2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

TransLocNet: Cross-Modal Attention for Aerial-Ground Vehicle Localization with Contrastive Learning

Phu Pham[1], Damon Conover[2], Aniket Bera[1]
[1]Department of Computer Science, Purdue University [2]DEVCOM Army Research Laboratory
{phupham, aniketbera}@purdue.edu, damon.m.conover.civ@army.mil

***Abstract*— Aerial–ground localization is difficult due to large viewpoint and modality gaps between ground-level LiDAR and overhead imagery. We propose TransLocNet, a cross-modal attention framework that fuses LiDAR geometry with aerial semantic context. LiDAR scans are projected into a bird's-eye-view representation and aligned with aerial features through bidirectional attention, followed by a likelihood map decoder that outputs spatial probability distributions over position and orientation. A contrastive learning module enforces a shared embedding space to improve cross-modal alignment. Experiments on CARLA and KITTI show that TransLocNet outperforms state-of-the-art baselines, reducing localization error by up to 63% and achieving sub-meter, sub-degree accuracy. These results demonstrate that TransLocNet provides robust and generalizable aerial–ground localization in both synthetic and real-world settings.**

I. INTRODUCTION

Vehicle localization is essential for autonomous navigation, advanced driver assistance systems, and mobile robotics [1], [2]. Conventional approaches often rely on Global Navigation Satellite System (GNSS)-based methods such as GPS/INS. However, these solutions degrade in GNSS-denied environments like urban areas, tunnels, and dense forests, where occlusion, multipath, and jamming can introduce multi-meter errors [3]. To address these issues, researchers have explored sensor-based alternatives such as visual odometry and Simultaneous Localization and Mapping (SLAM). These methods provide relative pose estimates but accumulate drift over long distances without global references [4], [5]. High-definition (HD) maps offer absolute localization by matching sensor data to compact semantic representations [6], [7], but their construction and maintenance at city scale are costly, requiring frequent re-collection and annotation.

Overhead imagery has emerged as an alternative global map source [8], [9]. Cross-view matching aligns ground-level sensor observations with aerial views to infer ego-vehicle pose. CVM-Net [10] introduced cross-view matching networks for viewpoint-invariant descriptors, while Highly-Accurate [8] formulated localization as direct pose estimation using CNNs and geometric projection. Transformer-based approaches [11] improved robustness by modeling spatial relationships across views.

Most existing methods focus on visual data and underutilize LiDAR, which provides dense 3D structure resilient to illumination and weather variations [12]. LiDAR complements aerial imagery with ground-level geometry, but prior fusion strategies typically rely on feature concatenation, which fails to capture spatial correspondences. They also lack robustness to large viewpoint shifts, scale discrepancies, and temporal misalignment. Attention mechanisms and contrastive learning have been explored independently for cross-modal alignment [13], [14], [15], but have not been jointly applied to aerial–ground localization.

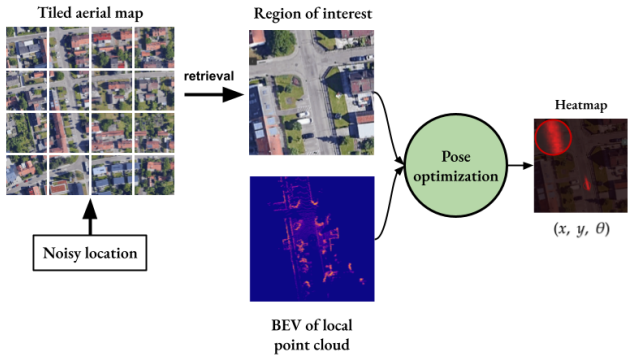

Fig. 1: Overview of our aerial-ground vehicle localization framework. Starting from a noisy initial position, a candidate region is retrieved from a tiled aerial map. The local LiDAR BEV projection is then jointly optimized with the aerial image patch to refine position (x, y) and orientation θ.

Despite these advancements, existing approaches face limitations in handling large-scale viewpoint shifts, modality discrepancies, and noisy real-world data, often relying on simplistic fusion strategies that fail to capture intricate cross-modal interactions. To address these gaps, we propose TransLocNet, a novel framework for aerial-ground vehicle localization that integrates cross-modal attention mechanisms with contrastive learning. By projecting local LiDAR point clouds to a top-view representation and aligning them with aerial images, TransLocNet employs attention-based fusion to model inter-modal dependencies and a contrastive loss to enforce semantic alignment, achieving robust pose estimation. An overview of our proposed framework is illustrated in Fig. 1.

The main contributions of this work are threefold:

- We propose a cross-modal attention module that dynamically attends to salient features across bird's-eye-view (BEV) point clouds and aerial imagery, enabling robust feature fusion and improved localization accuracy under large viewpoint variations.
- We integrate a contrastive learning objective that opti-

