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

Fusing Satellite Imagery and Planimetric Maps for Cross-View Localization

Quang Long Ho Ngo[1], Zimin Xia[1], Alexandre Alahi[1]

***Abstract*— Current cross-view localization methods predominantly rely on satellite imagery as the aerial modality. Although recent work explores planimetric maps (e.g., OpenStreetMap tiles), these approaches often lag in performance. Yet both modalities are widely available and possess complementary properties. Satellite images are closer to ground-level camera imagery, offering finer detail, whereas planimetric maps contain annotated objects (e.g., streetlamps) and remain informative in areas where the ground is occluded, such as by foliage. Despite this, only one prior work provides an end-to-end method to fuse the two modalities, and it does not demonstrate their potential within state-of-the-art methods. To combine the strengths of both modalities, we propose a new fusion module that augments standard encoders and demonstrates that integrating satellite imagery with planimetric maps improves state-of-the-art single-modality methods. The module comprises (i) cross-modal conditioning, which processes each modality's encoding with awareness of the other, and (ii) a patch-level fusion rule that controls the granularity of information exchange. We achieve state-of-the-art results, reducing the mean localization error by 30.13%. Qualitatively, the fusion adaptively selects the more informative modality, improving overall accuracy. https://github.com/lipefree/cross-view-fusion**

I. INTRODUCTION

Self-localization is fundamental to autonomous systems such as self-driving cars and mobile robots. These systems typically rely on Global Navigation Satellite Systems (GNSS) to estimate ego position, but in urban environments GNSS can incur errors of tens of meters [1], [2]. Cross-view localization aims to mitigate these errors by estimating the location and yaw orientation by matching a ground-level image of the current surroundings to a bird's-eye view (BEV) map of the local area identified using a noisy GNSS prior.

Current cross-view localization methods [3], [4], [5] typically use satellite imagery as the BEV map for its rich, color-consistent overhead views. However, it lacks explicit semantic labels and becomes less informative under strong occlusions, for example dense foliage. In contrast, planimetric maps such as OpenStreetMap (OSM) [6] provide explicit object annotations, remain informative under occlusion, and are widely available. As shown in Fig. 1, the two modalities encode similar content, such as roads, buildings, and object positions, but they present it differently. Satellite imagery can be uninformative under occlusion, whereas planimetric maps remain reliable. Conversely, the rich appearance detail in satellite imagery often makes it more informative than planimetric maps, as also illustrated in Fig. 1. Because both sources provide dense and complementary information about location, they are natural candidates for cross-view localization. Incorporating a fusion module into the training pipeline exploits these complementary strengths while suppressing redundancy, which yields more reliable pose estimates. However, most recent cross-view localization methods remain unimodal. Prior work [7] does include a fusion mechanism, but it is embedded within an end-to-end architecture, making it difficult to separate improvements due to architectural design from those attributable to combining the two modalities. The fusion rule operates at the scale level, which reduces design flexibility.

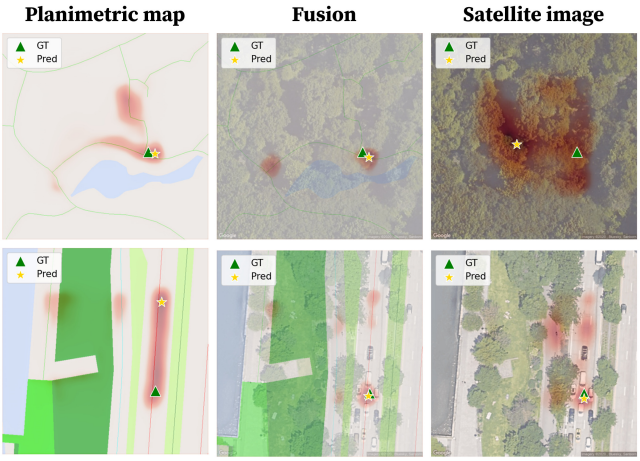

Fig. 1. Satellite images can lack information when overhead objects block the view (e.g., foliage) as seen in the first row, whereas planimetric maps constructed from OpenStreetMap data overcome these limitations. Applying our fusion module to a recent cross-view localization model yields higher confidence. The location-probability heatmap (red) is more concentrated. The fusion module also demonstrates that it adaptively selects the relevant modality for each scene.

We seek a fusion module that can be inserted into recent cross-view localization methods and that explicitly enables interaction between modalities. The fused features should be obtained through a fine-grained procedure, providing the precision required for fine-grained cross-view localization. We introduce a fusion module that integrates feature maps from satellite imagery and planimetric maps into a single representation. The module substitutes the unimodal encoder in existing cross-view localization pipelines. Features from the two BEV maps are further processed by a context-aware extractor with shared learnable queries and cross-deformable attention [8] in order to reduce redundancy across modalities and to promote cross-modal interaction. We then introduce a patch-level fusion rule, since different regions of the map may benefit from different modalities.

[1] École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

This patch-level fusion, in contrast to recent approaches that rely on simple scaled addition [9], adapts to the scene and exhibits interpretable behavior, for example relying on OpenStreetMap tiles in occluded or visually ambiguous areas while exploiting fine details provided by satellite imagery.

Our contributions are as follows: 1) We establish a new state-of-the-art on the KITTI dataset, achieving a mean localization error of 3.85 m on the challenging cross-area split, a 30.13% reduction compared to the best-performing single-modality method. 2) Our fusion module shows consistent improvements across three cross-view localization methods and two benchmarks.

