The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLIX-B2-2026
XXV ISPRS Congress 2026 “From Imagery to Understanding”, Commission II, 4–11 July 2026, Toronto, Canada

BEV-LOC: Real-Time and Lightweight Cross-View Localization via Online BEV Mapping

Jiyong Kwag, Charles Toth, Alper Yilmaz

Dept. of Civil Engineering, The Ohio State University, 281 W Lane Ave, Columbus, Ohio
{kwag.3, toth.2, yilmaz.15}@osu.edu

Keywords: HD map generation, cross-view geo-localization, computer vision, deep learning

Abstract

This paper presents a deep learning and classical computer vision framework for cross-view geolocalization using 360-degree multi-perspective view (PV) images and an offline global map. Recent studies on cross-view geolocalization typically rely on deep learning models to localize panoramic PV images by matching them with reference satellite imagery. However, such approaches face practical limitations in real-world deployments, due to their dependence on large-scale GPU resources and the need to store extensive satellite image datasets. To address these challenges, we propose BEV-LOC, a lightweight and real-time cross-view geolocalization method. BEV-LOC employs Bird’s Eye View (BEV) encoder that learns to transform 360-degree multi-PV images into a local high-definition (HD) BEV map. The localization is then performed using Intersection Over Union (IoU)-based template matching with an offline global map. Our architecture achieves real-time performance at 30 FPS without the need for high-end GPU hardware and delivers a high positioning accuracy of 1.2 meters.

1. Introduction

Determining geographic location remains a challenging problem, particularly for autonomous driving applications. Accurate location is critical for vehicle systems that navigate dense urban environments filled with dynamic elements such as vehicles and pedestrians. Although real-time kinematic (RTK) GPS offers centimeter-level precision, it is expensive and prone to signal degradation or spoofing in urban canyons. Infrastructure-based localization provides a potential alternative, but it often requires pre-installed devices and generally delivers lower accuracy than GPS, depending on the signal modality used.

As a cost-effective solution, PV image-based cross-view localization has emerged as a promising approach, especially in structured and repetitive environments. Recent advances in deep learning have demonstrated impressive performance in cross-view localization in large geographic regions. These methods typically train models to match a query PV image with the most similar satellite image to estimate its geographic location (Zhu et al., 2021); (Deuser et al., 2023) . However, real-world deployment of such models presents significant challenges: they require high-end GPUs for training and inference, and maintaining a large database of satellite imagery introduces substantial memory overhead. Since wider satellite coverage improves localization accuracy, these methods inherently demand considerable storage resources.

To address these limitations, we propose an alternative solution: matching a memory-efficient offline global High Definition (HD) map with a local HD map generated from multi-PV images. HD map is highly accurate digital representations of road environments, recording detailed elements such as roads, lane markings, and pedestrian crossings (Kwag and Toth, 2024). HD map is available in various formats, including the PV segmentation map (Perera et al., 2024) (Kwag et al., 2025), the BEV segmentation map (Li et al., 2024); (Chen et al., 2022) and the vectorized format (Liao et al., 2023); (Liu et al., 2023a). Crucially, HD maps are significantly more memory efficient than raw satellite images because they encode only essential road features in a sparse and structured format, drastically reducing the data required per unit area.

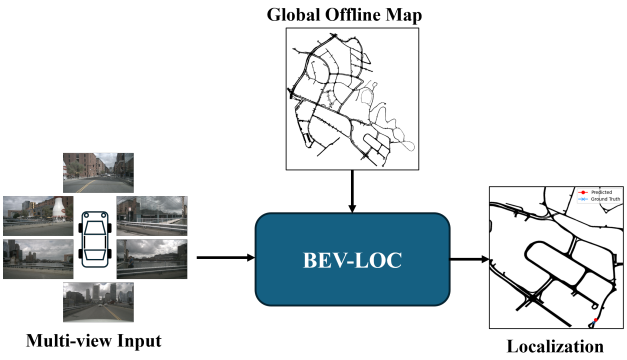

Figure 1. **Overview of BEV-LOC.** BEV-LOC leverages multi-view PV images from the vehicle and a pre-defined global offline HD map for lightweight cross-view localization using IoU-based template matching, without the need for high-end GPUs.

