SG-BEV: Satellite-Guided BEV Fusion for Cross-View Semantic Segmentation

Junyan Ye[1,2*], Qiyan Luo[1], Jinhua Yu[1], Huaping Zhong[3],
Zhimeng Zheng[2,4], Conghui He[2,3], Weijia Li[1†]
[1]Sun Yat-Sen University, [2]Shanghai AI Laboratory, [3]SenseTime Research, [4]Zhejiang University
{yejy53, luoqy26, yujh56}@mail2.sysu.edu.cn, zhonghuaping@sensetime.com,
{zhengzhimeng, heconghui}@pjlab.org.cn, liweij29@mail.sysu.edu.cn

Abstract

This paper aims at achieving fine-grained building attribute segmentation in a cross-view scenario, i.e., using satellite and street-view image pairs. The main challenge lies in overcoming the significant perspective differences between street views and satellite views. In this work, we introduce SG-BEV, a novel approach for satellite-guided BEV fusion for cross-view semantic segmentation. To overcome the limitations of existing cross-view projection methods in capturing the complete building facade features, we innovatively incorporate Bird's Eye View (BEV) method to establish a spatially explicit mapping of street-view features. Moreover, we fully leverage the advantages of multiple perspectives by introducing a novel satellite-guided reprojection module, optimizing the uneven feature distribution issues associated with traditional BEV methods. Our method demonstrates significant improvements on four cross-view datasets collected from multiple cities, including New York, San Francisco, and Boston. On average across these datasets, our method achieves an increase in mIOU by 10.13% and 5.21% compared with the state-of-the-art satellite-based and cross-view methods. The code and datasets of this work will be released at https://github.com/yejy53/SG-BEV.

1. Introduction

Fine-grained building attribute segmentation is a crucial task for urban planning, environment monitoring and residential management [11, 16, 26]. Satellite images offers a comprehensive outline of building footprints, while street-view images contribute detailed facade features. Integrating these two types of data has demonstrated significant potential in achieving precise attribute segmentation of buildings [18, 32, 38, 39]. In this paper, we focus on the cross-view semantic segmentation of fine-grained attributes using pairs of satellite and street-view images.

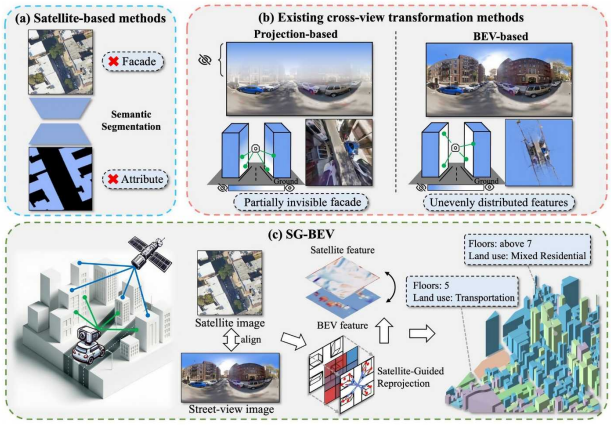

Figure 1. **Illustration of cross-view semantic segmentation of fine-grained building.** (a) Satellite imagery lacks information on building facades, making it difficult to distinguish detailed building attributes. (b) Existing cross-view transformation methods face issues with incomplete feature capture and uneven feature distribution. (c) Our method integrates satellite and street-view features to precisely segment building attributes and floor numbers.

Previous studies on the semantic segmentation of fine-grained attributes for buildings or other terrestrial objects have predominantly relied on satellite images [16, 19, 43]. However, as shown in Figure 1(a), the satellite perspective captures only the top and outline information, making it challenging to distinguish the fine-grained attribute differences between different buildings [15, 38, 39]. To address this issue, recent research has incorporated the street-view perspective to supplement facade information of buildings. The typical approach involves mapping street-view features to corresponding areas in the satellite view, thereby creating a link between satellite and street-view images [8, 32, 38, 39]. However, existing approaches often reflect only general characteristics near the area, struggling to map street-view building features precisely to specific loca-

*This work was partially done during the internship at Shanghai Artificial Intelligence Laboratory.
†Corresponding author.

tions in the satellite view, leading to subpar performance in fine-grained attribute segmentation at the individual building level. To more effectively convey facade features from the street view for each building, exploring a novel cross-view feature mapping method that can continuously and precisely map street-view features to specific satellite view locations is necessary. The significant difference between street and satellite views poses a substantial challenge for precise cross-view feature mapping.

To effectively map and align features from street and satellite imagery, some current studies employed cross-view geometric projection methods [29, 35, 44]. However, these methods are more suitable for analyzing central ground areas, such as road regions for applications like image localization and driving planning [29, 35, 42]. In these cross-view mapping approaches, geometric projection is typically conducted through ground assumption and 360° panoramic mapping relations [29, 35]. However, these methods often fail to effectively capture the facade features of taller buildings above the viewpoint, resulting in significant distortion of features away from the center area, as illustrated in Figure 1(b). This limitation leads to poor performance in comprehensively capturing the facade features of buildings, which is particularly evident in urban environments with high-rise structures. Such a constraint significantly restricts their capability in addressing cross-view fine-grained building attribute segmentation problems.

Bird's Eye View (BEV) methods represent another category of cross-view feature mapping, commonly used in autonomous driving or robot navigation [10, 12, 13, 17, 21–24, 33]. Compared with the geometric projection methods mentioned previously, BEV methods, leveraging 3D scene estimation, can capture more complete features of building facade. We plan to introduce the BEV approach to map street-view features onto satellite images, representing a novel attempt at fine-grained building segmentation tasks in cross-view scenarios. However, as street-view images are captured from a ground perspective, they struggle to fully perceive the complete outline of building footprints. When converting street-view images to BEV, features are mainly concentrated and stacked at the visible parts of roads and building wall edges [17, 23] (as shown in Figure 1(b)). This results in uneven BEV feature distribution, limiting its performance in fusion with satellite features. We note that satellite images provide complete building contours, hence we introduce a Satellite-Guided Reprojection (SGR) module. This module relocates features from building edges to interiors, effectively addressing uneven feature distribution.

