Neural Networks 203 (2026) 109099

Contents lists available at ScienceDirect

Neural Networks

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

Full Length Article

Window-to-window BEV representation learning for limited FoV cross-view geo-localization

Lei Cheng, Daikun Liu, Lingquan Meng, Teng Wang *, Changyin Sun

School of Automation, Southeast University, Nanjing, 210096, Jiangsu, China

ARTICLE INFO

Keywords:
Cross-view geo-localization
Limited field of view
Bird's eye view

ABSTRACT

Cross-view geo-localization confronts significant challenges due to large perspective changes, especially when the ground-view query image has a limited field of view (FoV) with unknown orientation. To bridge the cross-view domain gap, we explore learning the bird's eye view (BEV) representation directly from the ground feature. However, the unknown orientation of ground images combined with the absence of camera parameters leads to ambiguity between BEV queries and ground references. To tackle this challenge, we propose a novel **W**indow-**t**o-**W**indow **BEV** representation learning framework, termed **W2W-BEV**, which adaptively matches BEV queries to ground reference at window-scale and generates the BEV feature for each grid by attending to the semantically relevant ground features within the matched window. Additionally, we employ depth-guide BEV initialization to project the ground feature into BEV space, thus facilitating the high-quality window matching. By performing window-scale matching in a learnable manner, our W2W-BEV allows the BEV space to learn features by borrowing information from limited FoV, and aligning it with satellite imagery. Extensive experimental results on benchmark datasets demonstrate significant superiority of our W2W-BEV over previous state-of-the-art methods under challenging conditions of unknown orientation and limited FoV.

1. Introduction

Cross view geo-localization refers to the process of matching a query ground-level image against a database of reference satellite images annotated with accurate GPS labels to determine the geographic location of the ground-level image. It holds great prospect in robotic navigation (Shetty & Gao, 2019) and autonomous driving (Li et al., 2019). This task is usually cast as an image retrieval problem, with the aim of learning viewpoint-invariant feature embeddings. However, the great domain gap between ground and aerial views makes the above representation learning extremely challenging.

To bridge the domain gap, a body of works (Shen et al., 2024; Shi et al., 2019, 2020b; Zhang et al., 2023; Zhu et al., 2023b) utilize deep metric learning to implicitly align cross-view features extracted from dual-branch convolutional neural networks (CNNs). Some other methods (Fervers et al., 2023a; Regmi & Shah, 2019; Shi et al., 2019) explicitly align different views by transforming images from one view into another view in pixel (Shi et al., 2020a, 2022) or feature (Shi & Li, 2022; Shi et al., 2023) space. For instance, polar transform (Shi et al., 2019) and projection transform (Shi et al., 2022) are developed to map aerial images to ground-level panoramas, whereas cross-view feature transformation is usually achieved by predicting the rotation and translation matrix (Shi & Li, 2022; Shi et al., 2023). Significant performance improvement is observed by incorporating these geometric transformations into CNN-based frameworks. However, these geometric transformations easily cause image distortion, leading to the loss of valuable geometric information. Furthermore, all of these works assume that the ground-level query image is a panorama and is perfectly aligned with the reference aerial image. Nevertheless, in practical scenarios, the ground-level images typically exhibit a limited Field of View (FoV) with unknown orientations, which incurs obvious performance degradation.

More recent studies (Shi et al., 2020a, 2022; Wang et al., 2024; Yang et al., 2021; Zhai et al., 2017; Zhu et al., 2022, 2023b) have thus shifted their focus towards these more practical scenarios. To cope with the exacerbated domain disparity, Shi et al. (2020a) propose a Dynamic Similarity Matching (DSM) model to establish the patch-level correspondences between cross-view features, so that cross-view features could be matched in a more accurate manner in presence of orientation misalignment and limited FoV; L2LTR (Yang et al., 2021) and TransGeo (Zhu et al., 2022) resort to Vision Transformer (ViT) (Dosovitskiy et al., 2020) to learn robust global representations; DeHi (Wang et al., 2024) combines CNN and ViT to learn a multi-scale feature representation ro-

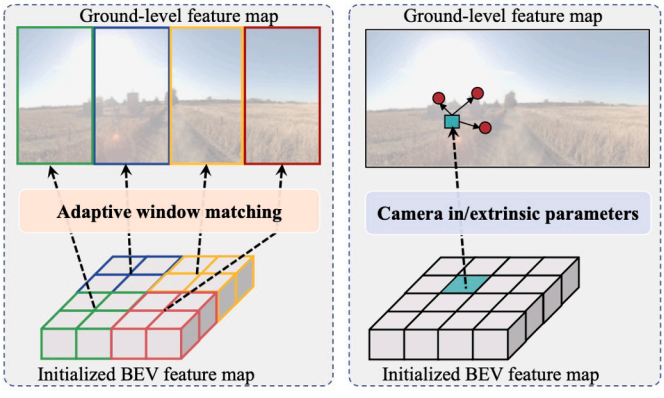
* Corresponding author.
E-mail addresses: leicheng@seu.edu.cn (L. Cheng), dkliu@seu.edu.cn (D. Liu), lingquanmeng@seu.edu.cn (L. Meng), wangteng@seu.edu.cn (T. Wang), cysun@seu.edu.cn (C. Sun).

https://doi.org/10.1016/j.neunet.2026.109099
Received 26 August 2025; Received in revised form 27 March 2026; Accepted 10 May 2026
Available online 14 May 2026

bust to orientation misalignment. Despite some improvements, the performance of these methods is still far from satisfactory when the FoV is reduced down to 90° or smaller, which is usually featured by standard cameras mounted on autonomous driving systems.

In this work, we attempt to leverage bird's-eye-view (BEV) representation learning (Li et al., 2023, 2022) to enhance cross-view geo-localization, especially under limited FoV. Transforming ground-level image features into BEV space is attractive, since it could explicitly align cross-view images in feature space. However, in practical scenarios where the orientation of ground-level image is unknown and camera parameters are missing, how to determine the point correspondences between the ground-level and BEV feature maps is challenging. One feasible solution might be to establish point-to-point correspondences in a learnable manner, which suffers from heavy computation. To this end, we develop a context-aware window matching strategy to determine the patch-level correspondences between BEV and ground-level feature maps in a learnable manner, as shown in Fig. 1(a). Thereby, each grid in a BEV window could generate its feature by attending to those semantically relevant ground features within the matched window. By matching and learning at window scale, we could improve matching efficiency while preserving the semantic integrity of the reference ground features. Furthermore, to facilitate high-quality window matching, we introduce a depth-guided BEV initialization method, which leverages the predicted depth to transform the ground-level features into BEV space. Extensive experimental results demonstrate that our proposed W2W-BEV outperforms the state-of-the-art (SOTA) by a large margin under limited FoV and unknown orientation. To summarize, our main contributions are as follows:

- We propose a novel window-to-window BEV representation learning framework to enhance cross-view geo-localization under limited FoV and unknown orientation. The BEV representations learned from ground-level image features could effectively capture the underlying spatial structures similar to that of aerial images even under limited FoVs, thereby reducing cross-view disparities.
- We develop a context-aware window matching strategy to establish the local correspondence between BEV and ground-level feature maps in a learnable manner, allowing BEV space to leverage relevant ground-level semantics within the limited FoV for reasoning and imagination. When combined with the proposed BEV initialization strategy, it could effectively facilitate the learning of BEV representation under limited FoV.



(a) Window-to-window BEV learning (b) Point-to-point BEV learning

Fig. 1. Regular methods utilize camera parameters to find the **locationally corresponding** point (green square) for each query point in BEV space and learn BEV features using deformable attention. In contrast, our method localizes the most **semantically relevant** ground window for each BEV window in a fully learnable manner. Here, FoV = 180°.

- Extensive experimental results on benchmark datasets show that our method achieve significant improvements under unknown orientation and limited FoV.

2. Related work

2.1. Feature extractor

Earlier works in cross-view geo-localization primarily employ a dual-branch CNN framework to extract features from ground-level and aerial images, respectively (Hu et al., 2018; Lin et al., 2015; Vo & Hays, 2016; Workman et al., 2015). In subsequent work, several methods for extracting global features have been introduced. SAFA (Shi et al., 2019) proposes a spatial-aware feature aggregation module. MCCG (Shen et al., 2024) introduces cross-dimensional feature interaction to enhance feature diversity. L2LTR (Yang et al., 2021) and TransGeo (Zhu et al., 2022) opt for transformer-based architectures as backbones to extract powerful global features. Dehi (Wang et al., 2024), leveraging the superior performance of ViT in cross-view geo-localization, proposes a hierarchical feature extraction framework that improves model capability through multi-level feature learning. Meanwhile, other researchers have focused on local-context feature extraction, particularly in Drone-Satellite cross-view geo-localization (Zheng et al., 2020; Zhu et al., 2023a). LPN (Wang et al., 2021) partitions feature maps into square-ring regions to extract both target and contextual information. However, its center-alignment assumption limits its applicability in more realistic scenarios (e.g., VIGOR Zhu et al., 2021). FSRA (Dai et al., 2021) employs heatmaps to segment features and achieve region alignment, which works well in Drone-Satellite scenarios(where shared information is abundant) but struggles in Ground-Satellite cross-view geo-localization, where viewpoint differences are more pronounced and shared features are scarce. To address domain-alignment challenges, DAC (Xia et al., 2024) performs cross-batch scene alignment, and CCR (Du et al., 2024) employs a counterfactual causal reasoning-based learning approach to extract effective alignment features. These methods are either unsuitable for ground-satellite cross-view geo-localization or rely on the assumption that street-view images are panoramic, making them difficult to generalize to real-world scenarios with unknown orientations and limited FoVs. DSM (Shi et al., 2020a) dynamically infers relative orientations through similarity matching to align cross-view features. ConGeo (Mi et al., 2024) leverages a contrastive learning strategy to capture the similarities and differences between panoramic images with known orientations and street-view images with limited FoV and unknown orientations, demonstrating excellent robustness to variations in both FoV and orientation. *These methods implicitly achieve cross-view alignment by learning global, viewpoint-invariant feature embeddings through deep metric learning. In contrast, our W2W-BEV adopts a fundamentally different strategy: we learn an Bird's-Eye View (BEV) representation from ground features, thereby aligning cross-view features in a clear and defined manner.*