In this paper, we introduce BEV-LOC, a hybrid framework that combines a lightweight deep learning model with classical computer vision techniques. BEV-LOC generates local HD maps from multi-PV images and matches them to a preconstructed, memory-efficient offline global HD map using IoU-based template matching. Our system is designed to operate efficiently without reliance on high-end GPUs or specialized hardware, enabling accurate and real-time localization. In experiments carried out in urban Singapore, BEV-LOC achieved a localization accuracy of 1.2 meters while maintaining real-time performance at 30 FPS. Key contributions of our work include:

- A lightweight, real-time online HD map generation model that operates without the need for high-end GPUs.
- A simple IoU-based template matching algorithm between offline global maps and online local HD maps, enabling efficient cross-view localization.

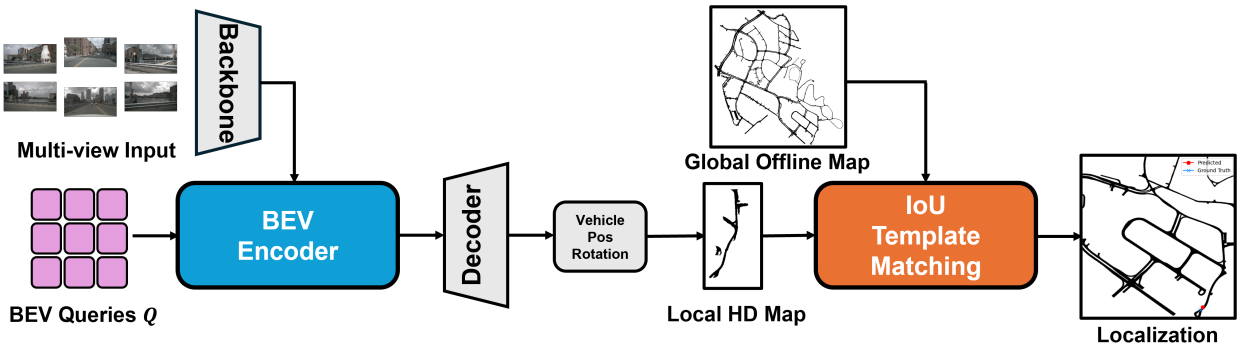

Figure 2. **BEV-LOC Architecture.** BEV-LOC consists of two main modules: BEV encoder and IoU-based Template Matching. BEV encoder transforms multi-PV images into a local HD map, This local HD map is then rotated to align with a preconstructed offline global map. Finally, IoU-based template matching is performed to identify the most probable location on the global map.

2. Related Work

2.1 BEV Map Generation.

BEV encoding represents multi-view images within a unified 2D BEV feature space, enabling consistent spatial reasoning across multiple camera perspectives (Liu et al., 2023b); (Huang et al., 2021); (Li et al., 2023); (Liao et al., 2023). This representation has become a fundamental component in autonomous driving perception systems because it facilitates the integration of information from different viewpoints into a common spatial frame.

Early approaches such as Lift-Splat-Shoot (LSS) and BEVFusion introduced depth-based frameworks to generate BEV representations from multi-camera images (Philion and Fidler, 2020); (Liu et al., 2023b). These methods estimate depth distributions for each image pixel and project the features into a BEV space, enabling tasks such as BEV map segmentation and object detection. More recent works, including BEVFormer and Cross-View Transformer (CVT), further advanced this paradigm by introducing attention-based architectures in a camera-only setting (Li et al., 2024); (Zhou and Krähenbühl, 2022). These transformer-based methods leverage spatial and temporal attention mechanisms to aggregate information across camera views and time steps, allowing the network to learn more expressive BEV representations without requiring explicit depth supervision.

A key advantage of BEV representations is that they inherently compress the vertical axis, projecting the 3D environment into a planar representation aligned with the road surface. This property significantly simplifies spatial reasoning and makes 2D BEV map generation computationally efficient, which is particularly beneficial for downstream tasks such as mapping, motion planning, and localization.

2.2 Retrieval Based Cross-View Localization.

Cross-view localization addresses the problem of localizing street-level images by matching them with GPS-referenced satellite imagery (Lin et al., 2013); (Workman et al., 2015) (Liu and Li, 2019). This task is particularly challenging due to the significant differences in viewpoint between aerial and street-level perspectives, scale, and environmental conditions (Hu et al., 2018); (Liu and Li, 2019); (Zhu et al., 2021) . In recent years, deep learning–based approaches have shown significant progress in tackling these challenges using the image retrieval feature matching paradigm. In this paradigm, satellite images and street-view images are mapped into a shared embedding space using deep neural networks. The goal is to learn feature representations in which images corresponding to the same geographic location are positioned close to each other in the embedding space, whereas images from different locations are pushed farther apart.