In this work, we introduce SG-BEV, a satellite-guided BEV fusion method for cross-view semantic segmentation. Unlike previous cross-view transformation approaches, our method establishes a clear spatial mapping relationship from the street-view to the satellite perspective, overcoming the limitations of geometric projection methods in capturing building facade features, and the uneven feature distribution issue of traditional BEV methods.

Our main contributions are summarized as follows:

- We innovatively apply BEV paradigm to the task of cross-view semantic segmentation of fine-grained building attributes, achieving a complete and continuous mapping of street-view features to a top-down perspective.
- We develop a Satellite-Guided Reprojection (SGR) module to further address the issue of features unevenly concentrated at the edges of buildings in BEV methods.
- Our method is evaluated on four cross-view datasets from cities including New York, San Francisco and Boston. On average across these datasets, it demonstrates an improvement of 10.13% and 5.21% in mIOU compared to the state-of-the-art satellite-based and cross-view methods.

2. Related work

2.1. Semantic Segmentation of Ground Objects

In the studies on semantic segmentation of ground objects, high-resolution satellite imagery has significantly contributed to the advancement of a variety of tasks, including urban road extraction [1, 3], land use classification [5, 41], and building extraction [4, 31]. Prior research focused on buildings and other terrestrial objects, has primarily utilized satellite imagery as the main source of data, achieving notable results [19, 43]. However, these approaches were somewhat limited in achieving fine-grained semantic segmentation, as they lacked facade information typically contained in street-view images. To address these limitations, Wojna et al. [37] introduced a method for projecting geometric attributes of buildings. Workman et al. utilized a backbone network to extract feature vectors representing the overall features of street views, using the spatial location of street-view images to diffuse these into the satellite view feature space [39]. Another approach involved creating a geospatial attention mechanism using distance and angle information, mapping street-view feature vectors onto the satellite view [38]. However, these feature mapping methods result in feature loss during the mapping process and sparse street-view features in the satellite perspective.Our method addresses these challenges by enabling accurate spatial mapping and efficient transfer of dense features from street views to the satellite view, thereby bridging the gap between different observational viewpoints.

2.2. Cross-View Projection Methods

Cross-view projection methods play a crucial role in bridging the gap between different perspectives in image localization and driving planning [20, 27, 28, 30, 34]. Techniques like polar transformations [28] were employed by to map features from satellite views to ground views. These

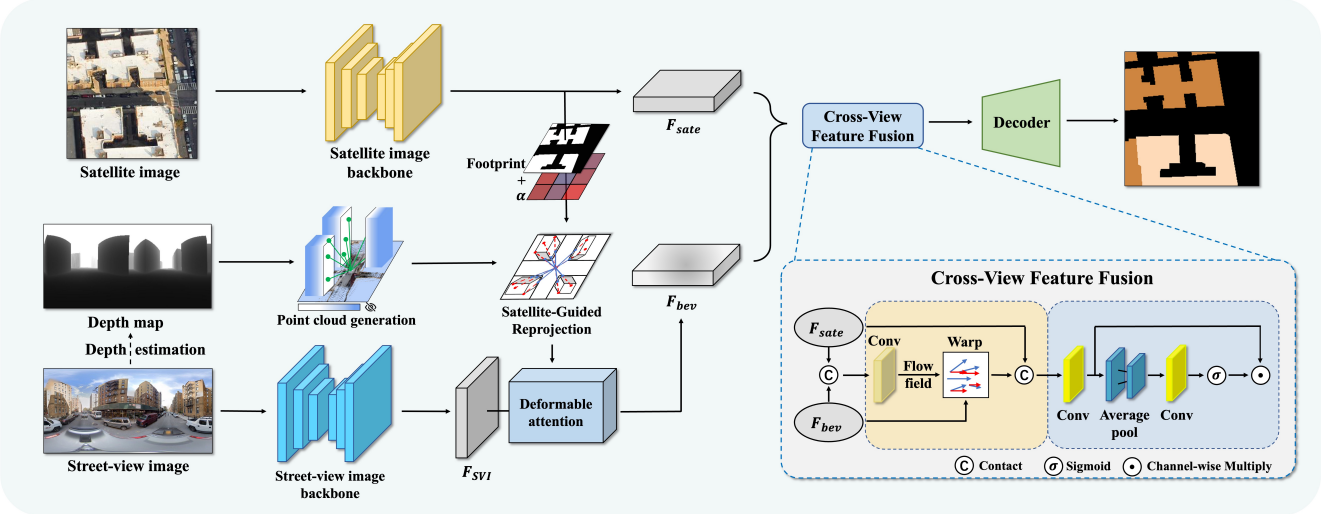

Figure 2. **Overview of our proposed SG-BEV framework.** In Satellite Feature Extraction branch, we extract features of input satellite imagery, meanwhile output building footprint segmentation results for further processing. In Street-View to BEV Conversion branch, we map street-view features to BEV space using estimated depth information combined with building footprints. In Cross-View Feature Fusion module, we align and fuse satellite features with BEV features to achieve fine-grained segmentation of building attributes.

transformations are crucial for tasks such as image retrieval and street-view generation. However, for the fine-grained building attribute segmentation from a top-down perspective that we aim to achieve, this method may not be suitable. In addition, several previous studies [29, 35], assumed a geometric relationship between the viewpoint and the ground plane to establish feature mapping relationship. However, these methods fail to capture features above the viewpoint and also create feature distortion away from the center area. Our approach overcomes the incomplete feature mapping in cross-view geometric projection methods, achieving comprehensive mapping of building facade information.

