}/\tau)}{\sum_{j=1}^{N}\exp(Q \cdot R_j/\tau)}\right) \tag{4}$$

In this formula, Q represents the query images, including the street view panorama query image Q_{Pan} and the BEV query image Q_{BEV}. R^{+} represent the positive reference images that are geographically consistent with the query images. R_j denote the negative samples and τ is a temperature parameter. During inference, the street view branch computes similarity S1, and the EP-BEV branch computes similarity S2. The sum of S1 and S2 determines the final retrieval result. To save memory, only the top-ranked results from the street view branch need to be used for similarity merging.

5 Experiments

5.1 Dataset and Evaluation Protocol

Following common experimental setting [15,17,32,35,3,27], we conducted extensive experimental evaluations on three widely used datasets: CVUSA [26], CVACT [8], VIGOR [33] and our proposed CVGlobal dataset, to validate the effectiveness of our model. We utilized the metric of top-k image recall rate to assess model performance. Specifically, given a street view panorama query, if the corresponding closest satellite image is within the top k retrieved images, then the query is considered "successfully retrieved." The percentage of query images that have been correctly localized is referred to as R@K.

5.2 Implementation Details

We employ the ConvNeXt-B [9] as the backbone network for encoding both ground and satellite images in the two retrieval stages, utilizing the AdamW optimizer with a learning rate set to 1.0×10^{-3}. The temperature parameter τ is a learnable parameter [3]. Our training spans 40 epochs, with a batch size of 128. Following the setup by Yu et al. [16], we set the camera shooting height to 1.5 m, and the BEV plane range is aligned with the satellite imagery range of CVACT, with the image size l set to 512×512 and a pixel resolution r of 14cm.

Table 2. Quantitative comparison between our approach and state-of-the-art approaches on CVUSA and CVACT. † denotes methods that use polar transformation.

Methods	CVUSA				CVACT Val				CVACT Test			
	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
SAFA†[15]	89.84	96.93	98.14	99.64	81.03	92.80	94.84	-	-	-	-	-
LPN[24]	85.79	95.38	96.98	99.41	79.99	90.63	92.56	-	-	-	-	-
LPN†[24]	92.83	98.00	98.85	99.78	83.66	94.14	95.92	98.41	-	-	-	-
DSM[17]	91.96	97.50	98.54	99.67	82.49	92.44	93.99	97.32	-	-	-	-
TransGeo [32]	94.08	98.36	99.04	99.77	84.95	94.14	95.78	98.37	-	-	-	-
GeoDTR[30]	93.76	98.47	99.22	99.85	85.43	94.81	96.11	98.26	62.96	87.35	90.70	98.61
GeoDTR†[30]	95.43	98.86	99.34	99.86	86.21	95.44	96.72	98.77	64.52	88.59	91.96	98.74
SAIG-D[35]	96.08	98.72	99.22	99.86	89.21	96.07	97.04	98.74	67.49	89.39	92.30	96.80
Samp4G[3]	98.68	99.68	**99.78**	**99.87**	90.81	96.74	97.48	98.77	71.51	92.42	94.45	98.70
Ours	**98.71**	**99.70**	**99.78**	99.86	**91.90**	**97.23**	**97.84**	**98.84**	**73.68**	**93.53**	**95.11**	**98.81**

Table 3. Quantitative comparison between our approach and the current state-of-the-art on VIGOR. † denotes methods that use polar transformation on the satellite input image.

Methods	Same-area				Cross-area			
	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
SAFA†[15]	33.93	58.42	68.12	98.24	8.20	19.59	26.36	77.61
TransGeo[32]	61.48	87.54	91.88	99.56	18.99	38.24	46.91	88.94
SAIG-D[35]	65.23	88.08	-	99.68	33.05	55.94	-	94.64
Samp4G[3]	77.86	95.66	97.21	99.61	61.70	83.50	88.00	98.17
Ours	**82.18**	**97.10**	**98.17**	**99.70**	**72.19**	**88.68**	**91.68**	**98.56**

Since CVUSA's panoramic images are not of regular size, ground images need to be padded to standard dimensions. All comparison methods were implemented according to their publicly available source code settings.