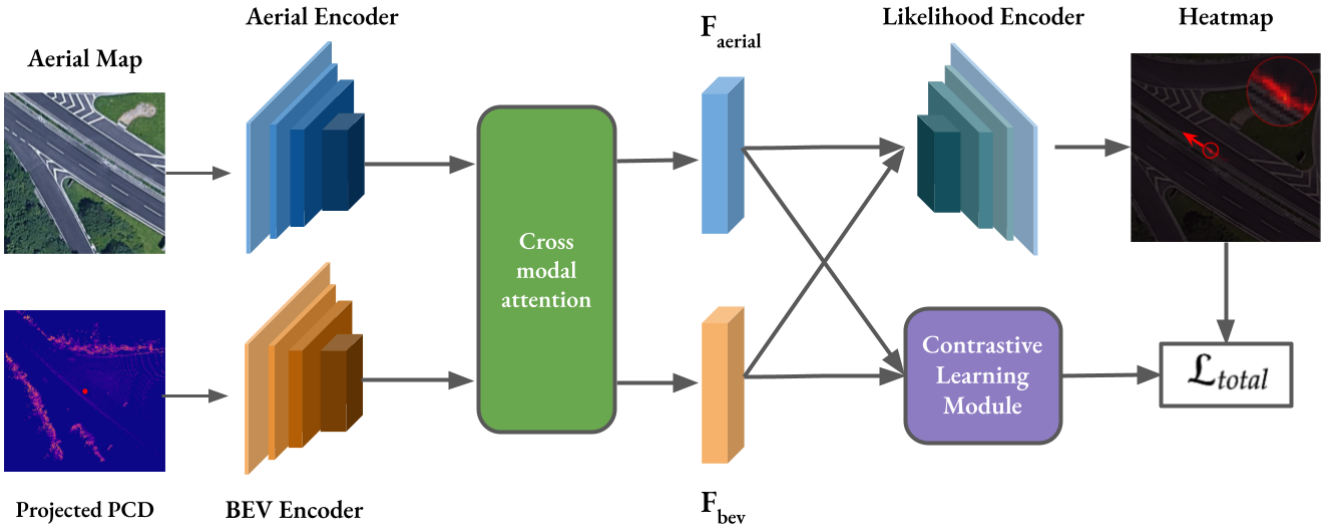

Fig. 2: **Overview:** TransLocNet localizes aerial and ground vehicles by encoding aerial images and LiDAR BEV projections into respective feature maps. A cross-modal attention aligns features, which are fused and decoded into likelihood maps for uncertain position and orientation estimates. Contrastive learning improves alignment by bringing matched aerial-BEV pairs closer and pushing mismatched pairs apart.

mizes feature embeddings across modalities, strengthening retrieval and refining metric pose estimation.

- We provide extensive empirical validation on both synthetic (CARLA [16]) and real-world (KITTI [17]) benchmarks, where our model achieves better performance, reducing location error by up to **63%** and orientation error by up to **81%** compared to prior methods.

II. RELATED WORK

A. Map-based localization without GNSS

Classical autonomous driving systems localize against prebuilt maps such as high-definition (HD) LiDAR reflectivity, semantic lane maps, or OpenStreetMap (OSM), typically using particle-filter or scan-matching pipelines. Landmark examples include HD-map particle filtering for urban driving, as well as a large body of Monte-Carlo/box-particle formulations that incorporate map-based priors [18]. When HD maps are not available, lightweight OSM data can be fused with onboard sensors to help a robot recover its global position [19], [20]. This is often achieved with learned image-to-map embeddings or constrained particle filtering.

B. SLAM-based localization

Simultaneous Localization and Mapping (SLAM) supports real-time map construction and pose estimation through onboard sensors. Classical visual SLAM pipelines [21], [22], [23] are well-established, while recent 3D Gaussian Splatting (3DGS) [24] methods [25], [26], [27], [28] enable real-time mapping and high-quality rendering, though primarily for indoor environments.

SLAM systems suffer from drift accumulation and require consecutive frames for tracking, making global localization difficult without GPS or global priors. This motivates cross-modal localization, where aerial imagery serves as a valuable complement to traditional SLAM.

C. Cross-view localization with aerial and ground data

Cross-view localization matches ground-level images with geo-referenced aerial databases under extreme viewpoint differences. CVM-Net [10] introduced deep Siamese networks with NetVLAD [29] descriptors and metric learning, while later works added rotation awareness and improved local feature matching [30], [31]. Recent methods integrate spatially aware correlation and multi-scale descriptors [32] for improved robustness.

Beyond image-only approaches, several works align LiDAR or BEV representations with aerial maps. BEVPlace++ [33] uses BEV images and CNNs with rotation equivariance for 3-DoF pose estimation. AGL-Net [16] aligns LiDAR point clouds with aerial maps using a two-stage matching network with scale and skeleton loss for robust, scale-invariant pose estimation.

D. Cross-modal attention and BEV-centric fusion

Transformer-based cross-attention [34] has become standard for fusing heterogeneous views into BEV representations. BEVFormer [35] learns BEV queries with spatial cross-attention to camera features and temporal self-attention, while Lift-Splat-Shoot [36] pioneered differentiable lifting of perspective images into BEV. BEVFusion [37] extends this paradigm by fusing multi-modal features from cameras and LiDAR in a unified BEV space, while MapTR [38] focuses on online mapping and localization using BEV representations. End-to-end driving systems such as TransFuser [39] couple LiDAR and camera streams with global attention.