2.3. Bird's Eye View methods

The Bird's Eye View (BEV) methods have been widely used in autonomous driving and robot navigation [13, 21, 22, 24], which are mainly for road area analysis and effective segmentation of targets like vehicles and lanes [10, 14, 22, 23]. The Lift, Splat, Shoot (LSS) [23] method mapped two-dimensional features to three-dimensional space by predicting depth distribution to acquire BEV features. BEVFormer [17] startded from BEV queries and maps back to two-dimensional features for interaction. Additionally, BEVFormer enhanced the ability to capture features of tall objects by selecting multiple three-dimensional reference points along the Z-axis. However, street-view images fail to capture the complete outline of building footprints, leading to effective features being concentrated at the edges of depth-estimated dense areas (the building walls). In the LSS method, effective features were primarily concentrated at wall locations with the highest depth probabilities, while in BEVFormer, after average pooling along the Z-axis, features were also predominantly focused on the walls. The inconsistent distribution between BEV and satellite features, with strong features on building walls but sparse elsewhere, may degrade performance in subsequent tasks. Our designed Satellite-Guided Reprojection (SGR) module utilizes the footprint information provided by satellite imagery, combined with the estimated depth information, to guide the concentrated BEV features towards the interior of building footprints, effectively overcoming the issue of uneven BEV feature distribution.

3. Methods

As shown in Figure 2, this paper introduces a novel method for cross-view fine-grained attribute segmentation of buildings, named SG-BEV. In our comprehensive workflow, we employ two distinct branches to extract features from satellite and street-view images, respectively, and then merge them using a feature fusion module. In the satellite branch, we apply a backbone network to extract the satellite feature F_{sat} and output preliminary segmentation of the building footprint to guide the subsequent BEV features (Section 3.1). In the street-view branch, we initially map using depth estimation information, then optimize the feature distribution with the Satellite-Guided Reprojection (SGR) module, , and finally produce BEV features F_{bev} with deformable attention [17, 36] (Section 3.2). In the Cross-View Feature Fusion module, we integrate F_{sat} with F_{bev} in a unified top-down view space (Section 3.3). The decoder then processes these integrated features to produce fine-grained building attribute segmentation results, effectively capturing detailed attributes in the cross-view scenario.

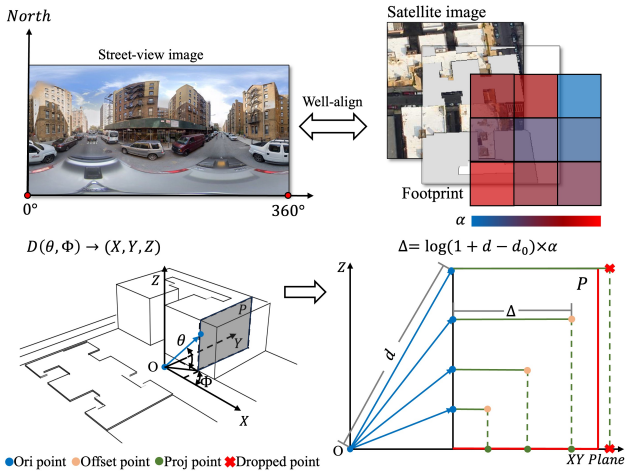

Figure 3. **Illustration of Satellite-Guided Reprojection Module.** We utilize satellite features to generate building footprint information, followed by calculating α. Based on depth information d and α, we calculate magnitude of the offset Δ to adjust the initial point cloud for uniform distribution within the building area and discard points that exceed the building's footprint.

3.1. Satellite Feature Extraction

Satellite images provide a comprehensive outline of buildings from a top-down perspective, effectively compensating for the limitations of street-view images in perceiving the overall form of buildings and capturing areas obscured in street views. We deployed a feature extractor to process satellite images, thus obtaining the satellite features F_{sat}. Since satellite images inherently offer a top-down perspective, they can be directly applied to subsequent Cross-View Feature Fusion. The final building contour features primarily originate from the satellite branch, which has a simpler structure. This shorter pathway design facilitates the backpropagation of loss.

Furthermore, the obtained satellite features will be processed through an additional decoder to produce segmentation results of building footprints, distinguishing between building and non-building areas. Our subsequent SGR module will leverage the advantages of multiple perspectives, guiding the sensible mapping of street-view features using the output building footprints to prevent BEV features from concentrating solely on building edges.

3.2. Street-View to BEV Conversion

Initial Point Cloud Generation. By utilizing established monocular depth estimation algorithms [2], we are able to derive depth maps from street-view images. Based on depth estimation results and the geometric relationship of panoramic images, we obtained a three-dimensional XYZ estimation of the scene [33]. With this, we established an index mapping relationship between the 3D scene and panoramic images, facilitating the preliminary mapping of street-view features.

$$\Theta_{i,j} = \frac{i\pi}{H}, \quad \Phi_{i,j} = -\frac{2\pi j}{W} + \pi \tag{1}$$
$$i = \{0, \ldots, H-1\}, j = \{0, \ldots, W-1\}$$

Here, Θ and Φ are angle matrices of panoramic images with size $H \times W$, consisting of two-dimensional Euler angular equivariant series, where i and j represent row and column numbers, respectively. Given the representation in spherical coordinate systems, each 3D point $(X_{i,j}, Y_{i,j}, Z_{i,j})$ in the camera coordinate system will be obtained through the calculation in Eq. (2), where $D_{i,j}$ is the panoramic depth information.

$$\begin{aligned} X_{i,j} &= D_{i,j} \cdot \sin(\Theta_{i,j}) \cdot \sin(\Phi_{i,j}), \\ Y_{i,j} &= D_{i,j} \cdot \cos(\Theta_{i,j}), \\ Z_{i,j} &= D_{i,j} \cdot \sin(\Theta_{i,j}) \cdot \cos(\Phi_{i,j}). \end{aligned} \tag{2}$$

Satellite-Guided Reprojection. In our cross-view semantic segmentation task, we aim to reproject street-view features into building interiors completely and continuously with minimal alterations. We observe that while BEV features concentrate on building walls in the XY plane, they are dispersed and extended in the Z-axis, corresponding to street-view features from the facade base to the top. Using depth information d as a positively correlated offset factor in this context can factor effectively maintains the visual continuity and integrity of the facade features. With this method, features at the bottom of the building facade are guided closer to the center area, while the top features are relatively distanced from the center area.