5.3 Evaluation results on existing datasets

Cross-view Image retrieval. Our method performed optimally on the classic CVUSA and CVACT datasets. On the more challenging VIGOR dataset, our method increased the top-1 recall by 4.32% in the Same-area task and by 10.49% in the Cross-area task, indicating our method's effectiveness for difficult tasks. As visualized in Fig. 5 , we observe a large number of similar images within the VIGOR dataset. This phenomenon allows other methods to perform well in terms of top-10 or top-5 recall rates but show lower performance in top-1 recall. In contrast, the newly added BEV branch in our method substantially enhances the ability to distinguish details in each image and emphasizes cross-view information near the shooting location, thereby significantly improving the capability to differentiate between similar images.

Generalisation Capabilities. Following the experimental setups in Samp4G [3] and L2LTR [27], we evaluated the generalization capability of the algorithm, as shown in Table 4. Compared to the settings of VIGOR-cross, CVUSA and CVACT present more challenging tasks due to significant differences in resolution, satellite image size, and the scope of street view imaging. As shown in Table

Table 4. Cross-dataset generalization capability test. Models are trained on the CVUSA training splits and tested on the CVACT validation and test splits. † denotes approaches that used the polar transformation. Only SVI denotes retrieval exclusively with the street view branch, Only BEV indicates retrieval solely via the BEV branch.

Methods	CVUSA → CVACT-Val				CVUSA → CVACT-Test			
	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
L2LTR[27]	47.55	70.58	-	91.39	-	-	-	-
L2LTR†[27]	52.58	75.81	77.39	93.51	-	-	-	-
GeoDTR[30]	47.79	70.52	-	92.20	11.24	18.69	23.67	72.09
GeoDTR†[30]	53.16	75.62	81.90	93.80	22.09	32.22	39.59	85.53
Samp4G[3]	56.62	77.79	87.02	94.69	27.78	52.08	60.33	94.88
Ours (Only SVI)	54.17	76.55	86.99	94.23	26.11	50.18	60.34	94.65
Ours (Only BEV)	61.92	81.33	85.81	93.88	33.51	61.25	68.97	94.18
Ours	**67.79**	**84.06**	**87.96**	**95.05**	**44.10**	**70.68**	**75.86**	**95.31**

Table 5. Ablation studies for the image retrieval on the VIGOR dataset. Only SVI denotes retrieval exclusively with the street view branch, Only BEV indicates retrieval solely via the BEV branch, and Ours refers to the combination of SVI and BEV, representing our Panorama-BEV Co-Retrieval Network.

Methods	Same-area				Cross-area			
	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
Only SVI	76.39	94.76	96.66	99.58	58.93	81.28	86.21	98.01
Only BEV	65.47	87.26	90.84	98.44	44.01	64.87	71.29	93.12
Ours	**82.18**	**97.10**	**98.17**	**99.70**	**72.19**	**88.68**	**91.68**	**98.56**

4, our method scores highly on CVACT, outperforming the current state-of-the-art method by over 10%, demonstrating strong generalizability. Both L2LTR[27] and GeoDTR[30] saw significant generalization improvements with polar transformation. Our method, using only transformed BEV images, also significantly outperformed street-view-only approaches. This is attributed to the embedding of cross-view information like roads through polar and EP-BEV transformations. Direct street-to-satellite retrieval, however, is more affected by variations in observation range and resolution across datasets.

Ablation experiment. We conducted ablation experiments on the VIGOR dataset to explore the effectiveness of our algorithm, as shown in Table 5. We compared the results of retrieval using each of the two branches independently with the outcomes of our collaborative retrieval that simultaneously employs both branches. The results show that using both street view and BEV queries simultaneously significantly improves recall compared to previously using only street view. This also indicates that using only the BEV branch is not ideal due to the limited observation range compared to satellite images, resulting in a significant loss of global layout information.