II. RELATED WORK

Cross-view localization for fine-grained pose estimation [10], [11] is often implemented with a Siamese architecture: one branch processes bird's-eye view map features and the other processes ground features. The resulting aerial and ground features are then combined to estimate the ground image's location and orientation.

Satellite imagery is a popular choice for bird's-eye view maps, as it offers a semantically rich representation. [12], [13] match orientation-aware ground and satellite descriptors to estimate the location and orientation simultaneously. Some works [14], [15], [16] sweep the BEV representation of the ground image over the features extracted from the satellite image, and then select the pose with the highest alignment. However, search over the full grid is computationally expensive. Instead, [3], [5], [17], [18], [19] directly matches aerial and ground features, densely or sparsely, to estimate the relative pose. Through iterative refinement of the initial pose estimation, [4] incorporates local, global and fine-grained contexts. A proportional-integral-derivative (PID) controller [20] inspired design is used to reduce the localization error at each iteration by comparing the difference in local, global and feature gradients of the previous pose.

Planimetric maps are an alternative BEV representation that is commonly used by humans. [21] introduces a planimetric representation using OpenStreetMap [6] data, where multiple node types (e.g., roads, buildings, street lamps) are included. [22] further explores the potential of OpenStreetMap tiles with a transformer-based coarse-to-fine approach. [23] leverages geometric cues via depth estimation to enhance localization performance. In indoor localization, recent works [24], [25], [26], [27] use floor plans as planimetric maps, since overhead imagery is typically unavailable; the pose is recovered by matching ground observations to the floor plan.

Fusing satellite imagery and planimetric maps potentially combines the rich visual detail of satellite images with the labeled, occlusion-robust structural information provided by planimetric maps (e.g., OpenStreetMap). Prior work [16] learns an abstract fused map with contrastive learning, but it requires large-scale training and is not available on the fly, which limits use in unseen areas. Another approach [7] fuses the modalities within an end-to-end architecture, making it difficult to separate architectural gains from fusion benefits. We instead propose a modular fusion component, integrated during training, that can be plugged into cross-view localization methods using bird's-eye view inputs. In the related task of cross-view image retrieval, [28] fuse planimetric maps and satellite imagery by partitioning both into patches, processing the patches with a Vision Transformer [29], and employing global and local losses to enforce consistency of the fused features.

III. METHODS

Task Formulation: Given a ground-level image G and a BEV map A, the goal of cross-view localization is to estimate the 3-DoF pose $P = (x, y, o)$ of the image G, where $(x, y) \in \mathbb{R}^2$ are the planar coordinates and $o \in [-\pi, \pi)$ the yaw. This is achieved by matching the ground image with the map.

While the majority of existing cross-view localization methods [3], [5], [4], [12], [17], [18] rely solely on satellite imagery as the BEV map, a limited number of approaches [21], [23] have explored the use of planimetric maps such as OSM. Despite the availability of both modalities, current methods remain largely unimodal.

Motivation: Satellite images provide rich appearance information but are sensitive to occlusion, while planimetric maps encode simple geometric structures and remain reliable under occlusion. Fusing both modalities thus leverages their complementary strengths. Therefore, we aim to design a generic fusion framework that can be seamlessly integrated into existing single-modality cross-view localization methods. Specifically, our fusion approach considers:

1) *Information exchange across modalities:* When each map modality is processed independently, feature extraction does not account for the presence of the other modality. This can lead to suboptimal representations that capture only limited complementary information. To mitigate this, we refine the outputs of a vanilla encoder into modality-specific representations that are explicitly designed for fusion (Sec. III-A).

2) *Fusion spatial granularity:* The importance of appearance and geometry varies across space. In dense vegetation, satellite imagery offers limited semantics while planimetric maps remain informative. Along featureless roads, maps lack detail while satellite images capture surrounding structures that anchor road position. To address this, we propose learning a location-dependent fusion strategy that produces a fused BEV representation serving as the BEV map A (Sec.III-B) where each location within the map is treated by the fusion mechanism differently.

A. Context-aware feature processing

Given the feature maps $f(S)$ and $f(O)$ from the satellite image and planimetric map (both have a resolution of $N \times N$), we further process the features by jointly considering information from both modalities. As shown in Fig. 2, we generate a shared BEV query that attends to each modality via deformable attention [8].

Let $Q \in \mathbb{R}^{N^2 \times J}$ denote the query matrix; the q-th query has content vector $z_q \in \mathbb{R}^J$ given by the q-th row of Q. For