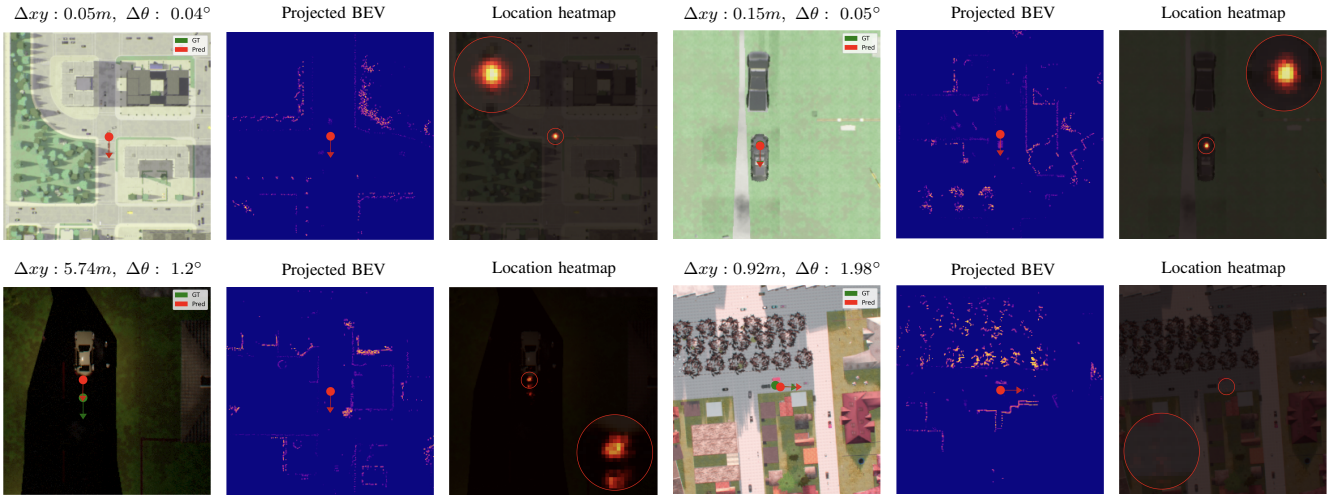

Fig. 3: Qualitative localization results on the CARLA dataset. Each example shows an aerial image with predicted pose, the projected BEV, and a heatmap of vehicle position likelihood. The synthetic environment provides aligned aerial-ground pairs for precise evaluation, achieving consistent sub-meter and sub-degree accuracy in various conditions.

III. METHOD

We propose TransLocNet, a cross-modal attention framework for aerial-ground vehicle localization that integrates LiDAR point clouds with aerial imagery through feature extraction, cross-modal attention, and contrastive learning.

A. Problem Formulation

Given a local LiDAR point cloud $\mathcal{P} = \{p_i\}_{i=1}^{N}$ where $p_i \in \mathbb{R}^3$ represents 3D coordinates, and a corresponding aerial image $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$, our goal is to estimate the vehicle's pose $\mathbf{T} = (x, y, \theta)$ in the aerial map coordinate system, where (x, y) denotes the 2D position and θ represents the heading angle.

The key challenge lies in bridging the viewpoint gap between ground-level LiDAR observations and overhead aerial imagery. We project the LiDAR point cloud to a BEV representation $\mathcal{B} \in \mathbb{R}^{H'\times W'\times 1}$, which naturally aligns with the aerial perspective and preserves geometric structure essential for localization.

B. Network Architecture

An overview of the proposed architecture is shown in Fig. 2. TransLocNet comprises modality-specific encoders, a cross-modal attention module, and a likelihood map decoder, complemented by a contrastive learning objective for robust feature alignment.

1) Feature Extraction: We employ separate encoders for the aerial map and the BEV representation of the ground LiDAR point cloud. This design choice allows each encoder to specialize in processing its respective modality while maintaining architectural consistency for effective cross-modal fusion.

BEV Encoder: Our BEV encoder processes the projected point cloud data using a ResNet-34 backbone with multi-scale feature fusion. The encoder extracts hierarchical features at three scales $(1/8,\ 1/16,\ 1/32)$ and fuses them through a refinement network with spatial and channel attention mechanisms.

The spatial attention mechanism highlights geometrically salient regions, while the channel attention mechanism emphasizes discriminative feature dimensions. To handle viewpoint variations, particularly rotation differences between LiDAR scans and aerial imagery, we incorporate a Fourier Transform layer that provides rotation-invariant representations. The magnitude spectrum of the 2D Fourier transform is invariant to rotations, as rotating an image corresponds to a phase shift in the frequency domain, leaving the magnitude spectrum unchanged.

This rotation invariance is particularly useful for BEV representations where the vehicle's heading angle may vary significantly relative to the aerial map orientation. The Fourier transform captures global structural patterns while being robust to rotational misalignment:

$$\mathcal{F}_{bev} = \text{FFT}(\text{Attention}(\text{Refine}(\text{MultiScale}(\mathcal{B})))) \quad (1)$$

where $\text{MultiScale}(\cdot)$ extracts multi-scale features, $\text{Refine}(\cdot)$ applies feature refinement with attention, and $\text{FFT}(\cdot)$ computes the log-amplitude spectrum of the Fourier transform.

Aerial Encoder: The aerial image encoder also utilizes the same ResNet-34 backbone to extract spatial features from the overhead imagery. This architectural consistency ensures that both encoders produce features in a compatible representation space, facilitating effective cross-modal attention. The encoder processes the aerial image $\mathcal{I}$ to produce feature maps $\mathcal{F}_{aerial} \in \mathbb{R}^{H/32\times W/32\times D}$ where D is the feature dimension. The aerial encoder captures semantic information such as road networks, building layouts, and vegetation patterns that aid localization.

2) Cross-Modal Attention: To bridge the modality gap between BEV and aerial features, we introduce a cross-modal attention mechanism that enables dynamic feature