2.2. Viewpoint transformation

Several studies (Lu et al., 2020; Regmi & Shah, 2019; Shi & Li, 2022; Shi et al., 2019, 2023, 2020a) focus on aligning cross-view perspectives to reduce the gap between different viewpoints. SAFA (Shi et al., 2019) proposes leveraging polar coordinate transformations to geometrically convert satellite images into street-view-like representations, which subsequent works (Shi et al., 2020a; Wang et al., 2024) have validated as effective for cross-view image geo-localization. However, it is constrained by the underlying assumption of orientation alignment. Works (Shi & Li, 2022; Shi et al., 2023) further employ neural networks to predict unknown rotation matrices and translation vectors, projecting satellite image features into ground-view feature spaces to mitigate cross-view discrepancies. On the other hand, Generative Adversarial Network-based approaches (Lu et al., 2020;

Regmi & Shah, 2019) synthesize satellite images from ground-view images and utilize these generated images for satellite image retrieval. While these methods demonstrate efficacy, they inevitably introduce image distortions or artifacts that interfere with robust feature extraction. *Different from these works, we learn BEV representation by borrowing related information from ground-level features via window-based cross-attention. The learned BEV feature has the capability to imagine the unseen contents.*

2.3. BEV transformation

BEV representation learning can effectively project front-view images into a bird's-eye view perspective. BEVFormer (Li et al., 2022) and LLS (Philion & Fidler, 2020) established exemplars for BEV representation learning, while subsequent methods like BEVFix (Li et al., 2025) further leverage 3D point cloud distributions to refine the BEV representations after view transformation. PolarFusion (Shi et al., 2025) replaces the spatial correspondences of Cartesian coordinates with those of polar coordinates, significantly enhancing the accuracy of BEV algorithms in object detection tasks. Recent studies start to introduce BEV into cross-view geo-locaziation, aims to reduce the cross-view domain gap. The work (Fervers et al., 2023b) utilizes camera in/extrinsic parameters to determine the reference points on ground-level feature maps for each BEV query grid. In the absence of camera parameters, C-BEV (Fervers et al., 2023a) employs column-wise attention to transform the panorama image features to a polar BEV representation, which is then projected to cartesian coordinates through bilinear resampling to generate BEV representation. Despite impressive results on street-view panoramas, C-BEV fails to handle limited FoV due to the use of column-wise cross attention for cross-view projection. The spherical transformation (Wang et al., 2024) is utilized for ground panorama images, projecting them onto the BEV perspective through geometric transformations, resulting in image distortion and significant information loss. *Crucially, unlike C-BEV and spherical transformation, which is limited in its ability to handle limited FoV, we develop a context-aware window matching strategy to establish semantic correspondence between the BEV and image feature space in a learnable manner. This window-scale matching allows each BEV grid to generate its features by attending to the most semantically relevant ground features. Therefore, our method can robustly handle various challenging conditions, including missing camera parameters, unknown orientation, and limited FoV.*

3. Methodology

The overall framework of our proposed W2W-BEV is shown in Fig. 2. It is mainly composed of an image feature extractor followed by a BEV representation learning method. The former utilizes CNN to extract multi-scale features from the input ground-level image, whereas the latter, which comprises three key components-depth-guided BEV initialization, context-aware window matching, and window-based BEV encoder-is responsible for generating the BEV representation from the ground-level CNN features. The obtained BEV feature map and ground-level CNN feature map go through an average pooling layer respectively to produce the feature embeddings, which are added together to form the final retrieval descriptor. Contrastive learning has been widely adopted for cross-view alignment tasks (Deng et al., 2025; Deuser et al., 2023); therefore, we employ the InfoNCE function (van den et al., 2018; Radford et al., 2021) as our loss function.

3.1. Multi-scale feature extraction

We follow Sample4Geo (Deuser et al., 2023) to harness ConvNeXt_base (Liu et al., 2022) to extract features from the input ground-level image. Similar to the previous works (Guo et al., 2020; Lin et al., 2017), we consider the feature activation output by four stages, and sequentially apply up-sampling and addition operations to generate the multi-scale features, denoted as $\boldsymbol{F}_{g,s}, s \in \{1,2,3,4\}$. Subsequently, we construct BEV representations by leveraging the extracted multi-scale features from ground-level images.

3.2. Window-to-window BEV learning

3.2.1. Depth-guided BEV initialization

We transform ground-level features into BEV space based on their geometric relations for the purpose of BEV intialization. This is achieved by first predicting a depth distribution for each ground-level pixel, re-projecting the ground-level feature back into 3D space based on the predicted depth, and finally aggregates the recovered 3D feature map along the height dimension through pooling. Specifically, the first-stage CNN feature $\boldsymbol{F}_{g,1} \in \mathbb{R}^{C\times H_1\times W_1}$ is considered for BEV initialization as it contains abundant fine details. For each ground-level image, we define a fixed-size BEV feature map $\boldsymbol{F}_b \in \mathbb{R}^{C\times N\times N}$, where N is set to be much smaller than the width W_1 of $\boldsymbol{F}_{g,1}$ extracted from FoVs ranging from 70° to 360°. Then, we predict the depth distribution over a set of discrete

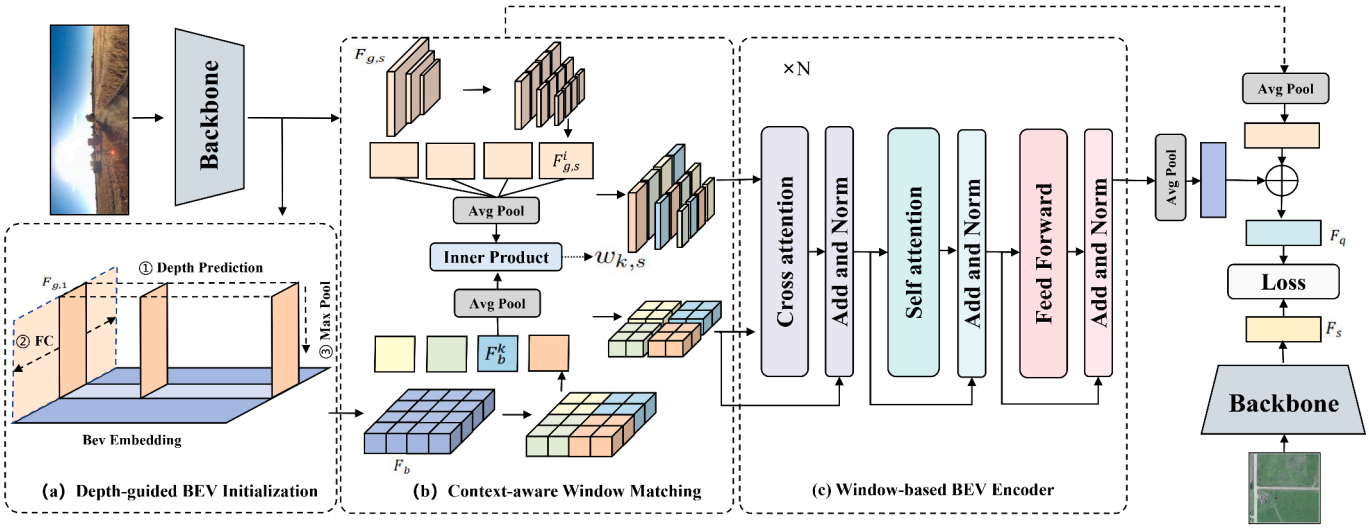


Fig. 2. Illustration of our method framework, which consists of three main components: depth-guided BEV initialization, context-aware window matching, and window-based BEV encoder.

depths for all pixels on $\boldsymbol{F}_{g,1}$ as follows:

$$\boldsymbol{P} = \mathrm{Softmax}(\mathrm{FC}(\boldsymbol{F}_{g,1})), \tag{1}$$

where FC refers to one fully connected layer; $\boldsymbol{P} \in R^{D\times H_1\times W_1}$ with each element $\boldsymbol{P}(d,h,w)$ denoting the probability that the pixel with coordinate (h,w) is at the discrete depth d. To facilitate subsequent projecting, D is set equal to N. Using the predicted depth distribution, the 3D feature map $\boldsymbol{F}_{3D} \in R^{C\times D\times H_1\times W_1}$ is constructed by performing weighted feature duplication along the depth dimension, which could be described as follows:

$$\boldsymbol{F}_{3D}(:,d,h,w) = \boldsymbol{P}(d,h,w) * \boldsymbol{F}_{g,1}(h,w). \tag{2}$$

Note that when the ground-level image has a limited FoV, transforming the ground-level feature directly into BEV space based on the predicted depth will lead to the initialized BEV space not aligned with the complete aerial view space. To handle varying FoVs, we diffuse the reconstructed 3D feature $\boldsymbol{F}_{3D}$ along the width dimension to obtain a complete feature $\boldsymbol{F}_c \in \mathbb{R}^{C\times D\times H_1\times N}$ as follows:

$$\boldsymbol{F}_c = \mathrm{FC}(\boldsymbol{F}_{3D}). \tag{3}$$

This diffusion process can be seen as imaging panoramic 3D features based on limited FoV features. Finally, max pooling is performed on $\boldsymbol{F}_c$ along the height dimension to produce the BEV feature $\boldsymbol{F}_b$.