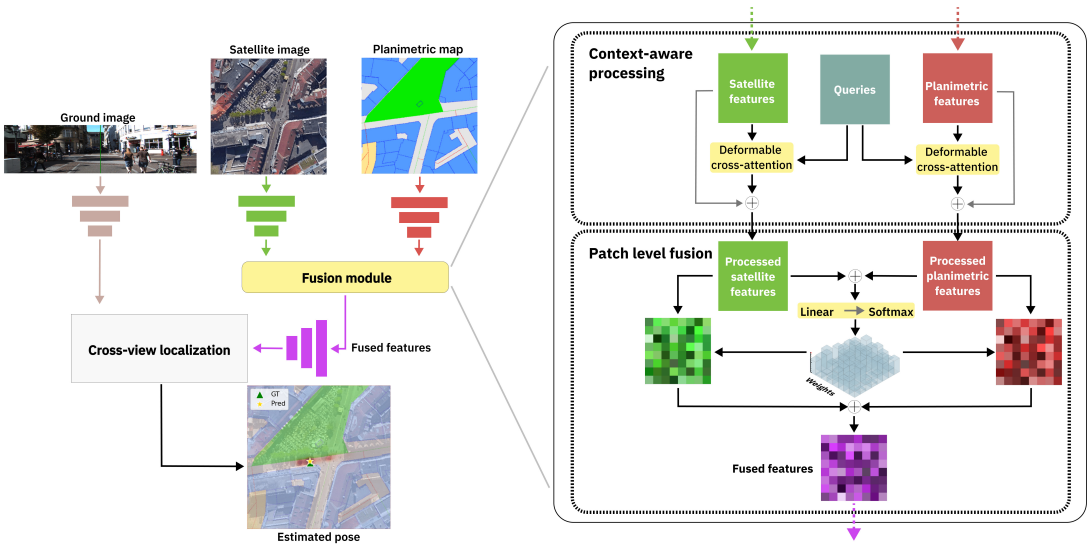

Fig. 2. Overview of the cross-view localization architecture and fusion module. Starting from feature maps extracted from the satellite image and the planimetric map, we apply context-aware processing that refines each map while conditioning on the other modality. The processed features are then fused using a patch-level rule. Across different regions of the BEV map, the model adjusts the contribution of each modality accordingly.

each modality feature map $x \in \{f(S), f(O)\}$, we treat x as a set of N^2 spatial tokens, one token per BEV cell, and use these tokens as the keys/values for attention. Concretely, we apply a 1×1 projection to obtain value embeddings $V^{(x)} \in \mathbb{R}^{J \times N \times N}$, use the same projected tokens for the internal key computation, and flatten them into a sequence $v^{(x)} \in \mathbb{R}^{N^2 \times J}$.

For each query index $q \in \{1, \ldots, N^2\}$, we associate a reference point p_q corresponding to one spatial coordinate on the BEV grid. We use the full-resolution grid at the current scale, yielding $N \times N$ reference points that tile the feature map. We provide explicit spatial information to the query by adding a 2D sine positional encoding to the query feature, so that $\tilde{z}_q = z_q + \mathrm{PE}(p_q)$.

The deformable-attention output for modality x at query q is

$$y_q^{(x)} = \mathrm{DeformAttn}\left(\tilde{z}_q, p_q, v^{(x)}\right),$$

which aggregates values from $v^{(x)}$ sampled at K learned offsets around p_q using bilinear interpolation, with attention weights conditioned on $\tilde{z}_q$. We denote the resulting context-aware features by $f_{\mathrm{DA}}(S)$ and $f_{\mathrm{DA}}(O)$.

B. Map Fusion

Our goal is to construct a single fused BEV feature map that serves as the BEV input A for localization. We therefore propose patch-level fusion: the method learns per patch weights that modulate each modality's contribution across the map, allowing different regions to rely on the most informative input. Let the patch size be $N_p \times N_p$, with $N_p \mid N$, yielding a grid of $\frac{N}{N_p} \times \frac{N}{N_p}$ patches per feature map. The fused feature for patch (i, j) is computed as a combination of the corresponding patches from the two modalities:

$$f^{i,j}(A) = \alpha^S_{i,j} f^{i,j}_{\mathrm{DA}}(S) + \alpha^O_{i,j} f^{i,j}_{\mathrm{DA}}(O), \tag{1}$$

where the patch-wise weights satisfy

$$\alpha^S_{i,j}, \alpha^O_{i,j} \geq 0, \qquad \alpha^S_{i,j} + \alpha^O_{i,j} = 1. \tag{2}$$

where $i, j \in \{1, \ldots, \frac{N}{N_p}\}$ and $f^{i,j}_{\mathrm{DA}}(\cdot)$ denotes the (i, j)-th patch after context-aware processing.

We form a patch descriptor by averaging the two modality patches,

$$\bar{f}^{i,j} = \frac{f^{i,j}_{\mathrm{DA}}(S) + f^{i,j}_{\mathrm{DA}}(O)}{2}.$$

and map it to two logits via a lightweight MLP:

$$\ell^{i,j} = \frac{1}{\sqrt{C}} \mathrm{MLP}\left(\bar{f}^{i,j}\right) \in \mathbb{R}^2. \tag{3}$$

The weights are then obtained with a softmax over the modality dimension, $[\alpha^S_{i,j}, \alpha^O_{i,j}] = \mathrm{softmax}(\ell^{i,j})$, which ensures $\alpha^S_{i,j}, \alpha^O_{i,j} \in [0, 1]$ and $\alpha^S_{i,j} + \alpha^O_{i,j} = 1$ for every patch.

C. Inference

The fused features $f(A)$ then serve as the BEV map A to the cross-view localization, which estimates the 3-DoF pose p of the ground image G. In practice, some methods [12] rely on multi-scale features and estimate the ground-truth location in a coarse-to-fine manner, using at each stage a feature map f_s from the same encoder at a different scale. For such multi-scale settings, we instantiate a separate fusion module per scale and generate scale-specific fused features, denoted $f_s(A)$. This design allows the fusion module to weight the modalities differently across scales. Compared to [30], we employ one deformable attention module per scale rather than a single multi-scale deformable attention formulation, which we find more flexible for multi-scale fusion. We do not introduce additional fusion-specific losses. Instead, the fusion module is optimized end-to-end solely through the baseline cross-view localization loss by training the entire system end-to-end.

IV. Experiments

We begin by describing the baseline models used for evaluating our approach, followed by detailing the datasets and evaluation metrics. We next present qualitative comparisons between the baselines and our proposed fusion variants. Finally, we present an ablation study.