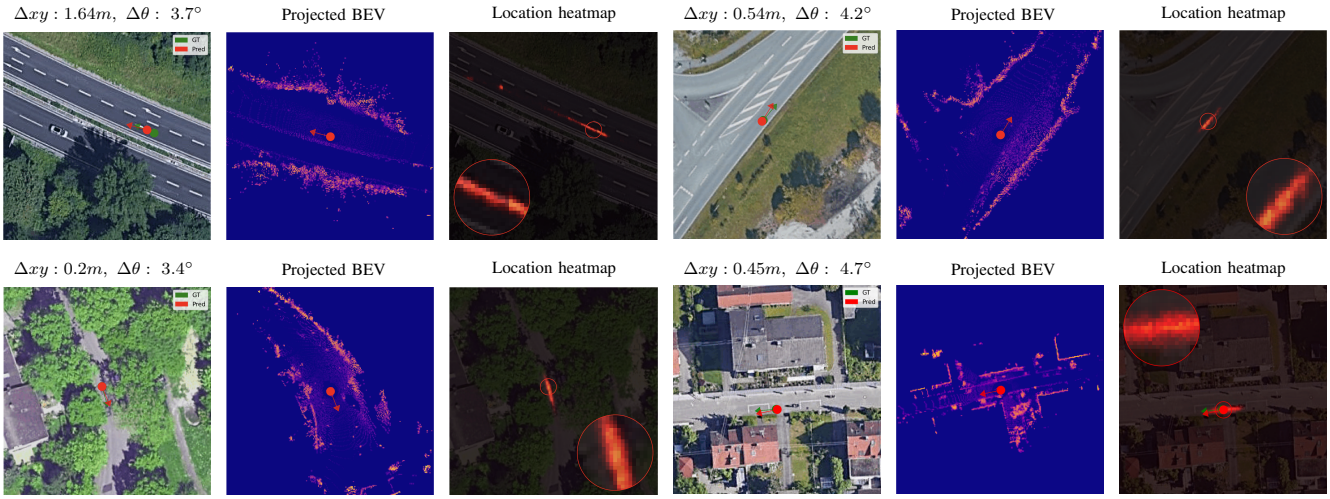

Fig. 4: Qualitative localization results on the KITTI dataset. Despite challenges from sensor noise and temporal misalignment, TransLocNet produces accurate localization with errors typically within a few meters and degrees.

alignment. Traditional fusion approaches, such as simple concatenation or element-wise operations, fail to capture the complex spatial correspondences between different viewpoints. Our attention mechanism addresses this by allowing each modality to selectively attend to relevant regions in the other modality.

This bidirectional attention enables mutual information exchange between aerial and BEV features. Each modality can selectively focus on relevant regions in the other modality. BEV features leverage aerial cues to locate corresponding structures such as roads or buildings, while aerial features exploit BEV geometry to validate ground-level details. The attention computation follows the standard transformer formulation [34]:

$$\text{Attention}_{a2b} = \text{Softmax}\left(\frac{Q_a K_b^T}{\sqrt{d_k}} + R_{pos}\right) V_b \tag{2}$$

$$\text{Attention}_{b2a} = \text{Softmax}\left(\frac{Q_b K_a^T}{\sqrt{d_k}} + R_{pos}\right) V_a \tag{3}$$

where Q_a, K_a, V_a and Q_b, K_b, V_b are query, key, and value projections for aerial and BEV features respectively, d_k is the dimension of the key vectors, and R_{pos} represents learnable relative positional encoding. The bidirectional design ensures that both modalities can influence each other, with aerial features providing global context to local BEV observations.

The relative positional encoding R_{pos} is crucial for maintaining spatial awareness across different scales and viewpoints. The attended features are then fused through a multi-scale fusion network that preserves both local and global information:

$$\mathcal{F}_{fused} = \text{Fusion}(\text{Concat}(\mathcal{F}_{aerial}, \text{Attention}_{a2b}(\mathcal{F}_{bev}))) \tag{4}$$

3) Likelihood Map Decoder: The likelihood map decoder generates spatial probability distributions for both location and orientation estimation. Unlike direct regression approaches that output point estimates, our probabilistic formulation provides uncertainty quantification and enables robust handling of ambiguous scenarios (e.g., symmetric intersections, repetitive building patterns). The decoder employs a U-Net style architecture with skip connections to preserve fine-grained spatial details while incorporating global context:

- **Location Map:** $\mathcal{M}_{loc} \in \mathbb{R}^{H \times W}$ representing the probability distribution over possible vehicle locations. This probabilistic representation naturally handles localization ambiguity and provides confidence measures.
- **Orientation Map:** $\mathcal{M}_{ori} \in \mathbb{R}^{K \times H \times W}$ where K is the number of orientation bins, encoding orientation probabilities at each spatial location. The discrete binning approach simplifies the orientation estimation task while maintaining sufficient angular resolution.
- **Confidence Score:** $c \in [0, 1]$ indicating the reliability of the prediction, learned from the accuracy of the predicted pose relative to ground truth.

The final pose estimate is obtained by finding the maximum likelihood location and orientation:

$$(x^*, y^*) = \arg\max_{(x,y)} \mathcal{M}_{loc}(x, y) \tag{5}$$

$$\theta^* = \arg\max_{k} \mathcal{M}_{ori}(k, x^*, y^*) \tag{6}$$

This approach provides interpretable uncertainty estimates and enables downstream systems to make informed decisions based on localization confidence.

C. Contrastive Learning Module

To enhance feature discriminability and improve cross-modal alignment, we integrate a contrastive learning module that learns shared feature representations between aerial and BEV modalities. The contrastive learning strategy addresses the inherent differences between LiDAR and visual data by learning a shared embedding space where semantically similar aerial-BEV pairs are close together, while dissimilar pairs are pushed apart.