3.2.2. Context-aware window matching

We perform window partition first before window matching. To be specific, we divide the fixed-size BEV feature map $\boldsymbol{F}_b \in \mathbb{R}^{C\times N\times N}$ evenly into K non-overlapping square windows, denoted as $\{\boldsymbol{F}_b^k\}_{k=1}^{K}$. In contrast, each ground-level feature map $\boldsymbol{F}_{g,s} \in \mathbb{R}^{C\times H_s\times W_s}$ is segmented only along the width dimension because the ground maps display the height information of objects, which helps reduce the compromise of semantic information integrity. Since the width W_s of the feature map $\boldsymbol{F}_{g,s}$ varies with FoVs of the ground-level image, the division occurs in different ways. If its width W_s could be divided by K, the feature map $\boldsymbol{F}_{g,s}$ will be evenly divided into K non-overlapping windows; Otherwise, it will be divided into K overlapping windows. Under the scenarios of an extremely small FoV or high feature level, the width W_s might be smaller than K. The implementation details are shown in Algorithm 1. In such cases, the feature map $\boldsymbol{F}_{g,s}$ is simply divided into W strips. The set of window-scale ground features is denoted as $\{\boldsymbol{F}_{g,s}^k\}_{k=1}^{K'_s}$, where K'_s refers to the number of windows. Once the window-scale features from both BEV and ground spaces are available, we find the best matching ground window for each BEV window by measuring the visual similarity between cross-view window pairs, and considering the ground window with the maximum similarity score as the matched window. In this work, we compute the inner product between the features from cross-view window pairs to measure their similarity. For instance, consider the scenario where we attempt to find the matched windows for the kth BEV window. The searching procedure over each ground-level feature map $\boldsymbol{F}_{g,s}, s \in \{1,2,3,4\}$ is described as follows:

$$f_b^k = \mathrm{Norm}(\mathrm{Avg}(\boldsymbol{F}_b^k)),\ f_{g,s}^i = \mathrm{Norm}(\mathrm{Avg}(\boldsymbol{F}_{g,s}^i)), \tag{4}$$

$$_{k,s} = \arg\max_{i\in\{1,\dots,K'_s\}} \left\langle f_b^k, f_{g,s}^i \right\rangle, \tag{5}$$

where Avg denotes the global average pooling, Norm represents the normalization operation, $\langle .,. \rangle$ is dot-product operation, and $w_{k,s}$ refers to the index of the matched window on feature map $\boldsymbol{F}_{g,s}$. This context-aware window matching strategy allows each BEV window to attend to those important ground windows containing more critical cues, such as roads, while filtering out less important windows, such as those dominated by sky semantics.

Algorithm 1 Window segmentation strategy.

Require: Feature tensor $\boldsymbol{F}_{g,s} \in \mathbb{R}^{B\times C\times W\times H}$, window count K
Ensure: Segmented windows $\{\boldsymbol{F}_{g,s}^k\}_{k=1}^{K'_s}$

```
Initialize B, C, W, H ← shape(F_{g,s})
start ← 0, end ← 1                                   ▷ Sliding window pointers
if W > K and W mod K = 0 then
    scale ← W//K
    for k = 1 to K do
        F^k_{g,s} ← F_{g,s}[:, :, scale·(k−1) : scale·k, :]
    end for
else if W > K and W mod K ≠ 0 then
    scale ← W//K + 1
    overlap_num ← K·scale − W
    for k = 1 to overlap_num do
        start ← end − 1                              ▷ 1-unit overlap
        end ← start + scale
        F^k_{g,s} ← F_{g,s}[:, :, start : end, :]
    end for
    for k = overlap_num + 1 to K do
        start ← end
        end ← start + scale
        F^k_{g,s} ← F_{g,s}[:, :, start : end, :]
    end for
else
    for k = 1 to W do
        F^k_{g,s} ← F_{g,s}[:, :, k − 1 : k, :]
    end for
end if
return {F^k_{g,s}}_{k=1}^{K'_s} where K'_s = min(K, W)
```

3.2.3. Window-based BEV encoder

To learn a robust BEV representation, we build a window-based BEV encoder, which reconstructs the feature of each BEV window by borrowing multi-scale ground features from matched windows. Our BEV encoder consists of a cross-attention module (Lin et al., 2022) followed by a self-attention module (Vaswani et al., 2017). The cross-attention module processes each BEV window independently. To generate the feature of the kth BEV window, it receives the BEV feature $\boldsymbol{F}_b^k$ combined with the relevant ground-level features $\boldsymbol{F}_m^k = \{\boldsymbol{F}_{g,s}^{w_{k,s}}\}_{s=1}^{4}$ as inputs with $\boldsymbol{F}_b^k$ and $\boldsymbol{F}_m^k$ acting as queries and keys/values respectively, and update the BEV features as follows:

$$\boldsymbol{F}_b^k = \mathrm{CrossAtt}(Q = \boldsymbol{F}_b^k, K = \boldsymbol{F}_m^k, V = \boldsymbol{F}_m^k), \tag{6}$$

where CrossAtt refers to the regular cross-attention block (Lin et al., 2022). Afterwards, the updated BEV feature map $\boldsymbol{F}_b$ goes through a self-attention block to model global interactions between BEV features, and a feedforward neural network (FFN) is used to provide non-linearity:

$$\boldsymbol{F}_b = \mathrm{SelfAtt}(Q = \boldsymbol{F}_b, K = \boldsymbol{F}_b, V = \boldsymbol{F}_b), \tag{7}$$

$$\boldsymbol{F}_b = \mathrm{FFN}(\boldsymbol{F}_b), \tag{8}$$

where SelfAtt refers to regular self-attention module (Vaswani et al., 2017). The learned BEV feature $\boldsymbol{F}_b$ and the high-level CNN feature $\boldsymbol{F}_{g,4}$ are pooled and summed to serve as the final query feature descriptor:

$$\boldsymbol{F}_q = \mathrm{Avg}(\boldsymbol{F}_b) \oplus \mathrm{Avg}(\boldsymbol{F}_{g,4}). \tag{9}$$

4. Experiments

4.1. Datasets and evaluation metrics

CVUSA (Zhai et al., 2017) and CVACT (Liu & Li, 2019) are two commonly used benchmarks in cross-view geo-localization. Both datasets

Table 1
The performance comparison with state-of-the-art methods on the standard CVUSA and CVACT datasets, where the orientation is aligned, and the ground map is a panoramic image. † denotes which models are using the polar transformation. ♠ indicates that the same data mining strategy Deuser et al. (2023) is used.

Approach	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	89.84	96.93	98.14	99.64	81.03	92.80	94.84	98.17	–	–	–	–
†DSM	91.96	97.50	98.54	99.67	82.49	92.44	93.99	97.32	–	–	–	–
†LPN	92.83	98.00	98.85	99.78	83.66	94.14	95.92	98.41	–	–	–	–
†L2LTR	94.05	98.27	98.99	99.67	84.89	94.59	95.96	98.37	60.72	85.85	89.88	96.12
TransGeo	94.08	98.36	99.04	99.77	84.95	94.14	95.78	98.37	–	–	–	–
†DeHi	94.34	98.63	99.22	99.82	84.96	94.48	95.98	98.58	–	–	–	–
†GeoDTR	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	96.34	99.10	99.50	99.86	89.06	96.11	97.08	98.89	67.49	89.39	92.30	96.80
OR-CVFI	98.02	99.50	99.68	–	91.24	96.54	97.31	–	–	–	–	–
Sample4Geo ♠	98.68	99.68	99.78	99.87	90.81	96.74	97.48	98.77	71.51	92.42	94.45	98.70
ConGeo ♠	98.30	–	–	**99.90**	90.10	–	–	98.20	71.70	–	–	98.30
C-BEV ♠	**98.72**	**99.77**	**99.83**	99.85	**91.68**	**96.62**	**97.42**	**98.75**	**75.94**	**93.23**	**94.85**	**98.77**
Ours♠	98.65	99.67	99.75	**99.90**	90.92	96.52	97.23	98.62	72.64	92.69	94.51	98.54

consist of pairs that involve one-to-one matching, including a panoramic image and a directionally aligned aerial image. The CVUSA dataset provides 35,532 pairs of images for training and 8884 pairs for testing. The CVACT dataset includes 35,532 pairs for training and 8884 pairs for validation (denoted as CVACT_val). Additionally, CVACT offers a larger version of the dataset for testing (denoted as CVACT_test), containing 92,802 pairs of images. VIGOR (Zhu et al., 2021) collects 90,618 satellite reference images and 105,214 street-view query images from four cities. The dataset is divided into training and testing sets according to two configurations: same-area and cross-area. In the same-area training setting, all four cities contribute 90,618 aerial images and 52,609 ground images, while the testing setting includes the same number of aerial images but 52,605 ground images. For the cross-area setting, the training set contains 44,055 aerial and 51,520 ground images from two cities, while the test set, sourced from the other two cities, contains 46,563 aerial and 53,694 ground images. VIGOR assumes that the ground query image can occur at arbitrary locations in the target area, so it is not spatially aligned to the center of any aerial reference images in both training and testing datasets. To be precise, for a ground query image, there is one positive reference image and three semi-positive references, all of which are partially covering the query location. The positive reference with the nearest GPS is considered as ground-truth. We employ recall accuracy at recall-k, denoted as R@k, as our primary evaluation metric, and report recalls at top-1, top-5, top-10 and top 1%.