Early approaches (Workman et al., 2015) relied primarily on convolutional neural networks (CNNs) to extract visual features from both modalities. However, recent research (Zhu et al., 2022) has increasingly adopted transformer-based (Dosovitskiy et al., 2021) architectures to capture global contextual information. These architectures improve the robustness of cross-view feature representations, enabling more reliable matching between aerial and ground-level imagery. A notable advancement in contrastive learning for cross-view localization is the introduction of hard negative mining strategies. For example, Sample4Geo (Deuser et al., 2023) proposes an effective training strategy that dynamically selects hard negative samples based on visual similarity between image embeddings using ConvNeXt (Liu et al., 2022). Instead of randomly sampling negative pairs, the model identifies negatives that are visually similar to the query but correspond to different locations. By focusing the training process on these challenging samples, the network learns more discriminative representations and improves retrieval accuracy.

Despite the impressive performance achieved by retrieval-based cross-view localization methods, several limitations remain. One major drawback is the computational and memory cost associated with large-scale feature matching. During inference, the system must compare the feature vector of a query street-view image against a large database of satellite image embeddings using similarity metrics such as cosine similarity. As the size of the satellite image database grows to cover large geographic areas, the number of similarity computations required increases significantly. This results in substantial memory consumption for storing feature embeddings and increased computational overhead during retrieval.

3. Method

3.1 BEV-LOC

Figure 1 illustrates an overview of BEV-LOC. BEV-LOC consists of two main components: BEV encoder and IoU-based

Template Matching module. BEV encoder transforms multi-PV image features into a unified BEV representation, capturing the surrounding environment in a consistent spatial format. The convolutional decoder then generates a semantic segmentation of drivable areas, resulting in a local HD map. Since the generated local HD map is in the ego-vehicle coordinate system, it is rotated to align with the preconstructed offline global map using IMU and odometry information. IoU-based Template Matching module then measures the similarity between the local HD map and the global map to accurately estimate the most probable vehicle location. Figure 2 and 4 illustrate BEV-LOC and BEV encoder pipeline.

3.1.1 BEV Encoder BEV encoder (Figure 4) in BEV-LOC follows a structure similar to that of BEVFormer (Li et al., 2024), but with key architectural modifications customized to our goals of real-time performance and lightweight design. The encoder is composed of three main components: BEV queries, global cross-attention layer, and 3×3 convolution layer. Unlike conventional Transformer architecture (Vaswani et al., 2017, Dosovitskiy et al., 2021), we adopt a simplified architecture that eliminates self-attention and feed-forward networks (FFNs), replacing them with a 3×3 convolution layer to significantly reduce computational overhead and model complexity.

BEV queries are grid-shaped learnable parameters that serve as anchors in the BEV space. These queries interact with multiscale features extracted from multi-PV images via a global cross-attention mechanism, enabling the model to aggregate spatial context from the surrounding environment into a unified BEV representation. Global cross-attention enables each BEV query to attend to features across different camera views, effectively capturing the spatial relationship between the ego-vehicle and its surroundings. Many existing methods (Li et al., 2024, Chen et al., 2022) leverage deformable attention and camera projection matrices (intrinsic and extrinsic) to reduce computation efficiency. Since we operate on low-resolution images for runtime efficiency, we found that incorporating deformable attention offered negligible benefits but degraded performance. Therefore, we opt to use the original global cross-attention mechanism directly between the BEV queries and multi-PV image features, achieving a favorable trade-off between performance and speed.

3.1.2 IoU Template Matching After generating the local HD map using the BEV encoder and ego-vehicle pose rotation, IoU-based template matching is performed to estimate the most likely location of the vehicle. However, a key challenge arises due to the large size of the global offline HD map: scanning every pixel location becomes computationally inefficient. Thus, we adopt two different methodologies: 2D Fast Fourier Transform (FFT) based scanning and previous observation-based scanning.