A. Baseline models

We demonstrate the generality of our proposed fusion module by evaluating it on three state-of-the-art cross-view localization methods [3], [12], [18], each with a distinct formulation. CCVPE [12] performs global feature matching, HC-Net [18] applies a geometric transformation to ground images to reduce the gap to the BEV map, and Loc^2 [3] explicitly matches BEV-to-ground view features. In all cases, we replace the satellite feature with our fused feature map. For CCVPE, which localizes in a coarse-to-fine manner, the replacement is applied at every scale.

B. Datasets and Evaluation Metrics

We first introduce the two datasets used and the protocol for obtaining planimetric maps, and then describe the evaluation metrics.

VIGOR: The VIGOR dataset [10] contains images from four major U.S. cities. It provides panoramic ground images and corresponding aerial images covering the ground-image locations. It defines two settings: *same-area*, where models are trained and tested on the same four cities, and *cross-area*, where training is done on two cities and evaluation on the other two. We report results for *unknown orientation* (the panoramic ground image is rolled by an unknown angle) and *known orientation* (the forward direction consistently points north). We use positive samples where the ground-truth location lies within the central 1/4 region of the aerial image.

KITTI: The KITTI dataset [32], recorded in Karlsruhe, Germany, provides ground-level images with a limited field of view. Shi et al. [33] supplement it with aerial imagery covering the vehicle's location and define both *same-area* and *cross-area* settings. For training and evaluation, the ground-truth location is perturbed by up to 20 m in the longitudinal and latitudinal directions, and random noise of up to 10° is added to the orientation.

Planimetric maps: VIGOR and KITTI contain only satellite imagery for cross-view localization. We augment them with planimetric maps by using the OrienterNet [21] pipeline to generate OpenStreetMap tiles that match the zoom level and are centered on the same coordinates as the satellite images in both datasets. The data are available on our GitHub.

Metrics We evaluate performance using mean and median errors in localization and orientation prediction. We also report recall at 1 m and 5 m for the lateral and longitudinal coordinates as well as recall at 1° and 5° for orientation error on KITTI.

C. Implementation details

We use the same schedulers and learning rates as the baseline models. We use a learning rate of 1e-4 with a linear scheduler for CCVPE and HC-Net on VIGOR only. We use the AdamW [34] optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for all experiments. For CCVPE and HC-Net, we train with batch size 16. For Loc^2, we train with a batch size of 28 on VIGOR and batch size of 30 on KITTI. We use a patch size of 16 for all scales. For Loc^2, we use patch size 18 and for HC-Net we use a patch size of 2. To select the patch size, we perform a grid search over all possible sizes.

D. Hardware and software

All training and inference are done on an NVIDIA Tesla V100 using CUDA 12.1 and PyTorch version 2.4.1. Other details are available on our GitHub.

E. Quantitative results

VIGOR: We evaluate baseline architectures with and without the proposed fusion module, and we also report their single-modality results. In addition, we benchmark against state-of-the-art cross-view localization methods [5], [12], [17], [18], [31] that operate on satellite imagery. As shown in Table I, the proposed fusion consistently improves baseline performance. Notably, CCVPE benefits the most, achieving reductions of 20.28% and 18.7% in localization error under the known- and unknown-orientation settings, respectively, by exploiting our multi-scale fusion. Orientation errors are also reduced, with a 5.78° reduction in mean error and achieves state-of-the-art performance. The performance improvement is also explained by the presence of covered area in the test set. However, some samples some samples contain tunnels or underpasses where no OpenStreetMap data was recorded. HC-Net exhibits positive gains, with a 0.14 m reduction in mean localization error, indicating that local feature-based methods also benefit from the proposed fusion module. CCVPE improves across all four settings, further supporting the efficacy of the fusion module and highlighting its particular strength for global-descriptor-based methods. These results suggest that leveraging complementary modalities enhances generalization to previously unseen areas.

KITTI: We compare against OpenStreetMap-based methods [23], [21], state-of-the-art satellite-imagery-based methods, and the only prior fusion approach [7]. Our fusion module improves both CCVPE and Loc^2 models, with the largest gains in the cross-area setting. CCVPE achieves a 58% reduction in mean localization error and also exhibits lower orientation error than the original model. It also achieves a recall at 1m and 5m of 77.01% and 90.41% in the longitudinale axis in both settings, which has proven difficult to achieve with recent methods [5], [3]. We observe that the fusion module helps mitigate overfitting, which is the main issue for single-modality models on the cross-area setting. For Loc^2 [3], the fusion module reduces mean and median localization errors by 0.85 m and 0.58 m, respectively, indicating that it also enhances performance for feature-matching approaches, which require relevant BEV features

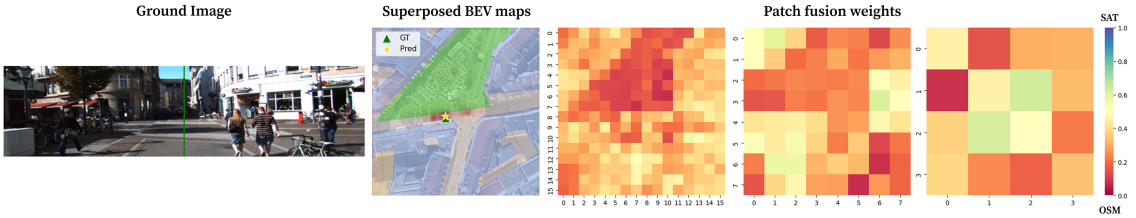

Fig. 3. The learned weights at different scales. When applying multi-scale fusion, for example in CCVPE [12], we see that different modalities are emphasized at different scales, demonstrating that both modalities are utilized. At the finest resolution, in areas corresponding to parks, the model tends to rely more on the planimetric map.