5.4 Evaluation results on the proposed CVGlobal dataset

Cross-region image retrieval. The experimental results from Table 6 show that the Regional-val, being a randomly divided validation set, provides extensive

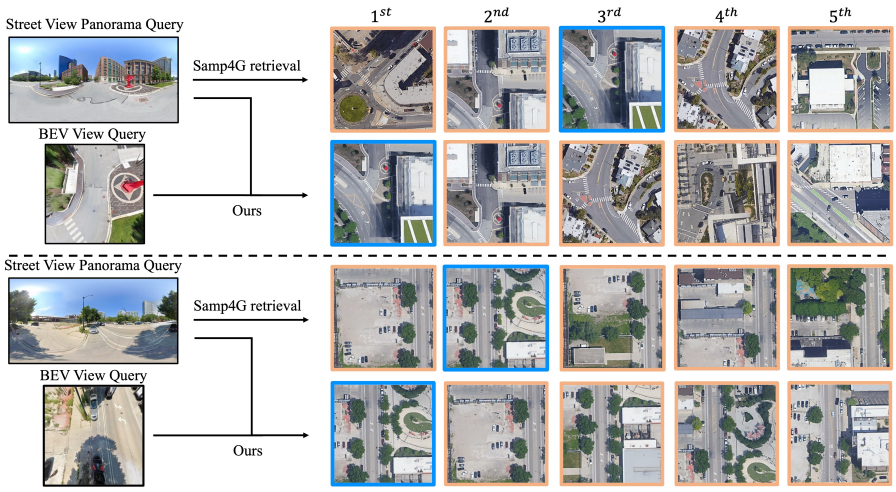

Fig. 5. On the VIGOR dataset, we compare our method with Samp4G's retrieval results, using blue and orange boxes to represent correct and incorrect retrievals, respectively.

Table 6. Quantitative comparison between our approach and the current state-of-the-art on CVGlobal Cross-regional image retrieval.

Methods	Regional val				Regional test				Cross-continent test			
	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
GeoDTR[30]	47.42	68.43	78.11	99.05	25.43	42.07	51.66	84.74	4.58	9.32	13.15	52.53
SAIG-D[35]	71.92	92.83	96.05	99.81	34.72	61.53	71.08	91.47	8.66	20.50	27.28	69.96
Samp4G[3]	97.20	99.43	99.69	**99.93**	84.66	92.66	94.42	98.10	51.21	70.95	76.02	92.58
Ours	**97.78**	**99.63**	**99.79**	**99.93**	**85.65**	**92.91**	**94.77**	**98.21**	**60.25**	**75.01**	**79.51**	**93.51**

Table 7. Quantitative comparison between our approach and the current state-of-the-art on CVGlobal Cross-temporal image retrieval.

Methods	2013			2016			2019			Mixing		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
GeoDTR[30]	10.05	19.33	25.93	12.10	22.74	30.46	13.76	24.94	32.82	7.10	13.81	19.22
SAIG-D[35]	13.72	30.83	40.46	18.77	39.23	49.31	22.52	46.26	56.56	12.52	28.59	36.92
Samp4G[3]	68.59	84.68	88.35	82.86	93.11	94.88	86.61	96.77	97.87	73.33	88.56	91.30
Ours	**74.52**	**89.02**	**91.89**	**86.44**	**94.96**	**96.35**	**88.74**	**97.56**	**98.45**	**78.18**	**91.50**	**93.52**

coverage of the training area, resulting in very ideal results. Meanwhile, the cross-regional evaluation performed on Regional-test, where the city style remains unchanged, demonstrates good algorithm performance. However, in the Cross-continent test, which faces scenes with significant style differences, performance noticeably declines. Compared to existing methods, our approach shows a clear improvement in performance, particularly standing out in challenging scenarios.

Table 8. Quantitative comparison between our approach and the current state-of-the-art on CVGlobal street view to map retrieval.