4.2. Implementation details

We crop ground-level panoramic images to 192 × 768 and aerial images to 384 × 384 on CVUSA and CVACT datasets. On VIGOR, the panoramic images are cropped to 384 × 768 and satellite images are cropped to 384 × 384. During training, we randomly translate panoramic images and further crop them along the width direction to simulate real scenarios with unknown orientation and limited FoV. We use ConvNext_base as the backbone model. We split the BEV embeddings and ground features into four windows. In our main experiment, the number of blocks in our BEV encoder is 3 and the number of attention heads is 4, and the size of the BEV embeddings is set to 28 × 28. Our models are trained using the AdamW (Loshchilov & Hutter, 2017) optimizer with initial learning rate of 0.0001 and cosine learning rate schedule. If not specified, all the experiments are conducted on two 24GB 4090 GPUs, with batch size of 16 and epoch of 60. Additionally, we employ the same data mining approach as Sample4Geo (Deuser et al., 2023), and also adopt its data augmentation strategy: empirically applying rotation with a probability of 0.75 and flip with a probablity of 0.5 to each image. For fair comparison, we reproduce Sample4Geo results under unknown orientation and limited FoV by using the same batch size and learning rate as ours.

4.3. Comparison with state-of-the-art methods

4.3.1. Localization with panoramic image

We first evaluate the performance of our W2W-BEV in standard setting with center-aligned panorama. we compare our method with some state-of-the-art methods, including SAFA (Shi et al., 2019), DSM (Shi et al., 2020a), L2LTR (Yang et al., 2021), TransGeo (Zhu et al., 2022), DeHi (Wang et al., 2024), GeoDTR (Zhang et al., 2023), SAIG-D (Zhu et al., 2023b), OR-CVFI (Cheng et al., 2025), Sample4Geo (Deuser et al., 2023), ConGeo (Mi et al., 2024), and C-BEV (Fervers et al., 2023a). As shown in Table 1, our method is competitive with the state-of-the-art approaches. Notably, C-BEV demonstrates a significant advantage on the CVACT_test dataset. This can be attributed to its two-stage design. And C-BEV preprocesses the spatial resolution of satellite imagery within the dataset to achieve uniformity, thereby enhancing the baseline performance. However, C-BEV necessitates annotations of the original spatial resolution of the satellite imagery, which imposes constraints on its practical applicability. Then, we evaluate the performance of our W2W-BEV in a challenging localization scenario, where the orientation of the ground-level query image is unknown. Several state-of-the-art methods, including DSM (Shi et al., 2020a), DeHi (Wang et al., 2024), Sample4Geo (Deuser et al., 2023), C-BEV (Fervers et al., 2023a) and ConGeo (Mi et al., 2024) are considered for performance comparison. DeHi employs polar transform to reduce the domain disparity and introduce the delicately designed modules to learn orientation-invariant features; Sample4Geo takes a data-centric approach, which enhances representation learning by proposing two effective data mining methods; DSM estimates orientation by exhaustively computing feature similarity across discrete angles. C-BEV, on the other hand, first projects ground features into a polar coordinate space via an attention mechanism, then transitions them to the BEV space, and finally computes the orientation estimation through similar discrete sampling to find the maximum similarity offset. Although both methods resolve orientation ambiguity for panoramic inputs, they neglect scenarios with a limited FoV, leading to a significant performance degradation or even rendering them unusable in limited FoV settings. ConGeo shifts the solution to a contrastive training framework to improve FoV robustness, but its reliance on additional panoramic reference samples leads to overfitting and poor cross-dataset generalization. In contrast, our W2W-BEV method addresses these limitations by diffusing and segmenting features and then performing adaptive feature similarity matching. This approach not only removes the dependency on orientation priors and eliminates the need for complex explicit orientation alignment, but also inherently robustly handles vari-

Table 2
The performance comparison with state-of-the-art methods on the CVUSA and CVACT datasets with unknown orientation. "†" denotes models that use polar transform to approximately align the two views. "♣" denotes that methods incorporates orientation prediction to explicitly align cross-view images. "∗" indicates the use of a single model trained with contrastive learning from ConGeo and "-" indicates that this metric was not reported in the original paper.

Approach	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%
†DSM ♣	78.11	89.46	92.90	98.50	72.91	85.70	88.88	95.28	–	–	–	–
†DeHi	82.38	93.53	96.30	99.36	77.94	90.62	93.26	**97.65**	–	–	–	–
Sample4geo	90.80	97.14	98.05	99.10	81.27	89.98	91.68	95.70	64.49	82.59	85.68	95.78
C-BEV ♣	**97.13**	**99.07**	99.16	99.29	**89.46**	**94.62**	**95.46**	97.17	**73.94**	**90.53**	**92.17**	**97.15**
ConGeo	96.60	98.90	**99.20**	**99.70**	83.00	90.60	92.40	96.30	–	–	–	–
Ours	93.22	97.99	98.72	99.35	83.09	91.20	92.98	96.45	65.91	84.32	87.28	96.55
ConGeo*	85.20	95.10	96.90	98.90	62.60	79.90	84.70	**93.90**	–	–	–	–
Ours*	**93.61**	**98.36**	**98.99**	**99.56**	**73.65**	**85.29**	**87.96**	93.75	–	–	–	–

Table 3
The performance comparison with state-of-the-art methods on the CVUSA, CVACT_val, and CVACT_test datasets with unknown orientation and limited FoV. † denotes models that use polar transform to approximately align the two views.

Dataset	Approach	FoV = 180°				FoV = 90°				FoV = 70°			
		R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
CVUSA	†DSM	48.53	68.47	75.63	93.02	16.19	31.44	39.85	71.13	8.78	19.90	27.30	61.20
	GAL	48.91	69.87	78.50	95.68	22.54	44.36	54.17	83.59	15.20	32.86	42.06	75.21
	†L2LTR	56.69	80.86	87.75	98.01	26.92	50.49	60.41	86.88	13.95	33.07	43.86	77.65
	TranGeo	58.22	81.33	87.66	98.13	30.12	54.18	63.96	89.18	–	–	–	–
	†DeHi	60.42	81.81	88.16	98.03	31.53	55.13	65.57	90.84	–	–	–	–
	†ArcGeo	–	–	–	–	44.18	70.33	78.84	–	–	–	–	–
	Sample4geo	83.51	94.54	96.23	98.83	47.24	71.24	79.14	93.96	36.77	62.20	71.92	92.06
	ConGeo	**92.30**	**97.90**	**98.70**	**99.70**	55.50	75.40	81.50	93.90	**49.10**	**70.80**	**78.00**	93.10
	Ours	86.55	95.63	97.13	98.95	**64.75**	**84.66**	**89.37**	**96.88**	43.33	68.73	77.41	**93.78**
	ConGeo*	**92.30**	**97.90**	**98.70**	**99.70**	55.90	73.20	79.00	90.90	37.10	55.70	62.80	81.40
	Ours*	85.60	95.72	97.48	99.34	**57.37**	**77.19**	**82.75**	**94.16**	**47.91**	**68.63**	**75.65**	**90.26**
CVACT_val	†DSM	49.12	67.83	74.18	89.93	18.11	33.34	40.94	68.65	8.29	20.72	27.13	57.08
	GAL	49.93	68.48	77.16	93.01	26.05	49.23	59.26	85.60	14.17	32.96	43.24	77.19
	Sample4geo	58.36	77.25	83.01	93.73	26.70	50.69	60.96	85.05	18.37	38.88	48.98	78.55
	ConGeo	**70.30**	**85.20**	**88.60**	**95.10**	**40.60**	**62.60**	**69.80**	86.60	**24.60**	45.30	54.30	80.60
	Ours	63.17	81.09	86.10	94.89	32.22	57.49	66.85	**88.16**	24.28	**47.70**	**58.16**	**84.87**
	ConGeo*	**70.30**	**85.20**	**88.60**	**95.10**	34.80	56.90	**69.80**	**86.60**	24.60	45.30	54.30	**80.60**
	Ours*	66.43	81.19	85.46	93.75	**36.68**	**58.14**	65.80	83.97	**29.05**	**49.96**	**58.18**	79.56
CVACT_test	Sample4geo	32.59	55.64	63.30	92.50	9.82	23.99	31.72	82.70	5.09	14.79	20.81	75.38
	Ours	**37.88**	**62.35**	**69.67**	**94.99**	**12.67**	**30.14**	**38.85**	**88.56**	**7.58**	**21.09**	**28.89**	**85.16**

ations in the FoV. The recall results of our W2W-BEV and all baselines are presented in Table 2. As can be seen, our W2W-BEV outperforms the first five compared methods by a large margin when the orientation is unknown. Due to the use of 360-aligned images for contrastive learning in ConGeo's training method, it performed well in specific angle training and testing. To fairly compare, we use the contrastive learning training method used in ConGeo, as shown in the last two lines of each experiment of Table 2. Specifically, our approach achieved an improvement on the CVUSA dataset from 85.20% to 93.61% (+8.41%), and on the CVACT dataset from 62.60% to 73.65% (+11.05%) on R@1 accuracy. This indicates that our approach has achieved a significant improvement in model performance. Subsequent experiments further demonstrate that the integration of our model with ConGeo exhibits superior generalization capabilities.

4.3.2. *Unknown-orientation localization with limited FoV*

A more realistic localization scenario with unknown orientation and limited FoV is considered. Please note that we add four more models including GAL (Rodrigues & Tani, 2022), L2LTR (Yang et al., 2021), TransGeo (Zhu et al., 2022) and ArcGeo (Shugaev et al., 2024) for comparison in addition to the previous six compared methods. The recall results of all the above methods on CVUSA and CVACT datasets are presented in Table 3. We observe that our W2W-BEV brings significant improvements across all different FoVs of 180°, 90°, and 70° on all three datasets compared to all model-based approaches. Overall, the relative improvements for R@1 increase as the FoV of the ground-level query image gets reduced on all of the three datasets. To be specific, we gain relative improvements of 16.2%, 29.0%, and 48.9% respectively under FoVs of 180°, 90°, and 70° on CVACT_test, 8.2%, 20.7%, and 32.2% respectively under FoVs of 180°, 90°, and 70° on CVACT_val, 3.6%, 37.1%, and 17.8% respectively under FoVs of 180°, 90°, and 70° on CVUSA. Our model demonstrates better generalization ability compared to ConGeo at smaller orientation angles, such as 90° and 70°, which illustrates the compatibility with advanced training methods. Due to the challenging nature of VIGOR, little attention has been paid to its limited Fov. Therefore, we only compared with the latest method, ConGeo. Even for ConGeo, which only reports robustness at 90° on the VIGOR dataset, to enable a more comprehensive comparison with ConGeo, we further evaluate its performance on VIGOR using their publicly released source code. We rigorously reproduce their experimental results under the FoV-specific training paradigm. Table 4 shows the performance of our method on VIGOR and compared with ConGeo. Our model demonstrates substantial improvements in R@1 accuracy, achieving improvements of 23.77%, 13.82%, and 12.08% respectively under FoVs of 180°, 90°, and 70° on Same-Area setting, 6.52%, 2.38%, and 0.5% respectively under FoVs of 180°, 90°, and 70° on Cross-Area setting. The above

Table 4
The performance on the VIGOR dataset with unknown orientation and limited FoV. '-' indicates that this metric was not reported in the original paper.