TABLE I
VIGOR TEST RESULTS

		Same-area				Cross-area			
		↓ Localization (m)		↓ Orientation (°)		↓ Localization (m)		↓ Orientation (°)	
Orien.	**Methods**	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Known	GGCVT [31]	4.12	1.34	–	–	5.16	1.40	–	–
	CCVPE [12]	3.60	1.36	–	–	4.97	1.68	–	–
	DenseFlow [17]	3.03	**0.97**	–	–	5.01	2.42	–	–
	HC-Net [18]	2.65	1.17	–	–	3.35	1.59	–	–
	FG^2 [5]	**1.95**	1.08	–	–	**2.41**	**1.37**	–	–
	CCVPE with fusion	2.87	1.24	–	–	4.05	1.82	–	–
	HC-Net with fusion	2.50	1.09	–	–	3.22	1.56	–	–
Unknown	CCVPE [12]	3.74	1.42	12.83	6.62	5.41	1.89	27.78	13.58
	DenseFlow [17]	4.97	1.90	11.20	1.59	7.67	3.67	**17.63**	**2.94**
	FG^2 [5]	8.95	7.32	15.02	2.94	10.02	8.14	31.41	5.45
	CCVPE with fusion	**3.04**	**1.24**	**7.05**	**2.37**	**4.53**	**1.78**	19.73	3.89

TABLE II
COMPARISON WITH STATE-OF-THE-ART METHODS ON SAME-AREA AND CROSS-AREA SETTINGS ON KITTI

	↓ Loc. (m)		↑ Lateral (%)		↑ Long. (%)		↓ Orien. (°)		↑ Orien. (%)	
Methods	Mean	Median	R@1m	R@5m	R@1m	R@5m	Mean	Median	R@1°	R@5°
Same-area										
GGCVT [31]	-	-	76.44	98.89	23.54	62.18	-	-	99.10	**100.00**
CCVPE [12]	1.22	0.62	97.35	99.71	77.13	97.16	0.67	0.54	77.39	99.95
HC-Net [18]	0.80	0.50	99.01	99.73	92.20	99.25	**0.45**	0.33	91.35	99.84
DenseFlow [17]	1.48	0.47	95.47	99.79	87.89	94.78	0.49	**0.30**	89.40	99.31
FG^2 [5]	0.75	0.52	**99.73**	**100.00**	86.99	98.75	1.28	0.74	61.17	95.65
Loc^2 [3]	1.11	0.76	99.60	**100.00**	65.86	98.04	1.97	1.43	36.68	92.84
Loc^2 with fusion	0.89	0.61	99.62	99.97	77.25	98.75	0.93	0.67	66.92	98.67
CCVPE with fusion	**0.72**	**0.43**	98.36	**100.00**	**95.49**	99.26	0.61	0.50	81.77	**100.00**
Cross-area										
GGCVT [31]	–	–	57.72	91.16	14.15	45.00	–	–	**98.98**	**100.00**
CCVPE [12]	9.16	3.33	44.06	92.89	23.08	64.31	1.55	0.84	57.72	96.19
HC-Net [18]	8.47	4.57	75.00	97.76	58.93	76.46	3.22	1.63	33.58	83.78
DenseFlow [17]	7.97	3.52	54.19	91.74	23.10	61.75	2.17	1.21	43.44	89.31
FG^2 [5]	7.45	4.03	**89.69**	**99.80**	12.42	55.73	3.33	1.88	30.34	81.17
Loc^2 [3]	5.51	2.97	90.97	99.22	13.70	65.06	3.32	2.12	26.03	80.68
Orienternet [21]	-	-	51.26	91.81	22.39	57.81	-	-	20.41	73.53
OSMLoc [23]	-	-	66.66	97.41	29.40	74.51	-	-	35.57	92.96
Hu et al. [7]	-	-	68.48	94.94	31.84	69.61	-	-	34.51	88.04
Loc^2 with fusion	4.66	2.39	88.24	98.06	16.87	70.66	3.40	1.73	29.62	80.67
CCVPE with fusion	**3.85**	**1.67**	86.99	96.75	**77.01**	**90.41**	**0.93**	**0.67**	66.73	99.32

to be matched to ground features. These improvements are particularly evident in cross-area experiments, where combining satellite imagery with planimetric mapping yields more robust features. We outperform all OpenStreetMap-based methods and surpass *Hu et al.* [7], validating our design choice of context-aware processing with patch-level fusion while remaining a modular fusion mechanism.

F. Qualitative results

Fusion behavior As shown in Fig. 1, some samples in the VIGOR test set contain regions where the satellite imagery