Method	Map Size (in pixels)	Avg. Loc. / Ori. Error (m / °) ↓	Lat. R@Xm ↑			Long. R@Xm ↑			Ori. R@X°↑		
			1m	3m	5m	1m	3m	5m	1°	3°	5°
OrienterNet	256 × 256	2.23 / 19.15	66.87	87.67	98.61	44.38	81.51	97.69	55.78	62.71	66.26
AGL-Net	256 × 256	1.83 / 3.76	76.58	92.14	100.0	44.07	88.44	99.54	57.47	68.41	74.42
TransLocNet	256 × 256	**1.57 / 3.22**	**78.12**	**95.38**	**100.0**	**47.83**	**91.37**	**100.0**	**57.92**	**70.86**	**78.95**
OrienterNet	1024 × 1024	18.75 / 82.71	33.44	40.99	56.55	13.41	18.64	22.50	30.35	31.43	32.51
AGL-Net	1024 × 1024	14.96 / 57.25	50.54	57.94	69.49	18.64	24.19	29.89	42.06	44.22	46.22
TransLocNet	1024 × 1024	**5.75 / 10.93**	**53.04**	**69.95**	**87.38**	**25.41**	**39.60**	**60.55**	**44.84**	**61.17**	**66.87**

TABLE I: Results on the synthetic CARLA dataset. The reported numbers for AGL-Net and OrienterNet are taken directly from the AGL-Net paper. TransLocNet consistently achieves more accurate and robust performance than the baselines across both resolutions.

The module projects aerial and BEV features to a shared embedding space through separate projection heads:

$$\mathbf{z}_{aerial} = \text{Project}_{aerial}(\text{GlobalPool}(\mathcal{F}_{aerial})) \quad (7)$$

$$\mathbf{z}_{bev} = \text{Project}_{bev}(\text{GlobalPool}(\mathcal{F}_{bev})) \quad (8)$$

where GlobalPool$(\cdot)$ performs global average pooling to aggregate spatial features, and the projection heads consist of two-layer MLPs with ReLU activation and layer normalization. The projected embeddings are normalized to unit length for cosine similarity computation.

The contrastive loss uses the standard InfoNCE formulation to encourage positive pairs (matching aerial-BEV pairs) to have high similarity while pushing negative pairs apart:

$$\mathcal{L}_{contrastive} = -\log \frac{\exp(\mathbf{z}_{aerial}^T \mathbf{z}_{bev}/\tau)}{\sum_{j=1}^{B} \exp(\mathbf{z}_{aerial}^T \mathbf{z}_{bev}^{(j)}/\tau)} \quad (9)$$

where τ is the temperature parameter (set to 0.07) and B is the batch size. The diagonal elements of the similarity matrix represent positive pairs, while off-diagonal elements serve as negatives. To improve training efficiency, we implement hard negative mining by selecting the top-k hardest negatives (highest similarity) from the batch for each sample, focusing the learning on the most challenging negative pairs.

D. Multi-Objective Loss Function

Our training objective combines multiple loss components to ensure robust localization:

$$\mathcal{L}_{total} = \mathcal{L}_{loc} + \mathcal{L}_{ori} + \lambda_1 \mathcal{L}_{reg} + \lambda_2 \mathcal{L}_{contrastive} \quad (10)$$

Location Loss: We employ a likelihood-based approach using KL divergence between predicted and target location distributions. This probabilistic formulation provides natural uncertainty quantification and handles ambiguous scenarios gracefully:

$$\mathcal{L}_{loc} = \text{KL}(\mathcal{M}_{loc} || \mathcal{M}_{loc}^{gt}) \quad (11)$$

The KL divergence loss encourages the predicted distribution to match the ground truth while penalizing overconfident predictions in uncertain regions.

Direct Regression Loss: We apply Huber loss directly on predicted coordinates for additional supervision. The Huber loss provides robust training by being less sensitive to outliers compared to L2 loss, while maintaining smooth gradients compared to L1 loss:

$$\mathcal{L}_{reg} = \text{Huber}((x^*, y^*), (x_{gt}, y_{gt})) \quad (12)$$

Orientation Loss: The orientation loss uses KL divergence between predicted and target orientation distributions:

$$\mathcal{L}_{ori} = \text{KL}(\mathcal{M}_{ori} || \mathcal{M}_{ori}^{gt}) \quad (13)$$

To determine optimal loss weights, we conduct systematic sensitivity analysis on the KITTI dataset. We evaluate different combinations of $\lambda_1 \in \{0.5, 1.0, 2.0, 4.0\}$ and $\lambda_2 \in \{0.05, 0.1, 0.2, 0.5\}$ while keeping the primary losses (location and orientation) at unit weight. Emperical results show that $\lambda_1 = 2.0$ and $\lambda_2 = 0.1$ provides the best performance. The higher weight for regression loss emphasizes direct coordinate supervision, while the lower weight for contrastive loss ensures it acts as a regularizer without dominating the primary localization objective.

IV. EXPERIMENTS AND RESULTS

To validate the effectiveness of our proposed TransLocNet framework, we conduct comprehensive experiments across multiple datasets and scenarios, demonstrating significant improvements in aerial-ground vehicle localization accuracy and robustness.

A. Datasets

We evaluate TransLocNet on three diverse datasets that represent different scenarios and data modalities for aerial-ground vehicle localization:

CARLA Dataset: We utilize the synthetic CARLA dataset introduced in the AGL-Net paper [16], which provides synchronized aerial and ground-level data collected by an ego vehicle in the CARLA simulation environment. The dataset encompasses multiple urban scenarios across different towns with varying driving conditions and scenarios. The aerial imagery has a scale of 1 pixel per meter.