Dataset	Approach	FoV = 180°				FoV = 90°				FoV = 70°			
		R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
Same-Area	ConGeo	13.24	29.18	37.11	87.42	5.93	15.35	21.02	72.03	4.88	13.23	18.49	70.20
	Ours	**37.01**	**63.19**	**70.59**	**93.14**	**19.75**	**41.95**	**50.79**	**90.36**	**16.96**	**37.15**	**45.93**	**89.34**
Cross-Area	ConGeo	3.14	8.72	12.49	55.41	1.36	4.19	6.30	43.01	0.87	3.10	4.95	38.58
	Ours	**9.66**	**22.18**	**28.81**	**72.88**	**3.74**	**10.81**	**15.54**	**63.58**	**1.37**	**4.57**	**7.26**	**48.89**

Table 5
The recall rates of our W2W-BEV on CVUSA and CVACT_val when it is trained and tested on different FoVs.

Dataset	train \ test	FoV = 180°				FoV = 90°				FoV = 70°			
		R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
CVUSA	180°	86.55	95.63	97.13	98.95	55.19	73.62	79.45	90.50	26.31	46.68	54.85	75.38
	90°	81.54	94.05	96.17	99.04	64.75	84.66	89.37	96.88	28.77	51.99	61.04	82.33
	70°	64.09	85.27	90.72	98.28	47.52	73.04	80.99	95.37	43.33	68.73	77.41	93.78
CVACT_val	180°	63.17	81.09	86.10	94.89	32.56	54.99	64.15	85.51	22.04	42.83	52.06	77.90
	90°	53.05	74.63	81.49	93.88	32.22	57.49	66.85	88.16	23.28	46.27	55.89	82.95
	70°	41.06	65.14	73.50	91.84	27.46	51.64	62.06	86.71	24.28	47.70	58.16	84.87

Table 6
The R@K in the CVUSA → CVACT and CVACT → CVUSA cross-dataset settings, which demonstrates the generalization ability of the model across the different datasets.

Datasets	FoV	Approach	R@1	R@5	R@10	R@1%
CVUSA	360°	L2LTR	47.55	70.58	77.39	91.39
		TransGeo	53.16	75.62	81.90	93.80
↓		Sample4Geo	56.62	77.79	87.02	94.69
		Ours	**63.15**	**82.38**	**87.14**	**95.90**
	90°	Sample4Geo	0.97	3.11	5.28	18.92
		ConGeo	1.29	4.00	6.20	23.00
CVACT		Ours	**1.63**	**5.10**	**6.80**	**25.46**
CVACT	360°	L2LTR	33.00	51.87	60.63	84.79
		TransGeo	44.07	**64.66**	72.08	90.09
↓		Sample4Geo	**44.95**	64.36	72.10	90.65
		Ours	44.84	64.28	**72.37**	**91.23**
	90°	Sample4Geo	1.09	3.37	5.20	18.45
		ConGeo	1.70	4.89	7.88	23.05
CVUSA		Ours	**1.85**	**6.80**	**9.70**	**28.03**

Table 7
Ablation studies on main components on CVUSA, with "✓" indicating the use of the respective method.

FoV	BEV	Initial	Matching	R@1	R@5	R@10	R@1%
70°				36.77	62.20	71.92	92.06
	✓			40.46	65.81	74.85	93.63
	✓	✓		42.59	67.94	76.20	93.66
	✓		✓	41.21	67.60	76.37	**94.02**
	✓	✓	✓	**43.33**	**68.73**	**77.41**	93.78
90°				47.24	71.24	79.14	93.96
	✓			60.57	82.32	88.77	**98.14**
	✓	✓		62.90	83.30	89.11	97.60
	✓		✓	62.48	82.96	88.69	97.58
	✓	✓	✓	**64.75**	**84.66**	**89.37**	96.88

superiority of our W2W-BEV could be explained by its capability to successfully imagine those important structure features, which fall outside the field of view, in BEV space by borrowing information from those known ground-level features.

4.3.3. Generalization across different datasets and FoVs

In cross-view geo-localization, cross-dataset evaluations such as CVUSA → CVACT and CVACT → CVUSA are commonly used to assess the model's generalization ability. In the Table 6, we report a comparison of our method with other state-of-the-art methods in terms of generalization across different datasets, under the condition of panorama with known orientation and 90° FoV. It is evident that our model shows consistent improvement in generalization ability. In real-world applications, the FoV of the camera mounted on autonomous systems may not equal to the one used for collecting training samples. Therefore, evaluating the generalization performance across different FoVs holds practical significance. Hence, we also investigate the generalization performance of our W2W-BEV by training it under a specific FoV and testing the models across various FoVs. As shown in Table 5, for a model trained under a specific FoV, the achieved recall results increase with the FoV of testing images. Furthermore, for a given testing FoV, the best performance is achieved by the model trained under the same FoV. Performance drop is observed when the training and testing FoVs are different. Despite performance degradation, the recall results are still promising. For instance, when the testing FoV is 70°, we gain 23.28% and 22.04% for R@1 on CVACT_val, respectively from models trained on FoV = 90° and FoV = 180°, which are still better than that achieved by Sample4Geo trained under FoV = 70° (i.e., 18.37% as shown in Table 3).

4.4. Ablation experiment

4.4.1. Effectiveness of main components

In this section, we validate the effectiveness of the key components by conducting ablation studies on the proposed BEV initialization and window matching modules. When they are not employed, we replace them with random initialization and random matching, respectively. As shown in Table 7, when the BEV initialization and window matching modules are separately incorporated, we gain an increase of 2.13% and 0.75% for R@1 respectively at FoV = 70°, and 2.33% and 1.91% respectively at FoV = 90°. When they are used together, our method performs the best, achieving 43.33% and 64.75% respectively for R@1 at FoVs of 70° and 90°. The observations demonstrate the effectiveness of these two key components. Another notable observation is that when the random initialization and matching techniques are considered in our BEV learning framework, we still achieve 40.66% and 60.57% for R@1 respectively at FoVs of 70° and 90°, outperforming the pure CNN-based baseline at row 1 by a large margin. We speculate this is due to the strong correlations between ground windows under small FoVs, which make random matching still work.

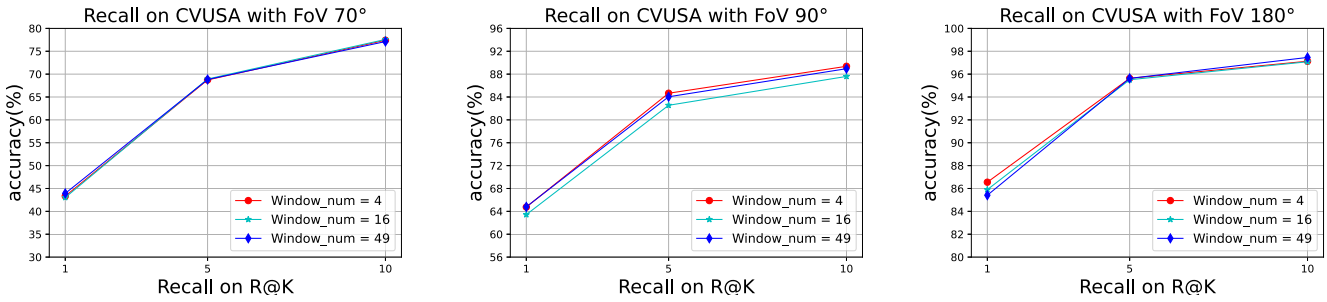


Fig. 3. The effect of the size of windows on recall results under different FoVs. The more windows there are, the less feature information each window contains.

Table 8
Effect of the depth extraction network on model performance. "Random" denotes replacing the depth value with a random value.

FoV	Depthnet	R@1	R@5	R@10	R@1%
70°	Random	41.50	64.14	72.17	88.82
	BEVDepth	**44.75**	**69.54**	**77.45**	93.58
	1 × 1 Conv	43.33	68.73	77.41	**93.78**
90°	Random	60.22	82.36	88.67	98.15
	BEVDepth	63.47	83.73	88.81	96.33
	1 × 1 Conv	**64.75**	**84.66**	**89.37**	**96.88**

Table 9
Comparison of different feature partitioning strategies under limited FoV and unknown orientation on the CVUSA Dataset.

FoV	Partition	R@1	R@5	R@10	R@1%
90°	Window	**64.75**	**84.66**	**89.37**	96.88
	Stripe	61.57	82.89	88.59	**97.42**
	Point	62.08	82.40	87.16	95.24
70°	Window	**43.33**	**68.73**	**77.41**	**93.78**
	Stripe	40.09	63.87	72.47	90.51
	Point	39.21	62.15	70.26	87.28

Table 10
Comparison of training memory consumption, training time, GLOPS, inference speed, and R@1 accuracy of different models on the CVUSA dataset under FoV = 90°.