Given a binary global map $G \in \{0,1\}^{H\times W}$ and a binary local map $T \in \{0,1\}^{h\times w}$, the IoU at each valid position (u,v) is computed as follows. Assume that we define FFT-based convolution as:

$$(f * g)(u,v) = \mathcal{F}^{-1}\left(\mathcal{F}(f)\cdot\mathcal{F}(g)\right)(u,v) \tag{1}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the 2D FFT and 2D inverse FFT, respectively. First, the local map T is flipped spatially for convolution:

$$T' = flip(T) \tag{2}$$

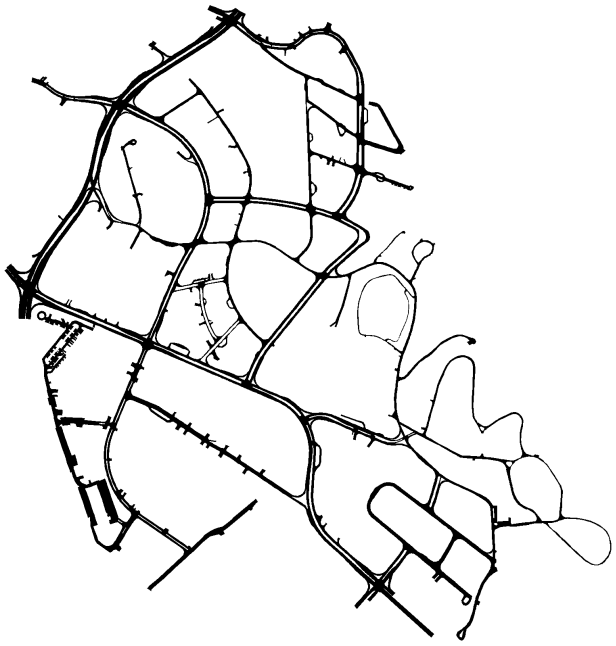

Figure 3. **Test Region HD Map.** The global offline map of the Singapore-onenorth region used for IoU-based template matching.

The intersection map is computed using FFT-based convolution:

$$I(u,v) = (G * T')(u,v) \tag{3}$$

The total number of foreground (1-valued) pixels in the template is:

$$A_T = \sum_{i,j} T[i,j] \tag{4}$$

The number of foreground pixels in each $h \times w$ window of the global map is computed via FFT-based convolution with an all-ones kernel:

$$A_G(u,v) = (G * \mathbf{1}_{h\times w})(u,v) \tag{5}$$

Finally, union is calculated using the inclusion-exclusion principle:

$$U(u,v) = A_T + A_G(u,v) - I(u,v) \tag{6}$$

Thus, IoU at each position is given by:

$$\text{IoU}(u,v) = \begin{cases} \frac{I(u,v)}{U(u,v)}, & \text{if } U(u,v) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

In the first stage, scanning the local map against the global map from scratch incurs significant computational cost, especially given the large size of the offline global map, with $20{,}250 \times 15{,}856$ pixels. Thus, we apply the 2D FFT to the global and local maps, leveraging the fact that convolution in the spatial domain becomes pointwise multiplication in the frequency domain. Typically, spatial domain convolution has a time complexity of $\mathcal{O}(N^2k^2)$, where $N \times N$ is the image size and $k \times k$ is the kernel size. By performing FFT-based pointwise multiplication, the run-time complexity is reduced to $\mathcal{O}(N^2 \log N)$, followed by inverse FFT to transform back to the spatial domain and identify the location with the highest response.

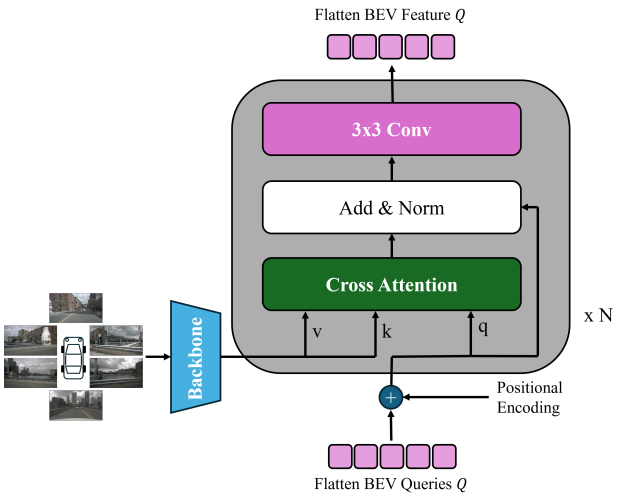

Figure 4. **Architecture of the BEV encoder.** The encoder comprises three components: BEV queries, cross-attention, and a 3×3 convolution layer, designed to efficiently capture features from multi-PV images while maintaining real-time performance.