Additionally, we extract building footprint information from satellite images to calculate the adjustment coefficient α to control the intensity of the offset. Specifically, the satellite image is divided into a 3×3 grid, and the proportion of building pixels in each grid is calculated to set the value of α. In our approach, a higher value of the adjustment coefficient α indicates a larger footprint area of the building, necessitating a greater degree of offset. The specific magnitude of the offset Δ is jointly constructed by depth and α, as illustrated in Eq. (3).

$$\Delta = \log(1 + d - d_0) \times \alpha \tag{3}$$

Here, d means depth, and d_0 is a predefined hyperparameter. Δ is adjusted using a logarithmic function, aiming to reduce discontinuities in the point cloud on the same building facade caused by the rapid increase in depth with height. When $d < d_0$, no offset occurs. Next, we determine the direction of the point cloud offset. Considering that the camera is situated at the center, the point cloud should offset away from the center. The offset direction is determined by the following Eq. (4):

$$\vec{D} = \begin{bmatrix} X_{i,j} - c_x \\ Y_{i,j} - c_y \end{bmatrix} \tag{4}$$

Here, c_x and c_y represent the coordinates of the center position. Utilizing the calculated offset direction and distance, we accordingly adjusted the XY coordinates of the point cloud. Our method combines information from satellite imagery and depth data to effectively optimize the distribution of the initial point cloud by shifting it as Eq. (5).

$$\begin{bmatrix} X'_{i,j} \\ Y'_{i,j} \end{bmatrix} = \Delta \cdot \vec{D} + \begin{bmatrix} X_{i,j} \\ Y_{i,j} \end{bmatrix} \tag{5}$$

Subsequently, we use the index information carried by the point cloud, which indicates the correspondence between the spatial positions in BEV space and the locations on the panorama. By integrating with deformable attention mechanisms [17, 36], we map the perspective features of the street-view onto the BEV plane. More visualization information can be found in the supplementary materials. More information on our street-view feature mapping visual comparison with other methods, and the α parameter can be found in the supplementary materials.

3.3. Cross-View Feature Fusion

By acquiring satellite features and street-view BEV features as described above, we have unified features from two different views under a top-down perspective. Recognizing the challenges posed by depth estimation errors and positional inaccuracies, our Cross-View Feature Fusion model first addresses aligning these diverse features [7]. The alignment process begins by generating a 2D flow field Ω, which is calculated based on the spatial discrepancies between F_{sat} and F_{bev}. This involves using convolutional layers to predict coordinate offsets. The calculated flow field Ω is then used to warp F_{bev} to align with F_{sat}, which can be mathematically formulated as:

$$F_{\text{aligned}} = [\text{Warp}(F_{\text{bev}}, \Omega), F_{\text{sat}}] \tag{6}$$

where $[\cdot, \cdot]$ denotes the concatenation along the channel dimension, Warp is a function applying the calculated offsets using bilinear interpolation, ensuring the spatial alignment of the features. Ω is the deformation field predicted by a convolutional network.

The next step is to the integration of the feature sets. We initiate a spatial fusion process by applying a 3×3 convolution layer, crucial for enhancing the spatial representation of the features. The output of this layer serves as the input for our adaptive integration function. As illustrated in Figure 2, the refined features from the convolution layer are first globally averaged pooled and then employ a linear transformation. This transformation is implemented via a 1×1 convolution. Our cross-view feature fusion module captures essential information from both satellite and BEV features while minimizing their alignment errors.

4. Experiments

4.1. Datasets

OmniCity Dataset (OmniCity) [15] encompasses street-level and satellite imagery from the Manhattan area in New York, with each street-view precisely corresponding to its satellite counterpart. This dataset provides detailed urban attributes such as land use and floor level, tailored for fine-grained cross-view semantic segmentation tasks.

Expanded Vigor Dataset (Vigor) [45] includes the nearest street-satellite image pairs centered on satellite imagery, which are selected from the original dataset and supplemented with land use information provided by DataSF[1] in the San Francisco area, to extend its application from geographic localization to cross-view semantic segmentation. As street-view and satellite images are not center-aligned, the provided offset values will be used for subsequent BEV feature translation to move them to the appropriate position.

Brooklyn-Manhattan Dataset (Brooklyn) combines street and satellite imagery from Brooklyn and Manhattan, utilizing land use and floor level information provided by PLUTO[2]. Building upon the OmniCity, we have optimized the data collection step size to reduce overlap in satellite imagery, expanded the coverage of dataset, and selected higher-quality street-view images.

Boston Dataset (Boston) contains street-satellite image pairs across the entire city of Boston, supplemented with land use information provided by Boston Maps[3]. The dataset maintains the same rigorous selection and division criteria as the Brooklyn.

To comprehensively evaluate the robustness and generalization capabilities of models, all datasets are strictly divided according to geographic regions, with a 4:1 training to test set ratio. We have provided a comparison of different data partition methods in the supplementary materials.

4.2. Experimental Settings

Evaluation Metrics. Our study utilizes mean intersection over union (mIOU) and overall accuracy (Acc) to evaluate our cross-view semantic segmentation task. mIOU measures the overlap between the model's predicted regions and the true regions, while Acc assesses the proportion of samples correctly predicted by the model. In our task, it involves perceptive segmentation of two fine-grained attributes of buildings: land use and the number of floors.