Method	Train Mem.	Train Time	GFLOPs	Infer. Speed	R@1
TransGeo	5G	10h	6.96	19ms	30.12
ConGeo	21.5G	9h	14.18	26ms	55.50
W2W-BEV	19.80G	9h	41.51	72ms	64.75
W2W-BEV(128)	14.6G	9h	–	55ms	59.39
W2W-BEV(256)	27.3G	10h	–	107ms	66.00

Table 11
Random seeds are chosen from [1, 2025]; results report the mid-range accuracy and the semi-range computed from the best and worst runs.

FoV	R@1	R@5	R@10	R@1%
90°	63.78±(1.01)	83.91±(0.85)	88.81±(0.72)	95.95±(0.05)
70°	43.54±(0.98)	68.90±(0.89)	77.60±(0.60)	93.83±(0.05)

4.4.2. *Depthnet for initialization*

Although the ablation study of the main components confirms the effectiveness of depth-guided initialization, we further investigate the influence of depth quality on W2W-BEV. Specifically, we replace the depth network with either random depth values or a more sophisticated depth estimator from BEVDepth (Li et al., 2023). As shown in Table 8, using random depth values leads to a clear performance drop, whereas introducing a more complex depth network does not yield notable improvements. We attribute this to the absence of a camera-parameter encoder in our framework, which makes BEVDepth difficult for depth estimators to be directly integrated and effectively utilized.

4.4.3. *Spatial partition strategies*

W2W-BEV establishes spatial correspondences via window-level matching, which identifies window-to-window associations based on maximal semantic similarity. Unlike the other two potential partitioning strategies-point-wise and stripe-wise, our window-level approach better preserves semantic completeness while remaining computationally efficient. Specifically, in the absence of camera parameters, point-wise partitioning suffers from incomplete and ambiguous semantic information, making it difficult to identify coarse spatial correspondences purely from semantic similarity. On the other hand, stripe-wise partitioning introduces geometric ambiguity in the correspondence mapping. Taking the depth direction as an example, stripe-wise partitioning groups features from different depths into the same region, forcing semantically distinct structures at different depths to share a single matching area. This assumption contradicts the real-world scenario, where semantic content at different depths is inherently diverse. Therefore, the proposed window-level partitioning preserves the semantic diversity of the BEV embedding. As shown in Table 9, on the CVUSA dataset under both 90° and 70° FoV settings, the window-wise partition consistently achieves the best performance. This indicates that window-wise segmentation is more conducive to effective BEV embedding learning.

4.4.4. *Window size*

We investigate the impact of the number of windows K on recall rates by varying the value of K. Specifically, we consider non-overlapping square partitions of the BEV feature map, with configurations of $K = 2 \times 2$, $K = 4 \times 4$, and $K = 7 \times 7$. The more windows there are, the less feature information each individual window contains. The recall rates on the CVUSA dataset for these three configurations are reported in Fig. 3. As observed, the performance decreases as the number of windows increases. This is because a larger number of windows results in smaller window granularity, causing the matching scheme to approach a point-wise pattern in which each window contains less semantic information. Consequently, the correspondence between windows becomes more ambiguous, making it difficult to identify the optimal match based on feature similarity, whereas a larger window increases the computational cost of attention-based interactions. Therefore, we choose 2×2 as the number of windows.

4.4.5. *Size of BEV space*

We conduct a comprehensive analysis of the impact of BEV space size on recall rates by evaluating different values of N, i.e., 20, 24, 28. As shown in Fig. 4, it demonstrates a consistent improvement in R@1 as the size of the BEV space increases. This improvement can be attributed to the ability of a larger BEV space to facilitate more effective learning of information from the ground features. However, increasing the size of the BEV space comes at the cost of higher computational overhead,

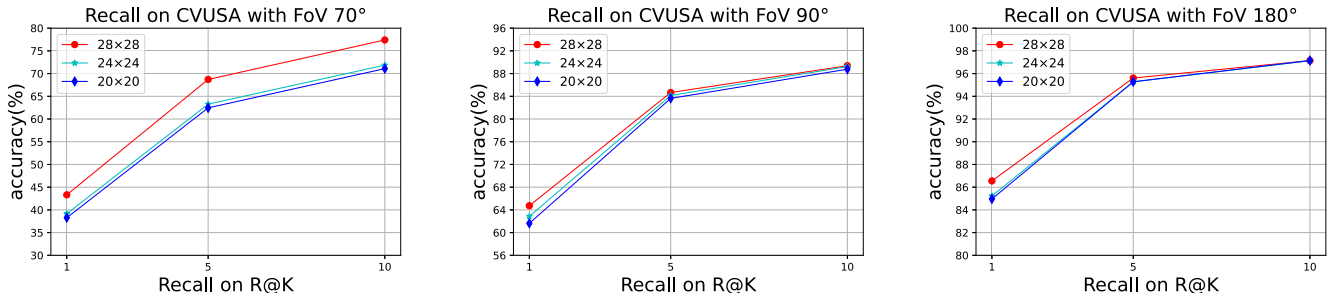


Fig. 4. The effect of the size of BEV space on recall results under different FoVs. A larger BEV space size enhances the ability to capture fine-grained details, but it also increases computational overhead and memory consumption.

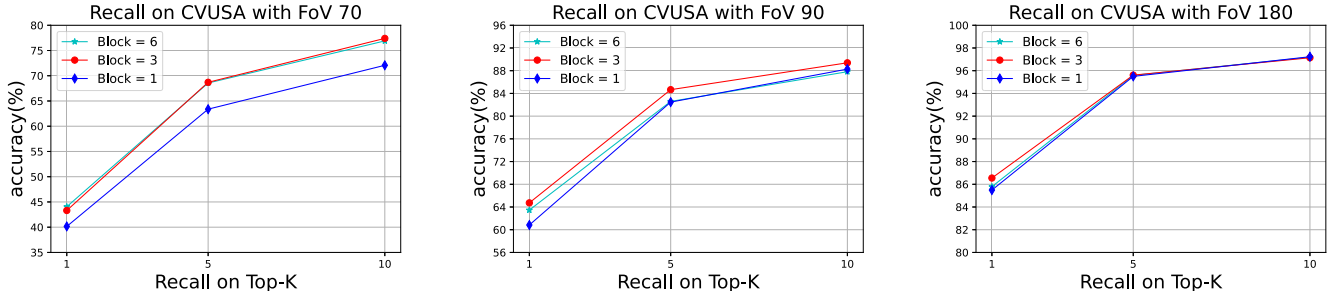
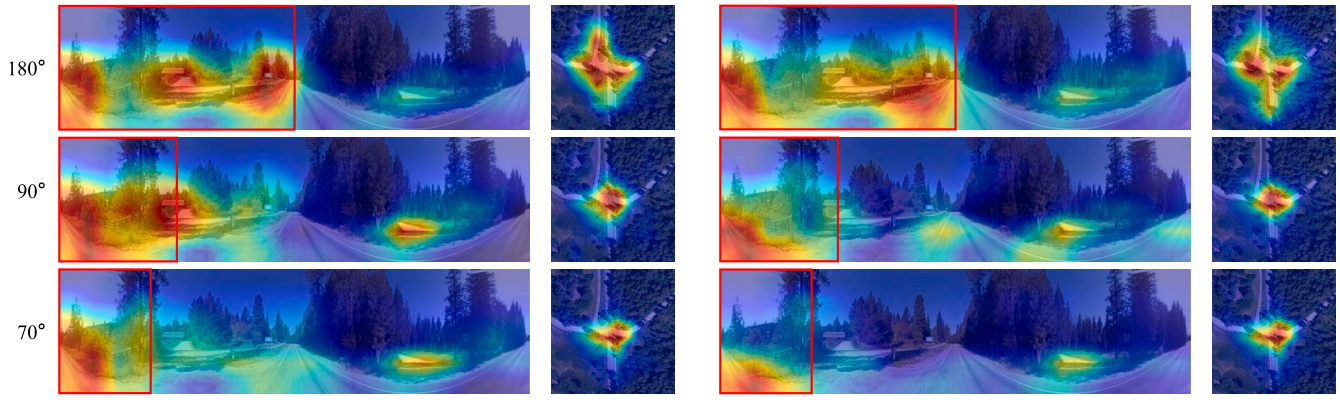


Fig. 5. The effect of the number of BEV encoder on recall results under different FoVs. Increasing the depth of BEV encoder typically raises the number of trainable parameters and architectural complexity. While this improves the model's capacity to learn feature representations, it also escalates computational costs, memory usage.



(a) Visualization of BEV feature (ours)

(b) Visualization of BEV feature (w/o depth)

Fig. 6. The figure shows the mapping of the BEV embedding's pooled features onto the panoramic image. The red boxes indicate the range of viewpoints used during both training and inference.

which includes greater memory usage and longer training times. After weighing these considerations, we choose a BEV space of 28×28 for all our primary experiments, as it offers a favorable compromise between recall rate improvements and computational feasibility.

4.4.6. Block of BEV encoder

We further investigate the impact of the number of blocks in the BEV encoder on the performance of the model by varying the number of blocks from 1 to 6. As shown in Fig. 5, a significant performance drop is observed across different FoVs when the number of blocks is reduced to just 1. This suggests that a single block is insufficient to capture the necessary global features. However, it is noteworthy that increasing the number of blocks to 6 does not improve the model's performance. We hypothesize that under limited FoV, the available effective information decreases, and simply adding more blocks does not enhance the model's capability.

