2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Orientation-Aware Multi-Modal Learning for Road Intersection Identification and Mapping

Qibin He†, Zhongyang Xiao†, Ze Huang, Hongyuan Yuan, and Li Sun*

***Abstract*— Accurate identification of road intersections is the pivotal task for automatic construction of high-definition maps, particularly in unstructured scenes. Existing methods predominantly rely on single-modal data and thus show an obvious unimodal limitation, *i.e.*, lack of contextual information. Moreover, these approaches overlook the benefits of leveraging multi-modal data fusion and representation learning that is crucial for generalizability. To this end, we propose a novel orientation-aware multi-modal learning paradigm, which formulates intersection identification as an oriented object detection task. Specifically, heterogeneous fusion is introduced to harmonize disparate data modalities, *i.e.*, vector maps, point clouds, and vehicle trajectories, into a unified feature space. Concurrently, we present trigonometry-induced adaptive regression to elevate orientation estimation, while mitigating issues related to scale imbalance and boundary confusion through dual-objective matching with spatial adaptation. To evaluate our methodology, we assemble the first-of-its-kind multi-modal benchmark tailored for complex low-speed environments, complete with fine-grained semantic annotations for intersections. Comprehensive empirical analyses, including ablation studies, affirm both the superior performance of our proposed framework and the efficacy of its constituent modules.**

I. Introduction

Road intersection is an essential element for interpreting road topology [1]–[3]. An accurate estimation of the geometry of road intersections can empower the connectivity and geometry inference of the crossed roads in automatic high-definition (HD) mapping, as shown in fig. 1. Road intersection identification is of significant challenge in unstructured scenes, such as parking lots, due to the lack of lane markings and arbitrary driving behaviors [4], [5].

Existing methods for intersection identification exhibit limitations despite leveraging multi-modal sensor data [6], [7]. Neural networks applied to vector maps lack robustness, relying heavily on pre-existing maps [8]. Image-based techniques are vulnerable to variable lighting and moving obstacles [9]. Other modalities, such as laser point clouds [10] and vehicle trajectories [11], [12], are susceptible to false positives due to intersecting paths and unexpected obstructions [13], [14]. These unimodal methods inherently face information gaps, and although multi-modal fusion could remedy this, current research remains largely single-modality focused.

This paper introduces an orientation-aware multi-modal learning framework, reframing road intersection identification as an oriented object detection problem. By harmonizing diverse modalities—including vector maps, vehicle trajectories, and LiDAR-derived occupancy grids—into a unified feature space, the framework enhances intersection representations with orientation semantics. This multi-modal feature is achieved through heterogeneous fusion with numerical discretization and distribution normalization, followed by a neighborhood search for feature association. In contrast to using horizontally acquired images, our approach leverages bird's-eye view (BEV) feature maps, addressing the challenges of irregular shapes and complex backgrounds in intersection identification. Traditional oriented object detectors often struggle with angular periodicity, leading to boundary discontinuities. This issue is exacerbated in low-speed, high-dynamic environments where small angular deviations cause significant confusion. To mitigate this, we employ trigonometry-induced adaptive regression for orientation refinement, performing dual-objective dynamic matching based on sine and cosine vectors. Coupled with a spatially adaptive factor, this allows the model to predict high-confidence bounding boxes, tailored to intersection size.

We summarize our contributions from four aspects:

- We propose an orientation-aware multi-modal learning framework that defines intersection identification as an oriented object detection task, achieving cross-modal collaborative processing and feature enhancement.
- A heterogeneous fusion technique is introduced to unify diverse modalities into a single feature space, *e.g.*, vector maps, vehicle trajectories, and LiDAR point clouds, ensuring scalablity and generalizability.
- A trigonometry-induced adaptive regression algorithm is presented to refine orientation perception through dual-objective dynamic matching, employing sine and cosine vectors. This also includes a spatial adaptation mechanism to address scale imbalance and boundary ambiguity.
- For performance evaluation, we construct the OpenInter dataset, the first multi-modal benchmark tailored for complex low-speed scenarios, offering fine-grained semantic labels for intersections.