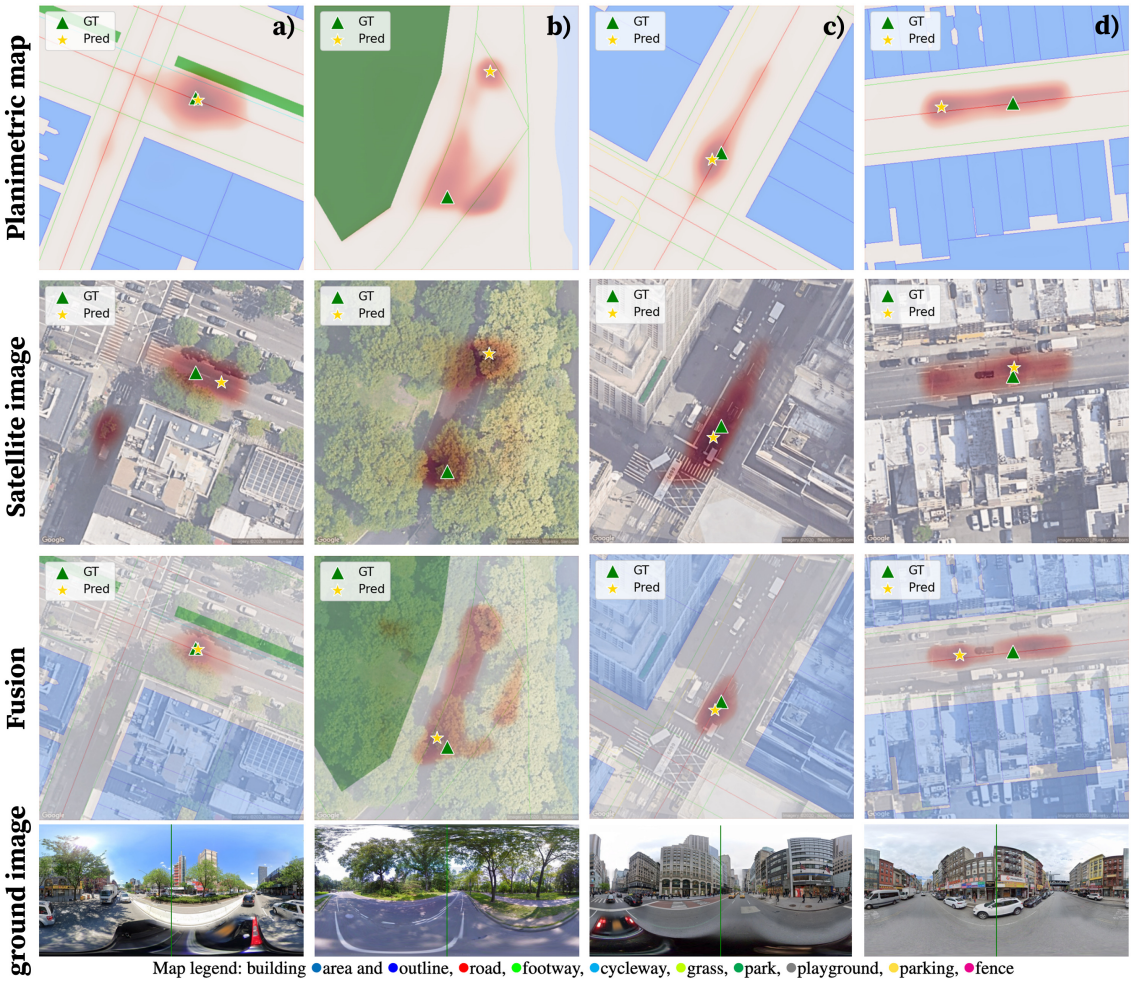

Fig. 4. **VIGOR [10] test set.** The red heatmap represents the estimated probability distribution over candidate locations indicating the uncertainty. The base model is CCVPE [12]. We compare three CCVPE variants: trained on planimetric map, on satellite imagery, and with our fusion module. The fusion variant is able to have a more concentrated heatmap and it is also able to perform localization in places where the two variants fail. However, in some scenes a single-modality CCVPE localizes correctly while the fused variant does not.

is occluded by overhead objects (e.g., tree foliage). Using satellite imagery alone often leads to localization failures in such areas, whereas OpenStreetMap tiles enable accurate pose estimates. The proposed fusion module adapts to these cases and recovers correct locations.

As observed in Fig. 4, the fusion also reduces predictive uncertainty (panels (a) and (c)), producing more concentrated location-probability maps by exploiting complementary information. We additionally observe failure cases that differ from those of the unimodal baselines (e.g., panel (b)), indicating that the model augmented with our fusion module can localize correctly in regions where both unimodal approaches fail. However, as observed in panel (d), the fusion module can occasionally underperform even when a single-modality variant localizes correctly. We further note that failure modes differ between VIGOR and KITTI. On KITTI, the module relies more heavily on planimetric maps, likely reflecting differences in OpenStreetMap coverage and quality between Germany and the United States.

Visualization of learned weights The patch-level weights in Fig. 3 show that the module adapts to scene content in this KITTI cross-area test case. In park areas, the planimetric map tends to be used more, while the rest is more balanced. In multi-scale settings, different scales emphasize different modalities, suggesting complementary roles across resolutions. Ablations that remove one modality at a time confirm these trends, yielding corresponding shifts in weight distributions and degraded performance. Notably, the weights never collapse to 0 or 1, indicating that both modalities contribute in most patches. Different regions of the map exhibit varying utilization of the available BEV maps, thereby validating our patch-level fusion approach.

Feature matching with fused features While feature-matching methods [5], [3] offer interpretable localization models, they constrain the system to identify features present in both the ground-view image and the BEV maps. In Fig. 5, the matches in panel (e) rely on structures visible in the satellite image, even when the location is occluded by buildings. The model tends to favor satellite cues that