4.5. Analysis

4.5.1. Model complexity analysis

To further evaluate the practical applicability of W2W-BEV, we analyze and compare the computational complexity of W2W-BEV with several open-source baselines. The Table 10 presents a comprehensive comparison in terms of training memory consumption, training time, GFLOPs, and inference speed. It is important to note that GFLOPs reflects the computational complexity of each model. Therefore, all methods are evaluated using the same input resolution (ground-view images at 112× 611 and satellite images at 256× 256 Zhu et al., 2022). Inference speed is measured on a single RTX 4090 GPU with a batch size of 16. TransGeo employs a lightweight DeiT-S backbone, achieving low parameter count and memory usage, but this comes at the cost of a substantial drop in accuracy. Moreover, Its two-stage training also requires offline storage of attention maps, adding overhead. ConGeo leverages

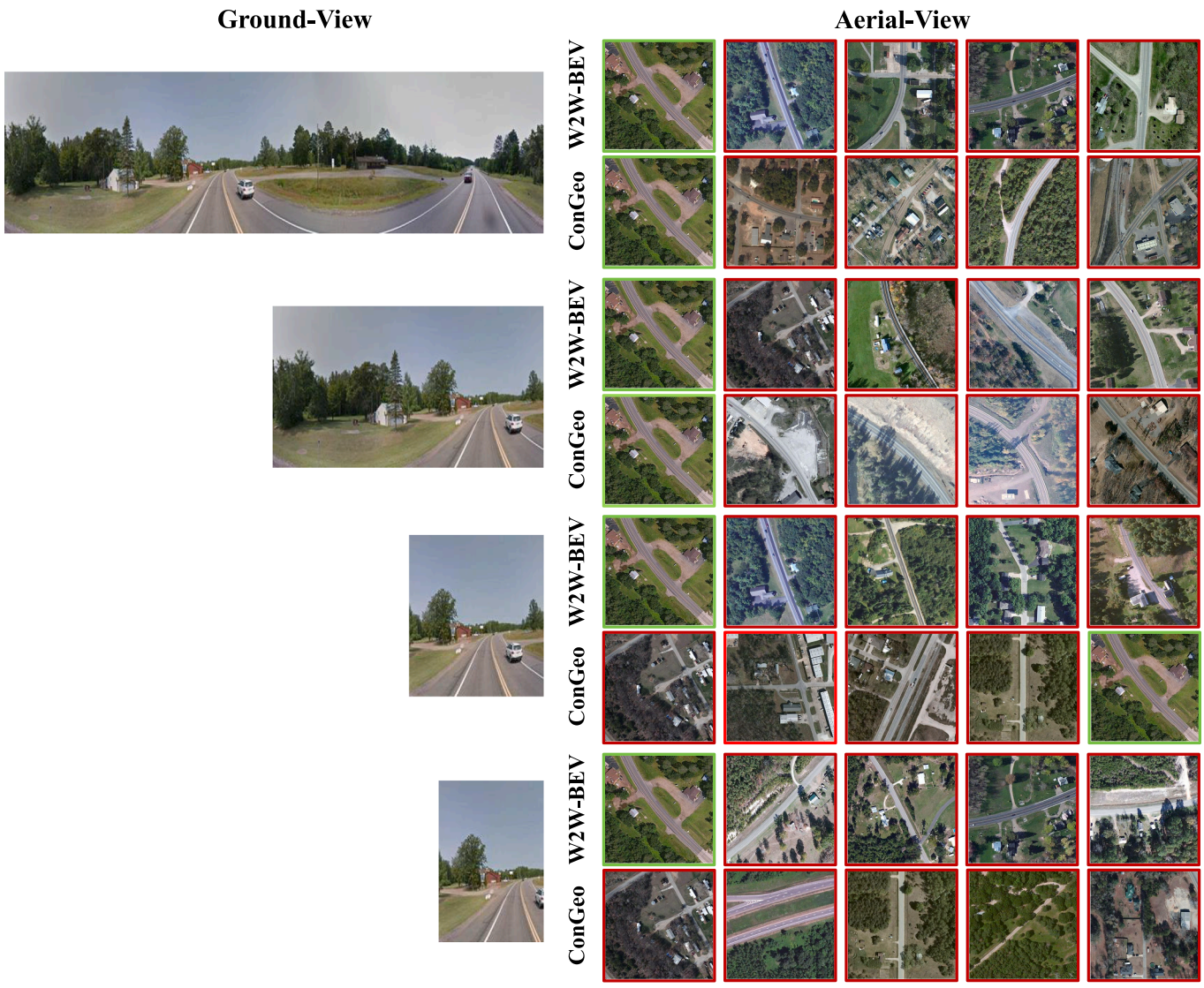


Fig. 7. Illustration of successful retrieval examples with different FoVs. The first column shows the ground query images with unknown direction and limited FoV, and the second to sixth columns represent the top-5 retrieved images. The green border indicates the ground-truth. The red border indicates mismatched references.

panoramic images for contrastive learning, which increases memory consumption. While its architectural simplicity allows for relatively fast inference, its performance under limited FoV remains insufficient, and the single-view contrastive learning design leads to severe degradation in cross-dataset generalization. For W2W-BEV, the introduction of BEV representation learning inevitably increases computational complexity and memory usage. Nevertheless, the additional overhead is justified by substantial gains in retrieval accuracy. In addition, We further evaluate the effect of street-view input resolution. As expected, when the resolution is reduced to 128× 512, the model exhibits a noticeable performance drop due to the loss of semantic detail. Increasing the resolution to 256× 1024 improves performance from 64.75% to 66.00%, but this gain comes with a significantly higher computational burden.

4.5.2. Model stability analysis

To evaluate the robustness of our model, we randomly select three integers from the range [1, 2025] as random seeds and report the mid-range accuracy together with the semi-range, computed from the best and worst runs. As shown in Table 11, the performance variation of W2W-BEV across three independent runs remains within a reasonable range, indicating that the effectiveness of W2W-BEV is consistent and not attributable to random fluctuations during training.

4.5.3. Visualization

We visualize the BEV representations of the regions of interest in both the panorama and aerial image using Grad-cam (Selvaraju et al., 2017) method. In Fig. 6 (a), it is evident that BEV embedding can distinctly recognize critical features such as roads. Furthermore, under the three different FoVs, the BEV embedding is capable of inferring similar road features beyond the FoV to varying extents. This observation helps explain why our method achieves notable improvements even with a limited FoV. Compared with Fig. 6(b), it is clear that depth-guided initialization significantly enhances the BEV embedding's ability to accurately recognize critical road regions in the ground map, thus optimizing the BEV learning process.

We present successful retrieval examples of our method under 360°, 180°, 90°, and 70° and compare them with ConGeo. As shown in Fig. 7, it can be observed that as the FoV decreases, the effective information provided by the ground-level images diminishes, yet W2WBEV still successfully retrieves matching aerial images. Additionally, we showcase examples of retrieval failures. As shown in Fig. 8, in the case of 360° with unknown orientation (first and second rows on the right), although both W2W-BEV and ConGeo achieve successful retrieval under the Top-1 metric, W2W-BEV returns top-5 candidates with extremely high similarity, whereas ConGeo retrieves top-2 and top-3 candidates that ex-

Fig. 8. Illustration of unsuccessful retrieval examples with different FoVs.

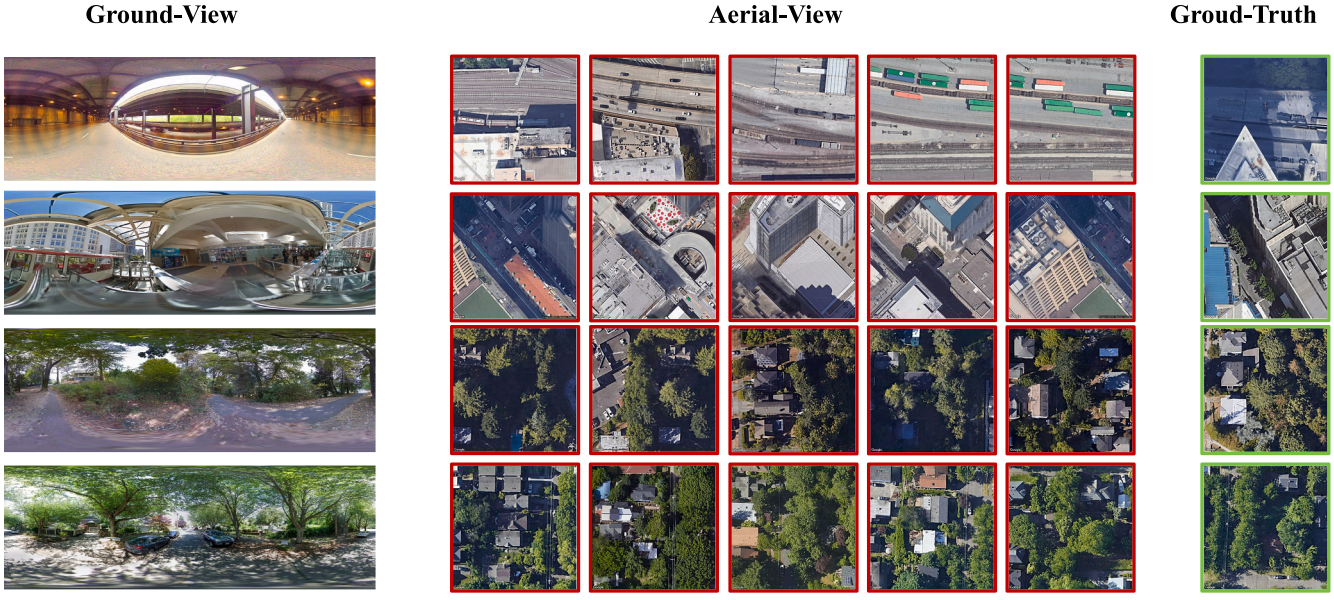


Fig. 9. Illustration of unsuccessful retrieval examples with different scenes.

hibit noticeable dissimilarity to the ground truth. However, when the FoV is further reduced to 90° and 70°, W2W-BEV fails to retrieve the correct satellite image under the top-1 metric. We speculate that this is because, at smaller FoVs, the effective visual information is drastically reduced, severely disrupting the integrity of semantic structures (roads, buildings, and terrain) and their spatial relationships. As a result, in semantically dense regions such as towns and urban areas (as shown in Fig. 8), relying solely on incomplete semantic and spatial information makes it difficult to reconstruct the global semantic layout, leading to retrieval failures. Nevertheless, it is worth noting that under a 90° narrow FoV, although W2W-BEV does not successfully retrieve the exact ground-truth image, its top-ranked satellite retrievals exhibit high semantic and structural similarity to the ground truth. This observation confirms that W2W-BEV is still able to learn meaningful semantic and spatial representations under small-FoV conditions, while also revealing that incomplete visual information limits the model's ability to distinguish globally complex layouts. When the FoV is further reduced to 70°, the semantic information becomes severely fragmented, rendering the retrieval unreliable. We attribute these failures to the highly extreme scene condition.