II. Related Work

Accurate identification of road intersections is the cornerstone of unstructured high-precision mapping, providing support for topological navigation and driving decisions [2],

†Equal contribution. *Corresponding author.
The authors are with the Autonomous Driving Division, NIO Inc., Beijing, China.
Qibin He is also with the University of Chinese Academy of Sciences, Beijing, China. qibin.he@outlook.com
Ze Huang is also with the School of Data Science, Fudan University, Shanghai, China.
Li Sun is also with the Department of Computer Science, University of Sheffield, Sheffield, UK. lisunsir@gmail.com
Work done by Qibin He and Ze Huang during their internship at NIO.

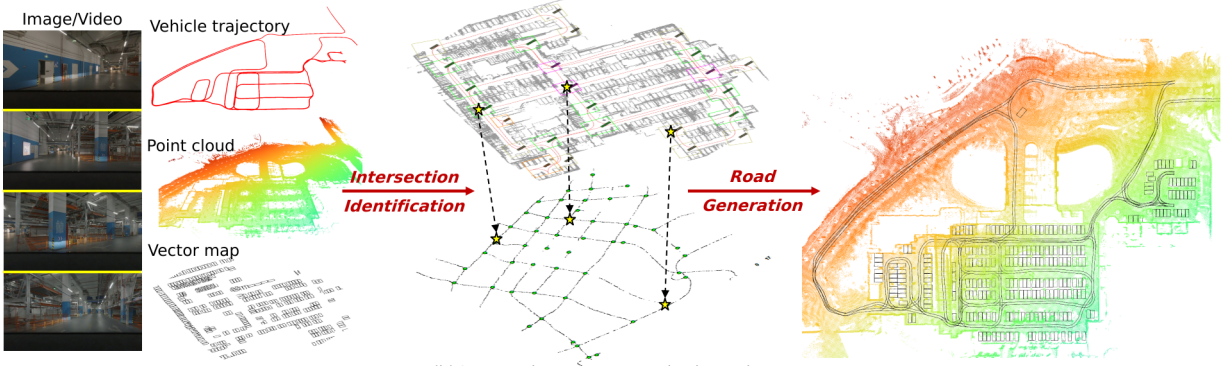


Fig. 1. The pipeline of road network generation. Intersections are nodes of road network topology, and their accurate identification is critical for HD mapping. Previous methods generally relied on a single modality (*e.g.*, image/video, point cloud, trajectory or vector map), which cannot circumvent inherent information defects. Thus we propose orientation-aware multi-modal learning to improve identification performance and robustness.

[15], [16]. Existing methods mainly rely on three types of geospatial data: images/videos, vector maps, LiDAR point clouds, and vehicle trajectories [4], [17]. Specifically, image/video and vector map related algorithms are mainly divided into rule-based [18], [19] and learning-based [20] identification methods. The former designs matching patterns through prior knowledge, *e.g.*, the configuration, location and direction of road sections, *etc*. The latter learns to extract features through iterative training, *e.g.*, shape descriptors and CNN-based detectors. Since images are easily affected by poor acquisition conditions and vector maps are limited by update frequency, some scholars try to capture the structural details of road intersections from LiDAR point cloud [10], [21]. Although LiDAR data alleviates the defects of spectral clutter and shadow occlusion, the high acquisition cost limits its practical application [22], [23]. On the contrary, thanks to the low cost and wide spatial coverage, vehicle trajectory data has been widely explored as an alternative, and has begun to be applied in research fields such as urban geography and intelligent transportation [11], [13]. The trajectory-based algorithm defines a road intersection as an area where vehicles gather and go straight, turn right, turn left, and turn around [5], [24]. The key to its success lies in the mining of the distribution pattern of turning trajectories [12], [14]. Although existing methods have made certain progress, most of them focus on a certain modality and inevitably encounter unimodal information defects. Thus, we propose a multi-modal road intersection identification framework to improve robustness and generalizability.