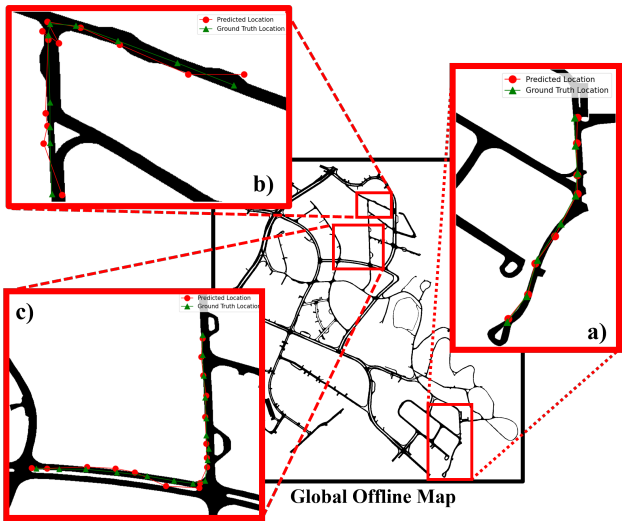

Figure 5. **Architecture of the BEV encoder.** The encoder comprises three components: BEV queries, cross-attention, and a 3×3 convolution layer, designed to efficiently capture features from multi-PV images while maintaining real-time performance.

After identifying the initial localization (first hit), we switch to a previous-observation-based scanning method instead of continuing with FFT-based scanning. Although FFT-based convolution is advantageous for large global maps, it becomes less efficient for smaller or localized regions. The observation-based scanning method significantly reduces the search space by focusing only on the area close to the vehicle's previous estimated location. This motion-based search space refinement takes advantage of the continuity of vehicle movement, assuming that the vehicle's current position is near its last known location. As a result, it further narrows the search area and improves runtime. To enhance stability and reliability, we also incorporate a geospatial filter that restricts geolocalization to drivable areas. This is implemented by assigning higher weights to drivable regions during IoU score calculation, thereby favoring plausible vehicle positions and reducing false matches in non-drivable regions.

3.2 Dataset

We evaluated BEV-LOC using the widely adopted nuScenes dataset (Caesar et al., 2020), which contains 1,000 unique driving scenes, each 20 seconds long, collected from two urban environments: Singapore and Boston. Each scene includes RGB images from six vehicle-mounted cameras, providing a full 360-degree horizontal field of view around the ego-vehicle. In addition to image data, nuScenes provides a human-annotated high-resolution (0.1 m / pixel) semantic map that includes road elements such as drivable areas, sidewalks, and pedestrian crossings. The semantic map is projected using the WGS 84 Web Mercator coordinate system. For our experiments, we focus on the Singapore-onenorth region, which spans approximately 20,250 m × 15,856 m. To create a more challenging evaluation scenario, we restrict our assessment to the drivable area annotations only.

3.3 Implementation Detail

BEV-LOC processes multi-PV images of resolution $256 \times 256 \times 3$ using a MobileViT-XXS (Mehta and Rastegari, 2021) backbone pretrained on ImageNet-1K (Russakovsky et al., 2015) to extract multi-scale features. Feature extraction is enhanced by an additional bottleneck layer composed of a MobileNetV2 (Sandler et al., 2018) block followed by a MobileViT block. The learnable BEV queries of shape $25 \times 25 \times 256$ interact with the backbone features using global cross-attention. A 3-stage decoder with bilinear upsampling and 3×3 convolutions reconstructs the BEV map to 200×200. For localization, a global FFT-based scan is followed by a local observation-based search within a 20 m radius using a 5 m stride for efficient refinement.

4. Result

Figure 5 illustrates the visualization of the cross-view geolocalization results from BEV-LOC on a large offline global map of the Singapore-onenorth region (Figure 3), demonstrating real-time performance at 30 FPS. The predicted vehicle locations closely follow the road network, indicating the accuracy and efficiency of the proposed method. Based on quantitative evaluation, our approach achieves a mean Euclidean distance error of 1.2 meters, with an across-track mean error of 0.76 meters and an along-track mean error of 0.86 meters. A typical passenger vehicle has a length of 4 to 5 meters and a width of 1.8 meters, which means that an error of 1.2 meters is well within the physical size of the vehicle. These results underscore the robustness of BEV-LOC in achieving precise localization across wide-area global map environments.