[1] https://data.sfgov.org/Housing-and-Buildings/Land-Use/us3s-fp9q

[2] https://www.nyc.gov/site/planning/data-maps/open-data/dwn-pluto-mappluto.page

[3] https://data.boston.gov/dataset/boston-buildings-with-roof-breaks

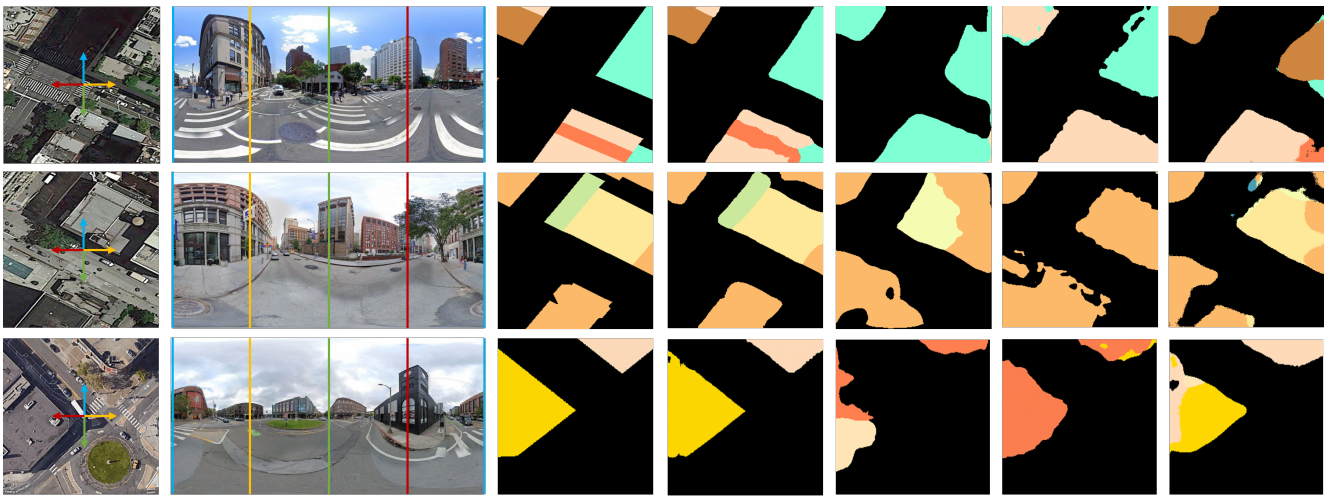

Figure 4. **Comparisons of SG-BEV (Ours) and Satellite-Based Methods for Fine-Grained Segmentation.** The first two rows show results of OmniCity on land use (first row) and floor level (second row) segmentation tasks. The third row presents land use predictions of Vigor. The street-view panoramas, from left to right, correspond to a 360-degree clockwise rotation starting from the north direction in the satellite imagery.

Table 1. Comparison with satellited-based semantic segmentation methods on different datasets, in terms of mIoU and Acc metrics (%).

Method	OmniCity				Brooklyn				Boston		Vigor	
	Land use		Floor		Land use		Floor		Land use		Land use	
	mIOU	Acc	mIOU	Acc	mIOU	Acc	mIOU	Acc	mIOU	Acc	mIOU	Acc
FarSeg [43]	26.27	68.62	17.42	71.75	31.71	**80.08**	26.93	72.31	28.62	72.08	28.49	74.49
SAN [40]	24.49	65.48	19.89	70.66	31.83	75.41	25.39	69.69	29.05	**77.63**	29.03	72.95
SegNext [9]	31.38	70.31	25.27	72.27	36.85	76.68	34.55	76.01	32.55	76.81	34.92	76.13
Ours	**37.54**	**76.13**	**40.64**	**77.82**	**47.19**	78.43	**49.51**	**79.00**	**39.72**	77.39	**41.70**	**76.81**

Comparison Methods. Our comparison methods are divided into two categories: The first category involves segmenting satellite imagery using state-of-the-art methods, including FarSeg [43], SAN [40] and SegNext [9]. The second category comprises cross-view methods that transform street view imagery into a satellite view and then combine it with satellite features, specifically Spherical Transform (ST) [35], Geometric Projection (GP) [29] and BEVFormer [17]. For these three cross-view methods, we consistently use the same satellite feature extraction and fusion module as in our approach. In our comparisons, we do not apply ST and GP to the Vigor dataset because these methods require the use of centrally aligned street view images for projection, which is not the case with the Vigor dataset. Furthermore due to the lack of temporal information in our data, we retain the Z-direction expansion capability in BEVFormer but only utilize its spatial attention module. All comparison methods follow the best settings.

Implementation Details. Our network employs SegNext with it variant MSCAN-B2 [9] with pre-trained weights on Cityscapes [6] as the feature extractors for street-view and satellite imagery utilizing non-share weights. Satellite images and the BEV transformations derived from ground images are uniformly sized at 256×256 across all datasets, corresponding to a sensing range of approximately 70×70 meters. Ours models are trained on Nvidia GeForce RTX 3090 GPUs, starting with an initial learning rate of $6e^{-5}$, which is adjusted according to a step strategy over 50 epochs. We use AdamW as the optimizer, with an epsilon of $1e^{-8}$, a weight decay of 0.01. Information about the depth estimation method used in the paper can be found in the supplementary material. More information on depth estimation and BEV dimension settings can be found in the supplementary materials.