III. Methodology

A. Problem Formulation

Given a vector map $\mathcal{X}_{\mathrm{V}} \subseteq \mathbb{Z}^2$, LiDAR point clouds $\mathcal{X}_{\mathrm{P}} \subseteq \mathbb{Z}^3$, and vehicle trajectories $\mathcal{X}_{\mathrm{T}} \subseteq \mathbb{Z}^3$ as inputs, we first project the heterogeneous data of varying dimensions onto a 2D grid layout. We then formulate the road intersection identification problem as an oriented object detection task based on multi-modal fusion features. This formulation is

Algorithm 1: Heterogeneous Information Fusion

Input: vector map $\mathcal{X}_{\mathrm{V}} \subseteq \mathbb{Z}^2$, point cloud $\mathcal{X}_{\mathrm{P}} \subseteq \mathbb{Z}^3$, vehicle trajectory $\mathcal{X}_{\mathrm{T}} \subseteq \mathbb{Z}^3$, scale S_{res}
Output: multi-modal grid layout $\mathcal{X}^* \subseteq \mathbb{Z}^2$

$\mathcal{X}_{\mathrm{O}} \leftarrow \mathtt{Voxelizer}(\mathcal{X}_{\mathrm{P}})$;
$\mathrm{x}_{\mathrm{ref}} \leftarrow \mathtt{FindNortheast}(\mathcal{X}_{\mathrm{V}})$;
foreach modality $\mathcal{M} \in \{\mathcal{X}_{\mathrm{V}}, \mathcal{X}_{\mathrm{O}}, \mathcal{X}_{\mathrm{T}}\}$ **do**
  $\mathbf{a} \leftarrow \mathtt{Mean}(\mathcal{M})$; $\mathbf{b} \leftarrow \mathtt{Std}(\mathcal{M})$;
  foreach $\mathrm{x}_{\mathrm{m}} \in \mathcal{M}$ **do** $\mathrm{x}_{\mathrm{m}} \leftarrow \mathtt{Round}((\mathrm{x}_{\mathrm{m}} - \mathbf{a})./\mathbf{b})$;
end
while $\exists\, \mathrm{x} \neq \mathrm{x}_{\mathrm{ref}}, \mathrm{x} \in \mathcal{X}_{\mathrm{V}} \cup \mathcal{X}_{\mathrm{O}} \cup \mathcal{X}_{\mathrm{T}}$ **do**
  $\mathrm{x}^* \leftarrow ||\mathrm{x} - \mathrm{x}_{\mathrm{ref}}||./S_{\mathrm{res}}$;
end
Return $\mathcal{X}^*$;

defined as follows:

$$\mathcal{T}, \mathcal{P} = \mathrm{OriDet}(\mathrm{Fusion}(\mathcal{X}_{\mathrm{V}}, \mathcal{X}_{\mathrm{P}}, \mathcal{X}_{\mathrm{T}})), \tag{1}$$

where $\mathcal{T}$ and $\mathcal{P}$ represent the sets of location coordinates and semantic categories for all intersections, respectively. The crux of the modeling lies in the multi-modal interaction within $\mathrm{Fusion}(\cdot)$ and the instance pattern mining within $\mathrm{OriDet}(\cdot)$. To address these challenges, we design projection-based heterogeneous information fusion and orientation-aware learning mechanisms. For more details, refer to the subsequent sub-sections.

B. Heterogeneous Information Fusion

With consideration of the complementary information offered by various modalities, we propose a heterogeneous fusion approach to enhance multi-modal collaborative learning. Specifically, we concentrate on vector maps, LiDAR point clouds, and vehicle trajectories, each of which provides distinct information on the global layout, local details, and drivable areas of the mapping scene, respectively. Using the northeast corner of the vector map $\mathcal{X}_{\mathrm{V}} \subseteq \mathbb{Z}^2$ as the reference point x_{ref}, we project both the point cloud $\mathcal{X}_{\mathrm{P}} \subseteq \mathbb{Z}^3$ and the trajectory $\mathcal{X}_{\mathrm{T}} \subseteq \mathbb{Z}^3$ from the 3D coordinate system onto a 2D grid layout, *i.e.*, $\mathbb{Z}^3 \rightarrow \mathbb{Z}^2$.