Figure 6 presents the temporal evolution of the Euclidean distance error, across-track error, and along-track error, corresponding to Figure 6 (a), (b), and (c), shown from top to bottom. In the across-track error plot, positive values indicate that the predicted vehicle location lies to the right of the ground-truth trajectory, while negative values indicate that the prediction lies to the left. Similarly, in the along-track error plot, positive values indicate that the predicted position is ahead of the ground truth along the direction of travel, whereas negative values indicate that it is behind.

For the scenario shown in Figure 6(a), a noticeable spike in the distance error occurs at time step 3. This spike is primarily

caused by inaccuracies in the local HD map generation produced by the BEV encoder. Despite this temporary deviation, the algorithm is able to recover quickly due to the sufficiently large search radius used in the previous observation-based localization step. This recovery behavior demonstrates the robustness of the localization framework against occasional perception errors. Furthermore, this observation suggests that scaling up the capacity of the BEV encoder could potentially improve the accuracy of local HD map generation, which would in turn enhance localization performance. However, such improvements would likely come at the expense of increased computational cost and runtime.

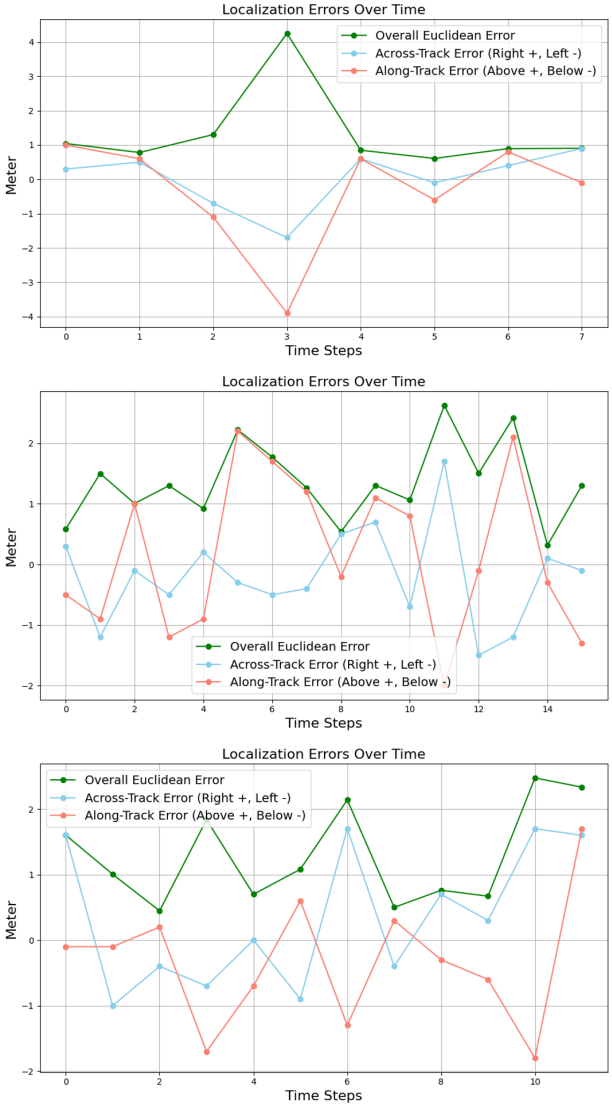

Figure 6. **Different distance error.** (Green) Euclidean distance error over time. (Blue) Across-track error over time — positive values indicate that the predicted location is to the right of the ground truth, while negative values indicate it is to the left. (Red) Along-track error over time — positive values indicate that the predicted location is ahead of the ground truth, while negative values indicate it is behind.

In contrast, the scenarios illustrated in Figures 6(b) and 6(c) present more complex environments, particularly around intersection regions where the road structure is more ambiguous. In these cases, the previous observation-based method does not significantly reduce the localization error. The increased environmental complexity introduces additional challenges for reliable feature matching and map alignment, which limits the ability of the algorithm to refine the predicted position. This highlights the inherent difficulty of cross-view localization in dense urban environments with complex road geometries and suggests potential directions for further improving the robustness of the localization system.