4.3. Performance Comparison

Compare to Satellite-Based methods. As shown in Table 1, our cross-view segmentation results demonstrate

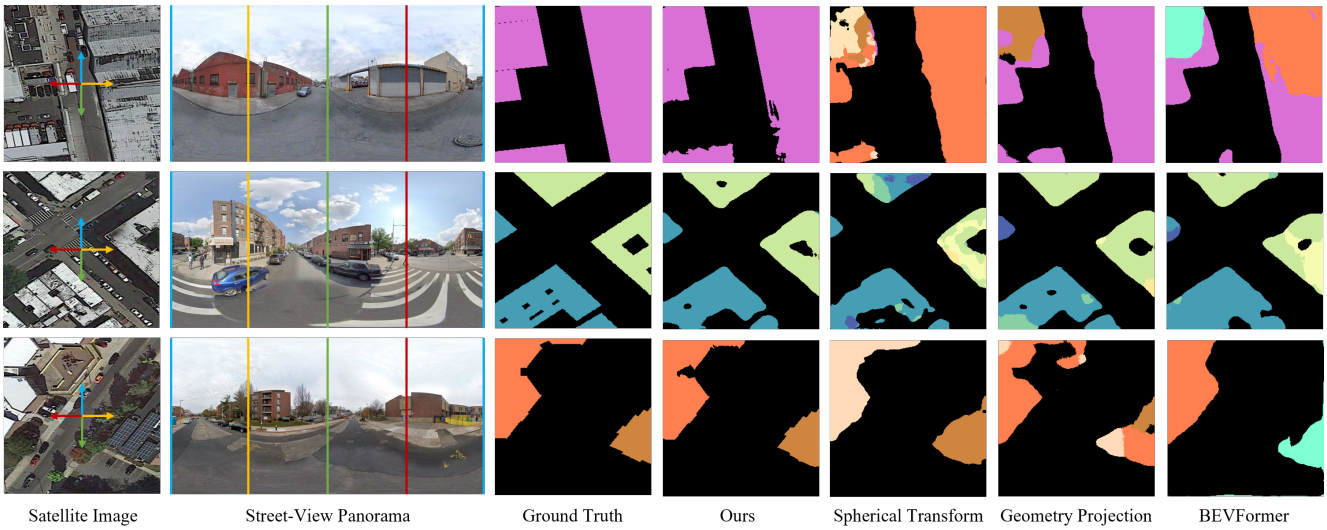

Figure 5. **Comparisons of SG-BEV (Ours) and Other Cross-View Methods for Fine-Grained Segmentation.** The first two rows display results of Brooklyn on land use (first row) and floor level (second row) segmentation tasks. The bottom row illustrates land use segmentation predictions of Boston. The street-view panoramas, from left to right, correspond to a 360-degree clockwise rotation starting from the north direction in the satellite imagery.

Table 2. Comparison with cross-view transformation methods on different datasets, in terms of mIoU and Acc metrics (%).

Method	OmniCity				Brooklyn				Boston		Vigor	
	Land use		Floor		Land use		Floor		Land use		Land use	
	mIOU	Acc	mIOU	Acc	mIOU	Acc	mIOU	Acc	mIOU	Acc	mIOU	Acc
ST [35]	27.07	70.06	18.06	69.11	37.65	79.11	32.72	75.23	27.02	76.95	-	-
GP [29]	33.28	71.98	27.66	74.91	39.70	**79.60**	36.09	77.05	32.43	78.53	-	-
BEVFormer [17]	32.17	71.17	30.81	75.88	42.96	78.32	44.59	76.11	37.16	**78.63**	37.33	76.23
Ours	**37.54**	**76.13**	**40.64**	**77.82**	**47.19**	78.43	**49.51**	**79.00**	**39.72**	77.39	**41.70**	**76.81**

significant improvements over segmentation using satellite imagery alone. In experiments across four datasets, our method showed an increase in mIOU for building category prediction and floor count prediction tasks by 7.61% and 15.16%, respectively, compared to the best satellite-based segmentation method. Predicting floor counts from a satellite view is more challenging than classifying building types, as floor count information is primarily concentrated on the facades. Our approach is more effective in this task, highlighting the efficacy of integrating street-view data for fine-grained building attribute segmentation tasks. As observed in Figure 4, other methods using only satellite imagery could delineate building outlines but struggled with fine-grained attribute perception of buildings. Our method achieves multi-perspective perception of building attributes, not only effectively segmenting building contours but also distinguishing between different building attributes.

Compare to Cross-View methods. As shown in Table 2, we compare our method with other Cross-View methods. Compared to geometric projection methods, our approach showed an average increase in mIOU of 6.35% and 13.20%, respectively. As evident from Figure 5, the two cross-view geometric projection methods maintain good geometric fidelity only near the camera, showing significant distortion in farther areas, leading to loss of street-view feature projection and reduced segmentation performance. Among the methods implemented through geometric projection, Geometry Projection (GP) performed better than Spherical Transform (ST), as it directly projects features instead of converting them into image projections.

Additionally, compared to BEVFormer, our method, with the addition of the SGR module, demonstrates a significant improvement, with an average increase in mIOU of 4.13% and 7.38%. From Figure 5, it is observed that BEVFormer approach may correctly identify the category at the building edges, but not accurately inside the building. This is linked to BEVFormer's limitation in effectively projecting features only to building edges. More visualization results will be shown in the supplementary materials.

Table 3. **Ablation study on Satellite-Guided Reprojection module, in terms of mIOU metrics (%).**

Method	OmniCity		Brooklyn	
	Land use	Floor	Land use	Floor
UNet [25]	27.07	34.13	37.04	39.61
UNet [25] + DGR	32.52	37.01	42.33	45.06
UNet [25]+ SGR	**36.67**	**40.53**	**47.10**	**48.55**
SegNext [9]	32.10	30.56	42.43	43.55
SegNext [9] + DGR	35.49	37.64	45.16	48.17
SegNext [9] + SGR	**37.54**	**40.64**	**47.19**	**49.51**

DGR: Depth-Guided reprojection, utilizes only depth.
SGR: Satellite-Guided reprojection, utilizes both satellite and depth.