KITTI + OSM: We employ the original KITTI dataset [17] combined with OpenStreetMap (OSM) data for map-based localization. The KITTI dataset provides real-world LiDAR point clouds collected from a vehicle equipped with a Velodyne laser scanner, along with precise GPS/INS ground truth poses. The OSM data serves as the reference map, containing vector-based road network information, building

Map	Method	Map Size (in pixels)	Avg. Loc. / Ori. Error (m / °) ↓	Lat. R@Xm ↑			Long. R@Xm ↑			Ori. R@X° ↑		
				1m	3m	5m	1m	3m	5m	1°	3°	5°
OSM	OrienterNet	256 × 256	28.01 / 8.30	2.86	8.80	15.37	4.03	11.47	18.48	8.62	24.11	36.38
	AGL-Net	256 × 256	18.02 / 8.59	6.55	18.88	31.17	5.16	15.00	24.54	6.25	17.94	31.87
	TransLocNet	256 × 256	**15.32 / 9.21**	**8.35**	**21.42**	**34.76**	**7.32**	**19.23**	**29.12**	**9.53**	**25.13**	**39.45**
Satellite	OrienterNet*	256 × 256	27.75 / 8.65	3.44	9.53	16.89	7.48	14.21	19.62	9.26	26.23	38.21
	AGL-Net*	256 × 256	17.32 / 9.76	8.73	19.21	29.54	6.89	17.29	27.54	6.67	15.86	30.75
	TransLocNet	256 × 256	**6.47 / 5.07**	**29.22**	**67.53**	**78.79**	**31.8**	**65.34**	**75.28**	**9.81**	**30.09**	**49.77**

TABLE II: Results on the KITTI dataset using OSM and satellite aerial maps. The asterisk (*) indicates models that we retrained with satellite aerial maps to ensure consistency across baselines. TransLocNet significantly outperforms existing methods, reducing position error by over 60% and achieving the most accurate orientation estimates in the challenging satellite setting.

footprints, and other geographic features, with a scale of 2 pixels per meter.

KITTI + Aerial Imagery: We extend the KITTI dataset by incorporating aerial imagery as the map reference, following the approach described in [8]. This dataset combines the original KITTI LiDAR point clouds with high-resolution aerial imagery at a scale of 5 pixels per meter. Unlike CARLA, the aerial imagery is not temporally aligned with LiDAR collection, introducing temporal and spatial misalignment challenges. This represents the most challenging scenario due to viewpoint differences, modality gaps, and lack of temporal synchronization.

B. Metrics

We evaluate our method using both recall and metric distance for evaluation.

Recall@Xm/°: This metric measures the percentage of predicted poses that fall within X meters or X degrees of the ground-truth poses. Specifically, position recall is reported at thresholds of 1m, 3m, and 5m, while orientation recall is evaluated at 1°, 3°, and 5°.

Metric Distance: This metric reports the average absolute error between the predicted and ground-truth poses, expressed separately in meters for position and degrees for orientation.

C. Implementation Details

TransLocNet is implemented in PyTorch. Both the aerial and BEV encoders are based on ResNet-34 backbones with multi-scale feature fusion and lightweight spatial/channel attention. The encoders output features with dimension 128, which are then processed by a cross-modal attention module with 4 heads to align aerial and BEV features. The likelihood decoder outputs both position and orientation likelihood maps, with orientation discretized into 360 bins.

For contrastive learning, we apply two-layer projection heads that map the 128-dimensional features to 64-dimensional embeddings, followed by an InfoNCE loss with temperature $\tau = 0.07$. Models are trained using the Adam optimizer with an initial learning rate of 1×10^{-4}. All experiments are conducted on an NVIDIA A100 (80GB) GPU.

Computational complexity: The computational complexity of TransLocNet is primarily determined by the cross-modal attention mechanism and feature extraction. The attention computation has complexity $O(N^2 \cdot d)$ where N is the number of spatial locations and d is the feature dimension. With our configuration ($N = 256 \times 256$ and $d = 128$), the model processes aerial images of size 256×256 and BEV maps of the same resolution. The total model size is approximately 45M parameters.

Training strategy: Given the limited size of the KITTI dataset compared to the synthetic CARLA dataset, we employ comprehensive data augmentation strategies specifically for KITTI training. Our augmentation pipeline includes random aerial map rotation (up to $\pm 60°$ with 70% probability), random map cropping around the vehicle position, and color jittering. These augmentations help the model generalize to various weather and lighting conditions.

D. Results

Qualitative results: Fig. 3 and 4 illustrate qualitative localization performance on the CARLA and KITTI datasets, respectively. On CARLA dataset, where aerial and ground-level data are synchronized and noise-free, TransLocNet achieves precise pose estimation, typically with sub-meter translation error and sub-degree orientation error. On KITTI dataset, real-world conditions introduce noise and temporal misalignment between aerial imagery and LiDAR scans, yet the model generates concentrated likelihood heatmaps and reasonable predictions, keeping errors within practical bounds. These results demonstrate both the accuracy of our method in ideal conditions and its robustness to real-world challenges.

Quantitative results. Table I summarizes performance on the CARLA dataset, comparing with two baselines: AGL-Net [16] and OrienterNet [20]. These baselines are the most recent and relevant approaches for aerial-ground (LiDAR) localization, ensuring fair comparison under identical experimental conditions. At 256×256 resolution (∼256 meters), TransLocNet achieves an average error of 1.57 m / 3.22°, improving over AGL-Net (1.83 m / 3.76°) and outperforming OrienterNet (2.23 m / 19.15°) in orientation prediction. In addition, TransLocNet attains **100% recall** at 5 m, indicating reliable predictions across different weather and lighting conditions. At the higher 1024×1024 resolution (∼1 kilometer), overall errors naturally increase due to the larger search space, yet TransLocNet still achieves notable gains.

Cross attn	MultiScale. features	Prob. decoder	Avg. Loc. / Ori. Error (m / °) ↓	Loc. R@Xm ↑		Ori. R@X° ↑	
				2m	5m	2°	5°
✗	✓	✓	15.37 / 11.25	22.35	47.96	16.24	38.56
✓	✗	✓	7.84 / 7.92	26.85	59.12	16.68	43.29
✓	✓	✗	8.73 / 7.16	27.34	57.78	18.14	45.26
✓	✓	✓	**6.47 / 5.07**	**32.60**	**67.13**	**20.81**	**49.77**

TABLE III: Ablation studies on model components. For cross-modal attention, we replace it with simple concatenation of aerial and BEV features. For the multiscale feature extraction, we compare models with multiscale BEV feature extraction enabled versus disabled. For the likelihood decoder, we replace it with direct (x, y, θ) regression to predict pose coordinates directly. Results are reported on the KITTI dataset.