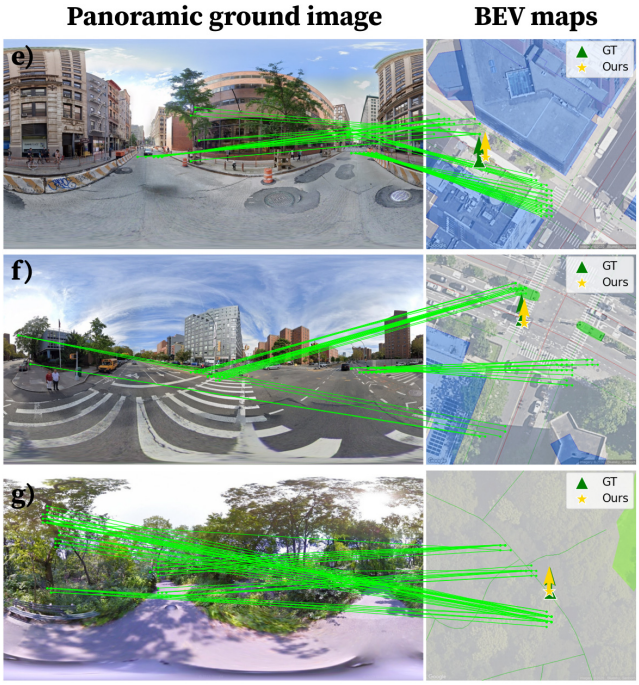

Fig. 5. Qualitative results for Loc^2 [3] augmented with our fusion module on the VIGOR same-area test set. The top 50 feature correspondences are shown, ranked by matching score. The model recovers correspondences that align with structures in the satellite imagery, particularly in panels (e) and (f). Panel (g) illustrates how road geometry from OpenStreetMap is exploited for localization in the occluded areas.

are easier to match; however, in regions such as panel (g), it falls back on the planimetric maps to localize the road observed in the image. We also observe that a nearby tree identified in the satellite image is assigned multiple feature correspondences, further underscoring the usefulness of our patch-level fusion design.

G. Ablation studies

Quantifying gains from fusion. Table IV compares CCVPE trained on a single modality with the same model augmented by our fusion module. Fusion consistently outperforms the single-modality baselines on both datasets: on VIGOR, where satellite imagery is generally advantageous, and on KITTI, where planimetric maps derived from OpenStreetMap perform better. Notably, the two models evaluated on KITTI were originally designed for satellite imagery, yet both achieve higher performance with planimetric maps, contrary to common practice in the literature. Results are reported on the VIGOR test set and the KITTI cross-area test set.

Patch size. Table III reports [3] with our fusion module on KITTI in both cross-area and same-area settings. Larger patch sizes induce greater variance in the learned fusion weights across samples, whereas smaller patches yield more stable adjustments. The optimal patch size is model- and dataset-dependent; a grid search is therefore required to obtain the best performance. We also report the relative increase in model size for each configuration compared with the original [3] model size, highlighting a crucial constraint for embedded deployment.

Model architectures. Table V evaluates the contribution of each component. In the context-aware variant, we fuse features by simple addition, whereas in the patch-level variant, we fuse features directly without the context aware stage. Both context aware processing and patch-level fusion individually improve upon the single-modality CCVPE baseline, and the full module, which combines context aware processing with patch-level fusion, achieves the largest gains, thereby supporting our design choices.

TABLE III
ABLATION STUDY OF PATCH SIZE

Patch size	**Model size**	**Position error (m) ↓**	
		Same-area	**Cross-area**
2	2.06×	0.91	4.80
6	1.13×	1.17	4.75
9	1.07×	1.06	4.83
18	1.03×	0.89	4.66
36	1.03×	1.03	4.99

TABLE IV
ABLATION STUDY OF MODALITIES

		Position error (m)	
Model	Dataset	mean	median
Satellite image			
CCVPE [12]	VIGOR	3.74	1.39
CCVPE [12]	KITTI	8.53	3.64
Loc^2[3]	KITTI	5.51	2.97
Planimetric map			
CCVPE [12]	VIGOR	4.99	2.19
CCVPE [12]	KITTI	4.55	1.77
Loc^2 [3]	KITTI	4.87	2.52
Fusion			
CCVPE [12]	VIGOR	**2.87**	**1.24**
CCVPE [12]	KITTI	**3.85**	**1.67**
Loc^2 [3]	KITTI	4.66	2.39

TABLE V
ABLATION STUDY OF FUSION COMPONENTS

	Position error (m)	
Variant	mean	median
Baseline (satellite only)	3.60	1.36
+ Context-aware processing	3.1	**1.21**
+ Patch-level fusion	3.23	1.33
Processing + Patch (full module)	**2.87**	1.24

V. CONCLUSION

Recognizing the need to leverage both modalities in cross-view localization, we propose a fusion module that combines the strengths of satellite imagery, such as color and fine building cues, with planimetric maps, which provide object

semantics and remain robust under aerial occlusions. The module refines input features by conditioning each modality on the other, and a patch-level fusion rule allocates modality contributions to the regions where they are most informative. This approach is effective across diverse scenarios: when one or both modalities degrade, the fusion incorporates complementary information to improve localization. It is also more robust in cross-area settings than single-modality models. In particular, we reduce the mean position error by 30.13% relative to the previous state-of-the-art. Our fusion module generalizes to multiple paradigms for cross-view localization, including global-descriptor-based [12], local-feature-based [18], and feature-matching approaches [3]. Qualitative analyses indicate that both modalities are actively exploited, and the learned patch-level weights provide interpretable insights into the fusion mechanism, revealing modality-specific usage across spatial regions of a scene.

REFERENCES

[1] J. Zidan, E. I. Adegoke, E. Kampert, S. A. Birrell, C. R. Ford, and M. D. Higgins, "Gnss vulnerabilities and existing solutions: A review of the literature," *IEEE Access*, vol. 9, pp. 153 960–153 976, 2021.

[2] B. Ben-Moshe, E. Elkin, H. Levi, and A. Weissman, "Improving accuracy of gnss devices in urban canyons," 01 2011.

[3] Z. Xia, C. Xu, and A. Alahi, "Loc^2: Interpretable Cross-View Localization via Depth-Lifted Local Feature Matching," *arXiv e-prints*, p. arXiv:2509.09792, Sept. 2025.