Table 4. **Ablation Study on Cross-View Feature Fusion module, in terms of mIOU metrics (%).**

Method	OmniCity		Brooklyn	
	Land use	Floor	Land use	Floor
ConcatFusion	31.78	29.34	42.12	42.28
Cross-View Fusion	**32.10**	**30.56**	**42.43**	**43.55**
DGR + ConcatFusion	33.85	35.13	43.08	46.19
DGR + Cross-View Fusion	**35.49**	**37.64**	**45.16**	**48.17**
SGR + ConcatFusion	36.74	39.69	46.73	49.28
SGR + Cross-View Fusion	**37.54**	**40.64**	**47.19**	**49.51**

4.4. Ablation study

Satellite-Guided Reprojection Performance. In our ablation experiments, we employed UNet [25] and SegNext [9] as image encoders to validate the contribution of our proposed Satellite-Guided Reprojection module (SGR). And in order to more fully verify the role of our SGR, we also extract the Depth-Guided Reprojection (DGR) module in SGR separately for experiments. When using the DGR module, α was set to a fixed value, and satellite footprint range restrictions were not applied. We will compare our method SGR with two different approaches: directly using BEV feature projection, and the second using DGR in the BEV feature projection, while maintaining the same subsequent feature fusion steps. As shown in Table 3, with the DGR module, the mIOU for U-Net improved by 4.77%, and for SegNext by 4.46%. This shows that using depth information effectively disperses street-view features, concentrated on building edges, throughout the building area. The addition of the Satellite-Guidance further increased the mIOU for U-Net by 3.98% and for SegNext by 2.40%, demonstrate the importance of the supporting role of satellite information. Additionally, the performance of our SGR module was significantly enhanced in both SegNext and UNet architectures, convincingly demonstrating the SGR module's outstanding role in optimizing BEV feature distribution, resolving the issue of concentrated building edge features, and substantially improving task performance.

Cross-View Feature Fusion. To demonstrate the effectiveness of our proposed fusion strategy, we conducted a series of ablation experiments on the OmniCity and Brooklyn datasets. We will use a feature fusion method that directly concatenates satellite and BEV features along the channel dimension (ConcatFusion) as the baseline for comparison. Additionally, we further explore the impact of the reprojection module on the feature fusion module. We compared the performance of the Cross-View feature fusion module without using any reprojection module, and when using DGR and SGR modules. As shown in Table 4, the Cross-View feature fusion module improved the performance of both tasks, with the mIOU averaging 0.78%, 2.05% and 0.61% improvements respectively. We found that the most significant improvements occurred when using the DGR module. This is because the offsets generated by the DGR module can be unstable, potentially causing features to exceed the boundaries of buildings, leading to misalignment. Our feature fusion module includes an alignment process that effectively mitigates this issue. However, as SGR is guided by satellite information, the misalignment between different features is less pronounced, leading to a reduced performance improvement from Cross-View fusion. These results validate the effectiveness of our feature fusion module, with more efficient feature integration enhancing task performance under various conditions.

5. Conclusion

In this paper, we proposed SG-BEV, a novel satellite-guided BEV fusion method for cross-view semantic segmentation, focusing on the fine-grained attributes of buildings. Utilizing BEV method coupled with our proposed Satellite-Guided Reprojection module, our method precisely converts features from the street view to satellite view, subsequently merging them with satellite imagery features, producing fine-grained building attribute segmentation results. Our SG-BEV demonstrates significant performance improvements compared to state-of-the-art satellite-based methods and cross-view methods, with an mIOU increase over 10.13% and 5.21% averaged on four datasets. SG-BEV represents a novel attempt at cross-view semantic fusion, achieving a comprehensive understanding of buildings from different perspectives. We hope that our work will inspire further research into cross-view semantic segmentation tasks.

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

References

[1] Favyen Bastani, Songtao He, Sofiane Abbar, Mohammad Alizadeh, Hari Balakrishnan, Sanjay Chawla, Sam Madden, and David DeWitt. Roadtracer: Automatic extraction of road networks from aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4720–4728, 2018. 2

[2] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 4

[3] Hao Chen, Zhenghong Li, Jiangjiang Wu, Wei Xiong, and Chun Du. Semiroadexnet: A semi-supervised network for road extraction from remote sensing imagery via adversarial learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 198:169–183, 2023. 2

[4] Jiankun Chen, Xiaolan Qiu, Chibiao Ding, and Yirong Wu. Cvcmff net: Complex-valued convolutional and multifeature fusion network for building semantic segmentation of insar images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021. 2

[5] Yujia Chen, Guo Zhang, Hao Cui, Xue Li, Shasha Hou, Jinhao Ma, Zhijiang Li, Haifeng Li, and Huabin Wang. A novel weakly supervised semantic segmentation framework to improve the resolution of land cover product. *ISPRS Journal of Photogrammetry and Remote Sensing*, 196:73–92, 2023. 2

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 6

[7] Hao Dong, Xianjing Zhang, Jintao Xu, Rui Ai, Weihao Gu, Huimin Lu, Juho Kannala, and Xieyuanli Chen. SuperFusion: Multilevel LiDAR-Camera Fusion for Long-Range HD Map Generation. *arXiv preprint arXiv:2211.15656*, 2022. 5

[8] Tian Feng, Quang-Trung Truong, Duc Thanh Nguyen, Jing Yu Koh, Lap-Fai Yu, Alexander Binder, and Sai-Kit Yeung. Urban zoning using higher-order markov random fields on multi-view imagery data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 614–630, 2018. 1

[9] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. *Advances in Neural Information Processing Systems*, 35:1140–1156, 2022. 6, 8

[10] Yanqin Jiang, Li Zhang, Zhenwei Miao, Xiatian Zhu, Jin Gao, Weiming Hu, and Yu-Gang Jiang. Polarformer: Multi-camera 3d object detection with polar transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1042–1050, 2023. 2, 3

[11] Ronald Kemker, Carl Salvaggio, and Christopher Kanan. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. *ISPRS journal of photogrammetry and remote sensing*, 145:60–77, 2018. 1