Compared to AGL-Net, TransLocNet reduces the localization error by **62%** (5.75 m vs. 14.96 m) and the orientation error by **80%** (10.93° vs. 57.25°).

Table II presents results on KITTI with map size of 256×256 for both OSM ($\sim$512 meters) and aerial maps ($\sim$1280 meters). For OSM-based localization, TransLocNet achieves the lowest position error of 15.32 m and a comparable orientation error of 9.21° relative to the baselines. In the more challenging aerial setting, the improvements are more pronounced: TransLocNet obtains the lowest errors for both position and orientation. Specifically, TransLocNet reduces the position error to 6.47 m, a 63% improvement over the runner-up AGL-Net. Orientation accuracy also improves to 5.07°, compared to 9.76° (AGL-Net) and 8.65° (OrienterNet).

In terms of location and orientation recall at different thresholds, TransLocNet consistently outperforms the baselines by large margins. These gains demonstrate the effectiveness of our model design in bridging modality and viewpoint gaps, especially in the aerial setting where image-only methods tend to struggle.

E. Ablation Studies

We conduct ablation studies to validate our design choices by evaluating three key components: (1) cross-modal attention versus simple concatenation, (2) multiscale versus single-scale BEV feature extraction, and (3) probabilistic decoder versus direct regression. Table III presents results on KITTI with 256×256 map size.

The ablation results demonstrate the contribution of each component. Removing cross-modal attention causes the most severe degradation with errors of 15.37 m / 11.25° and low recall rates (22.35% @ 2m, 16.24% @ 2°), indicating that simple concatenation fails to capture complex spatial relationships between aerial and BEV features.

Removing multiscale BEV feature extraction leads to drops in the model's performance. Replacing the probabilistic decoder with direct regression also shows similar moderate degradation with 8.73 m / 7.16° errors. With all three components enabled, TransLocNet achieves the best results across all metrics.

We further evaluate the impact of the direct regression loss, contrastive loss, and Fourier features in Table IV. For the Fourier features ablation, we compare models with and without the FFT layer applied to BEV features, where the FFT layer computes the log-amplitude spectrum to provide rotation-invariant representations. The direct regression loss alone yields moderate improvements while accelerating convergence. The contrastive loss acts as a regularizer, lowering error and enhancing recall. Fourier features particularly help with orientation prediction accuracy by encoding periodic patterns in the aerial imagery. The combination of all three components achieves the best overall performance, producing the lowest errors and highest recall across all metrics.

Direct reg. loss	Contrastive loss	Use Fourier features	Avg. Loc. / Ori. Error (m / °) ↓	Loc. R@Xm ↑		Ori. R@X° ↑	
				2m	5m	2°	5°
✗	✓	✓	7.15 / 6.54	**33.11**	66.35	18.63	45.11
✓	✗	✓	8.54 / 7.47	29.24	59.76	17.25	42.59
✓	✓	✗	6.82 / 7.34	32.71	65.57	16.21	41.75
✓	✓	✓	**6.47 / 5.07**	32.60	**67.13**	**20.81**	**49.77**

TABLE IV: Ablation of the loss function components and Fourier features on KITTI. We evaluate the contributions of direct regression loss, contrastive loss, and Fourier features to the total performance.

F. Limitations and Future Work

While TransLocNet improves aerial-ground vehicle localization, limitations include dependence on high-quality aerial imagery, sensitivity to LiDAR preprocessing, and computational complexity that may hinder real-time deployment. Future work will focus on developing robust features for low-quality data, lightweight attention methods, dynamic scene handling, and incorporating additional sensors like IMUs and aerial LiDAR.

V. CONCLUSION

In this work, we present TransLocNet, a cross-modal attention framework for aerial–ground vehicle localization that bridges the viewpoint and modality gaps between ground-level LiDAR point clouds and overhead aerial imagery. Our probabilistic design employs cross-modal attention to dynamically align BEV projections of LiDAR data with aerial features, complemented by a contrastive learning module that improves feature discriminability across modalities. TransLocNet consistently outperforms existing baselines, as shown by extensive experiments on both synthetic and real-world datasets. These results demonstrate the robustness and generalization of TransLocNet for reliable aerial–ground localization in both controlled and challenging real-world environments, marking a step forward for autonomous navigation systems.

ACKNOWLEDGMENTS

This material is based upon work supported in part by the DEVCOM Army Research Laboratory under cooperative agreement : W911NF2520170.

REFERENCES

[1] J. Levinson, M. Montemerlo, and S. Thrun, "Map-based precision vehicle localization in urban environments," in *Proceedings of Robotics: Science and Systems*, Atlanta, GA, USA, June 2007.

[2] J. Laconte, A. Kasmi, R. Aufrère, M. Vaidis, and R. Chapuis, "A survey of localization methods for autonomous vehicles in highway scenarios," *Sensors*, vol. 22, no. 1, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/1/247

[3] I. Jarraya, A. Al-Batati, M. B. Kadri, M. Abdelkader, A. Ammar, W. Boulila, and A. Koubaa, "Gnss-denied unmanned aerial vehicle navigation: analyzing computational complexity, sensor fusion, and localization methodologies," *Satellite Navigation*, vol. 6, no. 1, p. 9, Apr 2025. [Online]. Available: https://doi.org/10.1186/s43020-025-00162-z

[4] H. Lategahn, A. Geiger, and B. Kitt, "Visual slam for autonomous ground vehicles," in *2011 IEEE International Conference on Robotics and Automation*, 2011, pp. 1732–1737.