[4] W. Lee, J. Park, D. Hong, C. Sung, Y. Seo, D. Kang, and H. Myung, "Pidloc: Cross-view pose optimization network inspired by pid controllers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025, pp. 21 981–21 990.

[5] Z. Xia and A. Alahi, "Fgˆ2: Fine-grained cross-view localization by fine-grained feature matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025, pp. 6362–6372.

[6] OpenStreetMap contributors, "Planet dump retrieved from https://planet.osm.org ," https://www.openstreetmap.org, 2017.

[7] Y. Hu, Y. Liu, and B. Hui, "Combining openstreetmap with satellite imagery to enhance cross-view geo-localization," *Sensors*, vol. 25, no. 1, 2025.

[8] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," 2021. [Online]. Available: https://arxiv.org/abs/2010.04159

[9] 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 & Machine Intelligence*, vol. 47, no. 03, pp. 2020–2036, 2025.

[10] S. Zhu, T. Yang, and C. 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 (CVPR)*, June 2021, pp. 3640–3649.

[11] Z. Xia, O. Booij, M. Manfredi, and J. F. P. Kooij, "Visual cross-view metric localization with dense uncertainty estimates," in *Computer Vision - ECCV 2022*, vol. 13699, 2022, pp. 90–106.

[12] Z. Xia, O. Booij, and J. F. P. Kooij, "Convolutional cross-view pose estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 5, pp. 3813–3831, 2024.

[13] T. Lentsch, Z. Xia, H. Caesar, and J. F. P. Kooij, "Slicematch: Geometry-guided aggregation for cross-view pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 17 225–17 234.

[14] F. Fervers, S. Bullinger, C. Bodensteiner, M. Arens, and R. Stiefelhagen, "C-bev: Contrastive bird's eye view training for cross-view image retrieval and 3-dof pose estimation," 2023. [Online]. Available: https://arxiv.org/abs/2312.08060

[15] F. Florian, B. Sebastian, B. Christoph, A. Michael, and S. Rainer, "Uncertainty-aware vision-based metric cross-view geolocalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 21 621–21 631.

[16] P.-E. Sarlin, E. Trulls, M. Pollefeys, J. Hosang, and S. Lynen, "Snap: Self-supervised neural maps for visual positioning and semantic understanding," in *Advances in Neural Information Processing Systems*, vol. 36. Curran Associates, Inc., 2023, pp. 7697–7729.

[17] Z. Song, z. xianghui, J. Lu, and Y. Shi, "Learning dense flow field for highly-accurate cross-view camera localization," in *Advances in Neural Information Processing Systems*, vol. 36. Curran Associates, Inc., 2023, pp. 70 612–70 625.

[18] X. Wang, R. Xu, Z. Cui, Z. Wan, and Y. Zhang, "Fine-grained cross-view geo-localization using a correlation-aware homography estimator," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[19] S. Wang, Y. Zhang, A. Perincherry, A. Vora, and H. Li, "View consistent purification for accurate cross-view localization," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 8163–8172.

[20] G. Franklin, J. Powell, and A. Emami-Naeini, *Feedback Control Of Dynamic Systems*, 01 1994.

[21] P.-E. Sarlin, D. DeTone, T.-Y. Yang, A. Avetisyan, J. Straub, T. Malisiewicz, S. R. Bulò, 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 (CVPR)*, June 2023, pp. 21 632–21 642.

[22] H. Wu, Z. Zhang, S. Lin, X. Mu, Q. Zhao, M. Yang, and T. Qin, "Maplocnet: Coarse-to-fine feature registration for visual re-localization in navigation maps," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 13 198–13 205.

[23] Y. Liao, X. Chen, S. Kang, J. Li, Z. Dong, H. Fan, and B. Yang, "Osmloc: Single image-based visual localization in openstreetmap with geometric and semantic guidances," *arXiv preprint arXiv:2411.08665*, 2024.

[24] H. Howard-Jenkins, J.-R. Ruiz-Sarmiento, and V. A. Prisacariu, "Lalaloc: Latent layout localisation in dynamic, unvisited environments," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 10 107–10 116.

[25] Z. Min, N. Khosravan, Z. Bessinger, M. Narayana, S. B. Kang, E. Dunn, and I. Boyadzhiev, "Laser: Latent space rendering for 2d visual localization," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11 112–11 121.

[26] H. Howard-Jenkins and V. A. Prisacariu, "Lalaloc++: Global floor plan comprehension for layout localisation in unvisited environments," 2022.

[27] C. Chen, R. Wang, C. Vogel, and M. Pollefeys, "F^{3}loc: Fusion and filtering for floorplan localization," *CVPR*, 2024.

[28] Y. Tang, J. Zhang, J. Gong, Y. Li, and B. Yang, "City-level aerial geo-localization based on map matching network," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 229, pp. 65–77, 2025.

[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.

[30] S. Wang, H. Caesar, L. Nan, and J. F. P. Kooij, "Unibev: Multi-modal 3d object detection with uniform bev encoders for robustness against missing sensor modalities," in *2024 IEEE Intelligent Vehicles Symposium (IV)*, 2024, pp. 2776–2783.

[31] Y. Shi, F. Wu, A. Perincherry, A. Vora, and H. 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 (ICCV)*, October 2023, pp. 21 516–21 526.

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

[33] Y. Shi, X. Yu, L. Liu, D. Campbell, P. Koniusz, and H. li, "Accurate 3-dof camera geo-localization via ground-to-satellite image matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, pp. 1–16, 07 2022.

[34] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.