Acknowledgments

This work was supported in part by the U.S. Department of Transportation under Grant 69A3552348327 for the CARMEN+ University Transportation Center.

5. Conclusion

In this work, we introduced BEV-LOC, a lightweight and real-time cross-view geolocalization framework that integrates deep learning with IoU-based template matching. By projecting multi-perspective images into a local online HD map and aligning it with a global offline map, BEV-LOC enables accurate vehicle localization without relying on GNSS signals or satellite imagery. This design allows the system to operate robustly even in environments where GNSS signals are unreliable or unavailable. Experimental results demonstrate that the proposed approach achieves precise localization while maintaining real-time performance, making it suitable for deployment in practical autonomous driving systems and resource-constrained platforms, such as embedded automotive systems and edge devices.

For future work, we plan to investigate fully end-to-end learning architectures that directly learn feature-space similarity for map matching. Such approaches may further improve localization accuracy and generalization across diverse environments. Additionally, exploring more advanced BEV representation learning and robust feature matching strategies could further enhance the scalability and reliability of cross-view localization systems.

References

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O., 2020. nuscenes: A multimodal dataset for autonomous driving. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11621–11631.

Chen, S., Cheng, T., Wang, X., Meng, W., Zhang, Q., Liu, W., 2022. Efficient and robust 2d-to-bev representation learning via geometry-guided kernel transformer. *arXiv preprint arXiv:2206.04584*.

Deuser, F., Habel, K., Oswald, N., 2023. Sample4geo: Hard negative sampling for cross-view geo-localisation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16847–16856.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR*.

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. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7258–7267.

Huang, J., Huang, G., Zhu, Z., Yun, Y., Du, D., 2021. BEVDet: High-performance Multi-camera 3D Object Detection in Bird-Eye-View. *arXiv preprint arXiv:2112.11790*.

Kwag, J., Toth, C., 2024. A Review on End-to-End High-Definition Map Generation. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48, 187–194.

Kwag, J., Yilmaz, A., Toth, C., 2025. Lightweight road environment segmentation using vector quantization.

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. *Proceedings of the AAAI conference on artificial intelligence*, 37number 2, 1477–1485.

Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Yu, Q., Dai, J., 2024. Bevformer: learning bird’s-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liao, B., Chen, S., Wang, X., Cheng, T., Zhang, Q., Liu, W., Huang, C., 2023. Maptr: Structured modeling and learning for online vectorized hd map construction. *International Conference on Learning Representations*.

Lin, T.-Y., Belongie, S., Hays, J., 2013. Cross-view image geo-localization. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 891–898.

Liu, L., Li, H., 2019. Lending orientation to neural networks for cross-view geo-localization. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, Y., Yuan, T., Wang, Y., Wang, Y., Zhao, H., 2023a. Vectormapnet: End-to-end vectorized hd map learning. *International Conference on Machine Learning*, PMLR, 22352–22369.

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

Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D., Han, S., 2023b. Bevfusion: Multi-task multi-sensor fusion with unified bird’s-eye view representation. *IEEE International Conference on Robotics and Automation (ICRA)*.

Mehta, S., Rastegari, M., 2021. MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer. *arXiv preprint arXiv:2110.02178*.

Perera, S., Erzurumlu, Y., Gulati, D., Yilmaz, A., 2024. Mobileunetr: A lightweight end-to-end hybrid vision transformer for efficient medical image segmentation. *Proceedings of the IEEE/CVF European Conference on Computer Vision (ECCV)*, TBD.

Philion, J., Fidler, S., 2020. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. *European conference on computer vision*, Springer, 194–210.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. et al., 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3), 211–252.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C., 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4510–4520.

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.

Workman, S., Souvenir, R., Jacobs, N., 2015. Wide-area image geolocalization with aerial reference imagery. *Proceedings of the IEEE International Conference on Computer Vision*, 3961–3969.

Zhou, B., Krähenbühl, P., 2022. Cross-view transformers for real-time map-view semantic segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13760–13769.

Zhu, S., Shah, M., Chen, C., 2022. Transgeo: Transformer is all you need for cross-view image geo-localization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1162–1171.

Zhu, S., Yang, T., Chen, C., 2021. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3640–3649.