[12] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705, 2019. 2

[13] Hongyang Li, Chonghao Sima, Jifeng Dai, Wenhai Wang, Lewei Lu, Huijie Wang, Enze Xie, Zhiqi Li, Hanming Deng, Hao Tian, et al. Delving into the devils of bird's-eye-view perception: A review, evaluation and recipe. *arXiv preprint arXiv:2209.05324*, 2022. 2, 3

[14] Qi Li, Yue Wang, Yilun Wang, and Hang Zhao. Hdmapnet: An online hd map construction and evaluation framework. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 4628–4634. IEEE, 2022. 3

[15] Weijia Li, Yawen Lai, Linning Xu, Yuanbo Xiangli, Jinhua Yu, Conghui He, Gui-Song Xia, and Dahua Lin. Omnicity: Omnipotent city understanding with multi-level and multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17397–17407, June 2023. 1, 5

[16] Weijia Li, Wenqian Zhao, Jinhua Yu, Juepeng Zheng, Conghui He, Haohuan Fu, and Dahua Lin. Joint semantic–geometric learning for polygonal building segmentation from high-resolution remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 201:26–37, 2023. 1

[17] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Springer, 2022. 2, 3, 5, 6, 7

[18] Tsung-Yi Lin, Serge Belongie, and James Hays. Cross-view image geolocalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2013. 1

[19] Yinhe Liu, Sunan Shi, Junjue Wang, and Yanfei Zhong. Seeing beyond the patch: Scale-adaptive semantic segmentation of high-resolution remote sensing imagery based on reinforcement learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16868–16878, 2023. 1, 2

[20] Xiaohu Lu, Zuoyue Li, Zhaopeng Cui, Martin R Oswald, Marc Pollefeys, and Rongjun Qin. Geometry-aware satellite-to-ground image synthesis for urban areas. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 859–867, 2020. 2

[21] Bowen Pan, Jiankai Sun, Ho Yin Tiga Leung, Alex Andonian, and Bolei Zhou. Cross-view semantic segmentation for sensing surroundings. *IEEE Robotics and Automation Letters*, 5(3):4867–4873, 2020. 2, 3

[22] Lang Peng, Zhirong Chen, Zhangjie Fu, Pengpeng Liang, and Erkang Cheng. Bevsegformer: Bird's eye view semantic segmentation from arbitrary camera rigs. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5935–5943, 2023. 3

[23] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unpro-

jecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 194–210. Springer, 2020. 2, 3

[24] Lennart Reiher, Bastian Lampe, and Lutz Eckstein. 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)*, pages 1–7. IEEE, 2020. 2, 3

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 8

[26] Franz Rottensteiner, Gunho Sohn, Jaewook Jung, Markus Gerke, Caroline Baillard, Sebastien Benitez, and Uwe Breitkopf. The isprs benchmark on urban object classification and 3d building reconstruction. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences; I-3*, 1(1):293–298, 2012. 1

[27] Yujiao Shi and Hongdong Li. 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*, pages 17010–17020, 2022. 2

[28] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[29] Yujiao Shi, Fei Wu, Akhil Perincherry, Ankit Vora, and Hongdong Li. 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*, pages 21516–21526, 2023. 2, 3, 6, 7

[30] Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. 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*, pages 4064–4072, 2020. 2

[31] Aniruddh Sikdar, Sumanth Udupa, Prajwal Gurunath, and Suresh Sundaram. Deepmao: Deep multi-scale aware overcomplete network for building segmentation in satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 487–496, 2023. 2

[32] Shivangi Srivastava, John E Vargas-Munoz, and Devis Tuia. Understanding urban landuse from the above and ground perspectives: A deep learning, multimodal solution. *Remote sensing of environment*, 228:129–143, 2019. 1

[33] Zhifeng Teng, Jiaming Zhang, Kailun Yang, Kunyu Peng, Hao Shi, Simon Reiß, Ke Cao, and Rainer Stiefelhagen. 360bev: Panoramic semantic mapping for indoor bird's-eye view. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024. 2, 4

[34] Aysim Toker, Qunjie Zhou, Maxim Maximov, and Laura Leal-Taixé. 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*, pages 6488–6497, 2021. 2

[35] Xiaolong Wang, Runsen Xu, Zhuofan Cui, Zeyu Wan, and Yu Zhang. Fine-grained cross-view geo-localization using a correlation-aware homography estimator. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2, 3, 6, 7

[36] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pages 180–191. PMLR, 2022. 3, 5

[37] Zbigniew Wojna, Krzysztof Maziarz, Łukasz Jocz, Robert Pałuba, Robert Kozikowski, and Iason Kokkinos. Holistic multi-view building analysis in the wild with projection pooling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2870–2878, 2021. 2

[38] Scott Workman, M Usman Rafique, Hunter Blanton, and Nathan Jacobs. Revisiting near/remote sensing with geospatial attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1778–1787, 2022. 1, 2

[39] Scott Workman, Menghua Zhai, David J Crandall, and Nathan Jacobs. A unified model for near and remote sensing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2688–2697, 2017. 1, 2

[40] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2945–2954, 2023. 6

[41] Rongtao Xu, Changwei Wang, Jiguang Zhang, Shibiao Xu, Weiliang Meng, and Xiaopeng Zhang. Rssformer: Foreground saliency enhancement for remote sensing land-cover segmentation. *IEEE Transactions on Image Processing*, 32:1052–1064, 2023. 2

[42] Jiayu Yang, Enze Xie, Miaomiao Liu, and Jose M Alvarez. Parametric depth based feature representation learning for object detection and segmentation in bird's-eye view. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8483–8492, 2023. 2

[43] Zhuo Zheng, Yanfei Zhong, Junjue Wang, and Ailong Ma. Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4096–4105, 2020. 1, 2, 6

[44] Sijie Zhu, Mubarak Shah, and Chen Chen. Transgeo: Transformer is all you need for cross-view image geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1162–1171, 2022. 2

[45] Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3640–3649, 2021. 5