Methods	Regional map val				Regional map test			
	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
GeoDTR[30]	11.95	24.33	34.07	89.32	4.57	10.47	16.07	53.45
SAIG-D[35]	37.27	70.61	80.01	97.53	4.64	14.80	22.16	61.55
Samp4G[3]	75.38	92.80	95.67	99.54	46.94	70.46	77.91	91.47
Ours	**81.31**	**95.84**	**97.68**	**99.71**	**49.41**	**71.04**	**78.33**	**91.96**

Cross-temporal image retrieval. As shown in Table 7, we evaluate the model trained on 2023 data for cross-view assessment. We use ground data from 2013, 2016, and 2019, as well as a mix of these three years as queries, with 2023 satellite images as references. Our algorithm demonstrates superior performance in cross-temporal retrieval tasks. The model needs to capture long-term persistent features (e.g., road layouts and building structures) and reduce reliance on transient features (e.g., temporary buildings and vehicles) for this cross-temporal task. We also observe that the closer the data year is to 2023, the more ideal the evaluation results. Meanwhile, mixing data from different years introduces more disturbances to the results, but this scenario is closer to real-world application settings and represents a challenge that needs to be overcome.

Street view to map retrieval. We conducted training and cross-regional evaluations by replacing satellite images with map slice data, with results shown in Table 8. Due to the lack of specific appearance and texture information in map data, the performance of various algorithms was not optimal. However, in the validation set, because of the high quality of map data used for both training and testing and the dense coverage of training data, the top-1 recall using map tiles exceeds 70%, proving the effectiveness of applying map data in cross-view geolocalization tasks.

6 Conclusion

In our work, we use the ground plane assumption and geometric relationships to convert street view panorama images to the satellite perspective, effectively reducing the domain gap between them. Our Panorama-BEV Co-Retrieval Network captures global information and enhances local detail features simultaneously. Our method achieves outstanding results on existing datasets. In the VIGOR-cross testing task, compared to the state-of-the-art methods, our top-1 recall rate increased by 10.49%; in the CVUSA to CVACT task, it increased by 13.75%. At the same time, we introduce a dataset that is closer to real-world scenarios as a benchmark. We also provide various evaluation modes to explore the performance of our method in cross-regional, cross-temporal, and map data retrieval tasks. Our proposed dataset offers a new testing platform for cross-view geographic localization, fostering new research in the field.

Acknowledgements. This project was funded in part by National Natural Science Foundation of China (Grant No. 42201358) and Shanghai AI Lab.

References

1. Bansal, M., Sawhney, H.S., Cheng, H., Daniilidis, K.: Geo-localization of street views with aerial image databases. In: Proceedings of the 19th ACM international conference on Multimedia. pp. 1125–1128 (2011)
2. Cai, S., Guo, Y., Khan, S., Hu, J., Wen, G.: Ground-to-aerial image geo-localization with a hard exemplar reweighting triplet loss. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8391–8400 (2019)
3. Deuser, F., Habel, K., Oswald, N.: Sample4geo: Hard negative sampling for cross-view geo-localisation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 16847–16856 (October 2023)
4. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
5. Hu, S., Feng, M., Nguyen, R.M., Lee, G.H.: 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. pp. 7258–7267 (2018)
6. Hu, S., Lee, G.H.: Image-based geo-localization using satellite imagery. International Journal of Computer Vision **128**(5), 1205–1219 (2020)
7. Lin, T.Y., Belongie, S., Hays, J.: Cross-view image geolocalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 891–898 (2013)
8. Liu, L., Li, H.: Lending orientation to neural networks for cross-view geo-localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5624–5633 (2019)
9. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
10. Pan, B., Sun, J., Leung, H.Y.T., Andonian, A., Zhou, B.: Cross-view semantic segmentation for sensing surroundings. IEEE Robotics and Automation Letters **5**(3), 4867–4873 (2020)
11. Peng, L., Chen, Z., Fu, Z., Liang, P., Cheng, E.: Bevsegformer: Bird's eye view semantic segmentation from arbitrary camera rigs. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5935–5943 (2023)
12. Reiher, L., Lampe, B., Eckstein, L.: A sim2real deep learning approach for the transformation of images from multiple vehicle-mounted cameras to a semantically segmented image in bird's eye view. In: 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC). pp. 1–7. IEEE (2020)
13. Sarlin, P.E., DeTone, D., Yang, T.Y., Avetisyan, A., Straub, J., Malisiewicz, T., Bulò, S.R., Newcombe, R., Kontschieder, P., Balntas, V.: Orienternet: Visual localization in 2d public maps with neural matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21632–21642 (2023)
14. Sarlin, P.E., Trulls, E., Pollefeys, M., Hosang, J., Lynen, S.: Snap: Self-supervised neural maps for visual positioning and semantic understanding. Advances in Neural Information Processing Systems **36** (2024)
15. Shi, Y., Liu, L., Yu, X., Li, H.: Spatial-aware feature aggregation for image based cross-view geo-localization. Advances in Neural Information Processing Systems **32** (2019)