We further investigated the performance boundaries of W2W-BEV and presented examples where W2W-BEV completely fails in the top-5 under panoramic views, as shown in Fig. 9. We observed that W2W-BEV struggles with heavily occluded scenes, such as cross-view retrieval failures caused by dense outdoor trees and semi-outdoor scene.

5. Conclusion

Exploring cross-view image geo-localization under limited FoVs and unknown orientations is critical for real-world applications. We propose a novel window-to-window BEV representation learning method, termed W2W-BEV. By learning BEV grid embeddings from ground image features, the gap between different viewpoints is effectively reduced. This method introduces context-aware dynamic window matching to match BEV queries to ground references at the window scale, based on depth-guided BEV initialization, which eliminates the dependency of BEV mapping on camera intrinsic and extrinsic parameters. Extensive experimental results on benchmark datasets show that W2W-BEV achieves significant improvements under limited FoV and unknown orientation. Additionally, W2W-BEV can be flexibly adapted to video-based cross-view geo-localization. This can be achieved by performing temporal modeling on video frames, followed by representation learning across frames through cross-attention. Alternatively, BEV representation can first be learned for each individual frame, followed by temporal modeling on the resulting BEV representations. Therefore, flexibly and effectively transferring BEV to video-based cross-view geo-localization holds practical significance for future work.

CRediT authorship contribution statement

Lei Cheng: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation; **Daikun Liu:** Writing – review & editing, Visualization; **Lingquan Meng:** Writing – review & editing, Validation; **Teng Wang:** Writing – review & editing, Methodology, Investigation, Funding acquisition; **Changyin Sun:** Project administration, Funding acquisition, 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.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62273093).

References

Cheng, L., Wang, T., Li, J., & Sun, C. (2025). Offset regression enhanced cross-view feature interaction for ground-to-aerial geo-localization. *IEEE Transactions on Intelligent Vehicles*, *10*(1), 205–216.

Dai, M., Hu, J., Zhuang, J., & Zheng, E. (2021). A transformer-based feature segmentation and region alignment method for UAV-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(7), 4376–4389.

Deng, L., Kang, B., Zhong, Y., Wang, M., & Zhang, J. (2025). C3aptioner: Improving change captioning by leveraging momentum cross-view and cross-modality contrastive learning. *Neural Networks*, *193*, 107957.

Deuser, F., Habel, K., & Oswald, N. (2023). Sample4Geo: Hard negative sampling for cross-view geo-localisation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 16847–16856).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929.

Du, H., He, J., & Zhao, Y. (2024). CCR: A counterfactual causal reasoning-based method for cross-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, *34*(11), 11630–11643.

Fervers, F., Bullinger, S., Bodensteiner, C., Arens, M., & Stiefelhagen, R. (2023a). C-BEV: Contrastive bird's eye view training for cross-view image retrieval and 3-DoF pose estimation. arXiv:2312.08060.

Fervers, F., Bullinger, S., Bodensteiner, C., Arens, M., & Stiefelhagen, R. (2023b). Uncertainty-aware vision-based metric cross-view geolocalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 21621-21631).

Guo, C., Fan, B., Zhang, Q., Xiang, S., & Pan, C. (2020). AugFPN: Improving multi-scale feature learning for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12595–12604).

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* (pp. 7258–7267).

Li, A., Hu, H., Mirowski, P., & Farajtabar, M. (2019). Cross-view policy learning for street navigation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 8100–8109).

Li, W., Zhou, J., Chen, C., Yu, H., Du, B., & Zou, Q. (2025). BEVFix: Deep feature enhancement for robust 3D object detection. *Neural Networks*, *190*, 107675.

Li, Y., Ge, Z., Yu, G., Yang, J., Wang, Z., Shi, Y., Sun, J., & Li, Z. (2023). BEVDepth: Acquisition of reliable depth for multi-view 3D object detection. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 1477–1485).

Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., & Dai, J. (2022). BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision* (pp. 1–18). Springer.

Lin, H., Cheng, X., Wu, X., & Shen, D. (2022). Cat: Cross attention in vision transformer. In *2022 IEEE International conference on multimedia and expo (ICME)* (pp. 1–6). IEEE.

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* (pp. 5007–5015).

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2117–2125).

Liu, L., & Li, H. (2019). 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).

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11976–11986).

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv:1711.05101.

Lu, X., Li, Z., Cui, Z., Oswald, M. R., Pollefeys, M., & Qin, R. (2020). Geometry-aware satellite-to-ground image synthesis for urban areas. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 859–867).

Mi, L., Xu, C., Castillo-Navarro, J., Montariol, S., Yang, W., Bosselut, A., & Tuia, D. (2024). ConGeo: Robust cross-view geo-localization across ground view variations. arXiv:2403.13965.

Oord, A.v.d., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv:1807.03748.

Philion, J., & Fidler, S. (2020). Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer vision–ECCV 2020: 16th european conference, glasgow, UK, august 23–28, 2020, proceedings, part XIV 16* (pp. 194–210). Springer.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763). PMLR.

Regmi, K., & Shah, M. (2019). Bridging the domain gap for ground-to-aerial image matching. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 470–479).

Rodrigues, R., & Tani, M. (2022). Global assists local: Effective aerial representations for field of view constrained image geo-localization. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 3871–3879).

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., 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* (pp. 618–626).

Shen, T., Wei, Y., Kang, L., Wan, S., & Yang, Y.-H. (2024). MCCG: A convnext-based multiple-classifier method for cross-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, *34*(3), 1456–1468.

Shetty, A., & Gao, G. X. (2019). Uav pose estimation using cross-view geolocalization with satellite imagery. In *2019 International conference on robotics and automation (ICRA)* (pp. 1827–1833). IEEE.

Shi, P., Ge, R., Dong, X., Chakir, C., Liang, T., & Yang, A. (2025). Polarfusion: A multi-modal fusion algorithm for 3D object detection based on polar coordinates. *Neural Networks*, *190*, 107704.

Shi, Y., & Li, H. (2022). Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 17010–17020).

Shi, Y., Liu, L., Yu, X., & Li, H. (2019). Spatial-aware feature aggregation for image based cross-view geo-localization. *Advances in Neural Information Processing Systems*, *32*, 10090–10100.

Shi, Y., Wu, F., Perincherry, A., Vora, A., & Li, H. (2023). 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).

Shi, Y., Yu, X., Campbell, D., & Li, H. (2020a). 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).

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 transactions on pattern analysis and machine intelligence*, *45*(3), 2682–2697.

Shi, Y., Yu, X., Liu, L., Zhang, T., & Li, H. (2020b). Optimal feature transport for cross-view image geo-localization. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 11990–11997). (*vol. 34*).

Shugaev, M., Semenov, I., Ashley, K., Klaczynski, M., Cuntoor, N., Lee, M. W., & Jacobs, N. (2024). ArcGeo: Localizing limited field-of-view images using cross-view matching. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 209–218).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*, 6000–6010.

Vo, N. N., & Hays, J. (2016). 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.

Wang, T., Li, J., & Sun, C. (2024). DeHi: A decoupled hierarchical architecture for unaligned ground-to-aerial geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, *34*(3), 1927–1940.

Wang, T., Zheng, Z., Yan, C., Zhang, J., Sun, Y., Zheng, B., & Yang, Y. (2021). Each part matters: Local patterns facilitate cross-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(2), 867–879.

Wang, X., Xu, R., Cui, Z., Wan, Z., & Zhang, Y. (2024). Fine-grained cross-view geo-localization using a correlation-aware homography estimator. *Advances in Neural Information Processing Systems*, *36*, 5301–5319.

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* (pp. 3961–3969).

Xia, P., Wan, Y., Zheng, Z., Zhang, Y., & Deng, J. (2024). Enhancing cross-view geo-localization with domain alignment and scene consistency. *IEEE Transactions on Circuits and Systems for Video Technology*, *34*(12), 13271–13281.

Yang, H., Lu, X., & Zhu, Y. (2021). Cross-view geo-localization with layer-to-layer transformer. *Advances in Neural Information Processing Systems*, *34*, 29009–29020.

Zhai, M., Bessinger, Z., Workman, S., & Jacobs, N. (2017). Predicting ground-level scene layout from aerial imagery. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 867–875).

Zhang, X., Li, X., Sultani, W., Zhou, Y., & Wshah, S. (2023). Cross-view geo-localization via learning disentangled geometric layout correspondence. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 3480–3488). (*vol. 37*).

Zheng, Z., Wei, Y., & Yang, Y. (2020). 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).

Zhu, R., Yin, L., Yang, M., Wu, F., Yang, Y., & Hu, W. (2023a). SUES-200: A multi-height multi-scene cross-view image benchmark across drone and satellite. *IEEE Transactions on Circuits and Systems for Video Technology*, *33*(9), 4825–4839.

Zhu, S., Shah, M., & Chen, C. (2022). 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).

Zhu, S., Yang, T., & Chen, C. (2021). 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).

Zhu, Y., Yang, H., Lu, Y., & Huang, Q. (2023b). Simple, effective and general: A new backbone for cross-view image geo-localization. arXiv:2302.01572.