[5] J. Cheng, L. Zhang, Q. Chen, X. Hu, and J. Cai, "A review of visual slam methods for autonomous driving vehicles," *Engineering Applications of Artificial Intelligence*, vol. 114, p. 104992, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197622001853

[6] B. Ebrahimi Soorchaei, M. Razzaghpour, R. Valiente, A. Raftari, and Y. P. Fallah, "High-definition map representation techniques for automated vehicles," *Electronics*, vol. 11, no. 20, 2022. [Online]. Available: https://www.mdpi.com/2079-9292/11/20/3374

[7] Y. Gong, X. Zhang, J. Feng, X. He, and D. Zhang, "Lidar-based hd map localization using semantic generalized icp with road marking detection," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 3379–3386.

[8] Y. Shi and H. 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*, 2022, pp. 17 010–17 020.

[9] S. Wang, Y. Zhang, A. Vora, A. Perincherry, and H. Li, "Satellite image based cross-view localization for autonomous vehicle," *arXiv preprint arXiv:2207.13506*, 2022.

[10] S. Hu, M. Feng, R. M. H. Nguyen, and G. H. Lee, "Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7258–7267.

[11] H. Yang, X. Lu, and Y. Zhu, "Cross-view geo-localization with layer-to-layer transformer," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 29 009–29 020. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/f31b20466ae89669f9741e047487eb37-Paper.pdf

[12] Y. Yang, X. Zhao, H. C. Zhao, S. Yuan, S. M. Bateman, T. A. Huang, C. Beall, and W. Maddern, "Evaluating global geo-alignment for precision learned autonomous vehicle localization using aerial data," *arXiv preprint arXiv:2503.13896*, 2025.

[13] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *CoRR*, vol. abs/1807.03748, 2018. [Online]. Available: http://arxiv.org/abs/1807.03748

[14] J. Thoma, D. P. Paudel, and L. V. Gool, "Soft contrastive learning for visual localization," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 11 119–11 130. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/7f2cba89a7116c7c6b0a769572d5fad9-Paper.pdf

[15] Z. Deng, D. Lee, A. Adkins, J. Quattrociocchi, C. Ellis, and J. Biswas, "Spatiotemporal contrastive learning for cross-view video localization in unstructured off-road terrains," *arXiv preprint arXiv:2506.05250*, 2025.

[16] T. Guan, R. Xian, X. Wang, X. Wu, M. Elnoor, D. Song, and D. Manocha, "Agl-net: Aerial-ground cross-modal global localization with varying scales," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 8161–8161.

[17] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013.

[18] J. Levinson, M. Montemerlo, and S. Thrun, "Map-based precision vehicle localization in urban environments," in *Robotics: Science and Systems*, 2007.

[19] M. Zhou, X. Chen, N. Samano, C. Stachniss, and A. Calway, "Efficient localisation using images and openstreetmaps," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 5507–5513.

[20] P.-E. Sarlin, D. DeTone, T.-Y. Yang, A. Avetisyan, J. Straub, T. Malisiewicz, S. R. Bulo, R. Newcombe, P. Kontschieder, and V. Balntas, "Orienternet: Visual localization in 2d public maps with neural matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 632–21 642.

[21] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: A versatile and accurate monocular slam system," *IEEE transactions on robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[22] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 834–849.

[23] Z. Teed and J. Deng, "Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras," *Advances in neural information processing systems*, vol. 34, pp. 16 558–16 569, 2021.

[24] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering." *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.

[25] H. Matsuki, R. Murai, P. H. Kelly, and A. J. Davison, "Gaussian splatting slam," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 039–18 048.

[26] V. Yugay, Y. Li, T. Gevers, and M. R. Oswald, "Gaussian-slam: Photo-realistic dense slam with gaussian splatting," *arXiv preprint arXiv:2312.10070*, 2023.

[27] L. Zhu, Y. Li, E. Sandström, S. Huang, K. Schindler, and I. Armeni, "Loopsplat: Loop closure by registering 3d gaussian splats," in *2025 International Conference on 3D Vision (3DV)*. IEEE, 2025, pp. 156–167.

[28] P. Pham, D. Conover, and A. Bera, "Flashslam: Accelerated rgb-d slam for real-time 3d scene reconstruction with gaussian splatting," *arXiv preprint arXiv:2412.00682*, 2024.

[29] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5297–5307.

[30] Y. Shi, X. Yu, L. Liu, T. Zhang, and H. Li, "Optimal feature transport for cross-view image geo-localization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 990–11 997.

[31] L. Liu and H. Li, "Lending orientation to neural networks for cross-view geo-localization," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5617–5626.

[32] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, "Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 141–14 152.

[33] L. Luo, S.-Y. Cao, X. Li, J. Xu, R. Ai, Z. Yu, and X. Chen, "Bevplace++: Fast, robust, and lightweight lidar global localization for unmanned ground vehicles," *IEEE Transactions on Robotics*, 2025.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

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

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

[37] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 2774–2781.

[38] B. Liao, S. Chen, X. Wang, T. Cheng, Q. Zhang, W. Liu, and C. Huang, "MapTR: Structured modeling and learning for online vectorized HD map construction," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=k7p_YAO7yE

[39] K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz, and A. Geiger, "Transfuser: Imitation with transformer-based sensor fusion for autonomous driving," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 11, pp. 12 878–12 895, 2022.