16. Shi, Y., Wu, F., Perincherry, A., Vora, A., Li, H.: Boosting 3-dof ground-to-satellite camera localization accuracy via geometry-guided cross-view transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21516–21526 (2023)
17. Shi, Y., Yu, X., Campbell, D., Li, H.: Where am i looking at? joint location and orientation estimation by cross-view matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4064–4072 (2020)
18. Shi, Y., Yu, X., Liu, L., Campbell, D., Koniusz, P., Li, H.: Accurate 3-dof camera geo-localization via ground-to-satellite image matching. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(3), 2682–2697 (2022)
19. Shi, Y., Yu, X., Liu, L., Zhang, T., Li, H.: Optimal feature transport for cross-view image geo-localization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11990–11997 (2020)
20. Thoma, J., Paudel, D.P., Chhatkuli, A., Probst, T., Gool, L.V.: Mapping, localization and path planning for image-based navigation using visual features and map. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7383–7391 (2019)
21. Tian, Y., Chen, C., Shah, M.: Cross-view image matching for geo-localization in urban environments. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3608–3616 (2017)
22. Toker, A., Zhou, Q., Maximov, M., Leal-Taixé, L.: Coming down to earth: Satellite-to-street view synthesis for geo-localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6488–6497 (2021)
23. Vo, N.N., Hays, J.: Localizing and orienting street views using overhead imagery. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. pp. 494–509. Springer (2016)
24. Wang, T., Zheng, Z., Yan, C., Zhang, J., Sun, Y., Zheng, B., Yang, Y.: Each part matters: Local patterns facilitate cross-view geo-localization. IEEE Transactions on Circuits and Systems for Video Technology **32**(2), 867–879 (2021)
25. Wang, X., Xu, R., Cui, Z., Wan, Z., Zhang, Y.: Fine-grained cross-view geo-localization using a correlation-aware homography estimator. Advances in Neural Information Processing Systems **36** (2024)
26. Workman, S., Souvenir, R., Jacobs, N.: Wide-area image geolocalization with aerial reference imagery. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3961–3969 (2015)
27. Yang, H., Lu, X., Zhu, Y.: Cross-view geo-localization with layer-to-layer transformer. Advances in Neural Information Processing Systems **34**, 29009–29020 (2021)
28. Ye, J., Luo, Q., Yu, J., Zhong, H., Zheng, Z., He, C., Li, W.: Sg-bev: Satellite-guided bev fusion for cross-view semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27748–27757 (2024)
29. Zhai, M., Bessinger, Z., Workman, S., Jacobs, N.: Predicting ground-level scene layout from aerial imagery. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 867–875 (2017)
30. Zhang, X., Li, X., Sultani, W., Zhou, Y., Wshah, S.: Cross-view geo-localization via learning disentangled geometric layout correspondence. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 3480–3488 (2023)
31. Zheng, Z., Wei, Y., Yang, Y.: University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In: Proceedings of the 28th ACM international conference on Multimedia. pp. 1395–1403 (2020)

32. Zhu, S., Shah, M., Chen, C.: Transgeo: Transformer is all you need for cross-view image geo-localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1162–1171 (2022)
33. Zhu, S., Yang, T., Chen, C.: Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3640–3649 (2021)
34. Zhu, Y., Chen, S., Lu, X., Chen, J.: Cross-view image synthesis from a single image with progressive parallel gan. IEEE Transactions on Geoscience and Remote Sensing (2023)
35. Zhu, Y., Yang, H., Lu, Y., Huang, Q.: Simple, effective and general: A new backbone for cross-view image geo-localization. arXiv preprint arXiv:2302.01572 (2023)