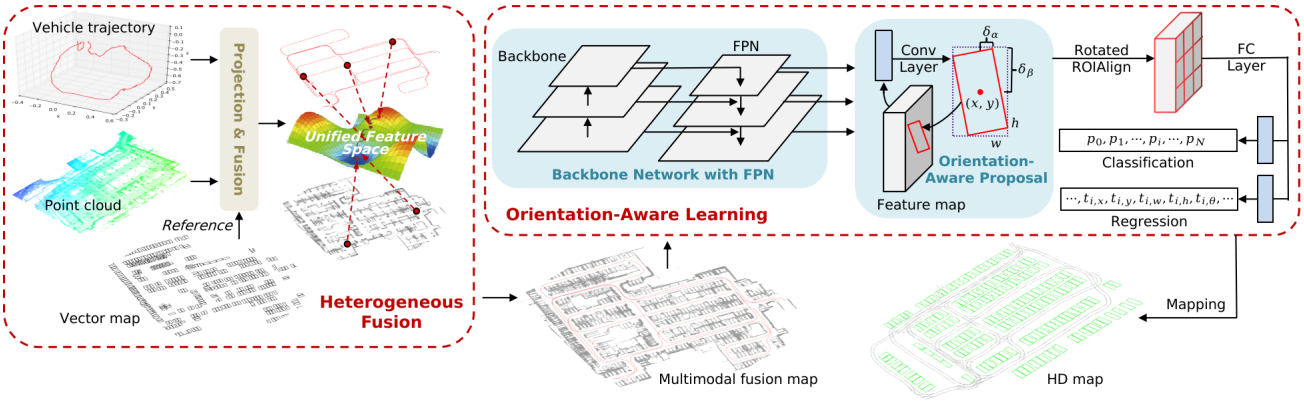


Fig. 2. The architecture of the proposed orientation-aware multi-modal learning paradigm. Vehicle trajectories, point clouds, and vector maps are first projected into a unified feature space through heterogeneous fusion. The multi-modal fusion map is then input into the backbone network with FPN to extract features, and generate orientation-aware region proposals to assist intersection semantic classification and coordinate regression. Finally, the road network of HD map is produced according to the result of intersection identification.

The projection can be mathematically formulated as:

$$\frac{||\mathrm{x}-\mathrm{x}_{\mathrm{ref}}||}{S_{\mathrm{res}}},\ \mathrm{s.t.}\,\mathrm{x}\neq\mathrm{x}_{\mathrm{ref}},\forall\,\mathrm{x}\in\mathcal{X}_{\mathrm{V}}\cup\mathcal{X}_{\mathrm{P}}\cup\mathcal{X}_{\mathrm{T}}, \quad (2)$$

where S_{res} denotes a fixed-resolution scale. To prevent overlap as multiple 3D points may map to the same 2D location, we refine eq. (2). The point cloud $\mathcal{X}_{\mathrm{P}}$ is voxelized to create an occupancy grid $\mathcal{X}_{\mathrm{O}}$ with identical voxel resolution. No additional preprocessing is required for $\mathcal{X}_{\mathrm{T}}$ due to its relative sparsity. After discretizing $\mathcal{X}_{\mathrm{O}}$ and $\mathcal{X}_{\mathrm{T}}$ on the 2D grid, we normalize the grid data as a Gaussian distribution with zero mean and a uniform standard deviation. Each grid point is then rescaled by S_{res} using a rounding operation to minimize topological changes. Finally, points are relocated to the nearest empty grid cell for refined positioning. More details are shown in algorithm 1. With the grid points' positions finalized, we can project all three modalities onto a unified feature map for subsequent orientation-aware learning.

C. Orientation-Aware Learning

1) Model Architecture: Having the multi-modal feature map obtained, we formulate the road intersection identification problem as an oriented object detection task. The overall model architecture is based on the classic two-stage detector structure [25], [26]. As depicted in fig. 2, the first stage aims to generate orientation-aware region proposals with minimal loss. In contrast, the second stage is responsible for proposal classification and regression. Specifically, referring to [27], the first stage utilizes five feature levels $\{P_2, P_3, P_4, P_5, P_6\}$ from FPN [28] as input and assigns an aspect ratio of three horizontal anchors to each spatial position across all levels. Each anchor is defined as $\mathbf{a} = \{a_x, a_y, a_h, a_w\}$, which consists of the center point coordinates $\{a_x, a_y\}$, height a_h, and width a_w. Lightweight convolutional layers are employed to predict the offset $\sigma = \{\sigma_x, \sigma_y, \sigma_h, \sigma_w, \sigma_\alpha, \sigma_\beta\}$ for decoding orientation-aware proposals:

$$\begin{aligned} &x = a_x + \sigma_x \cdot a_w, y = a_y + \sigma_y \cdot a_h, \\ &w = a_w \cdot e^{\sigma_w}, h = a_h \cdot e^{\sigma_h}, \\ &\delta_\alpha = \sigma_\alpha \cdot w, \delta_\beta = \sigma_\beta \cdot h, \end{aligned} \quad (3)$$

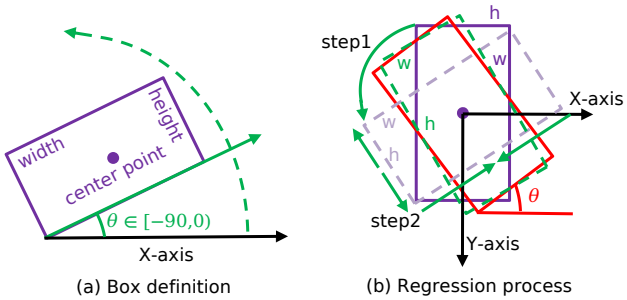


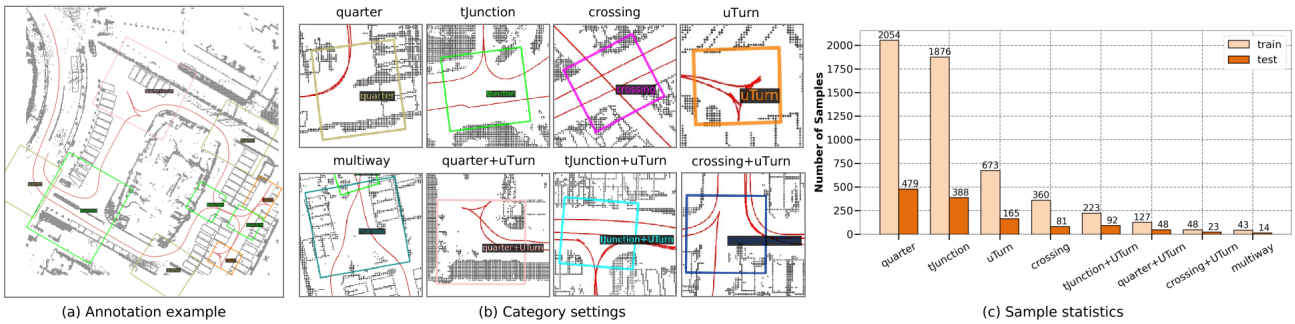
Fig. 3. (a) oriented bounding box definition and (b) example of regression process, where purple, green, and red represent proposal/anchor, prediction, and ground truth box [30].

where $\{w, h\}$ indicate the width and height of the proposal's outer horizontal box, $\{x, y\}$ denote the center coordinates, and $\{\delta_\alpha, \delta_\beta\}$ indicate the offset relative to the midpoint of the top and right edges of the horizontal box. The definitions are similar to those in [29]. The tuple $\{x, y, h, w, \delta_\alpha, \delta_\beta\}$ effectively captures the orientation information for each intersection proposal. Following proposal generation, the second stage employs rotated RoI alignment [27] to extract feature tensors of corresponding dimensions from the features $\{P_2, P_3, P_4, P_5\}$. Subsequently, each feature tensor serves as input to fully parallel connected layers, which generate coordinate offsets and class probabilities for each predicted proposal.

2) Trigonometry-Induced Adaptive Regression: Given that slight angular deviations can significantly confuse road intersections in low-speed scenarios, we introduce trigonometry-induced adaptive regression to refine orientation-aware learning. Specifically, we adopt the multi-task loss function [31] frequently used in object detection as our learning objective:

$$\mathcal{L}(\{t_i\},\{p_i\}) = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{\mathrm{reg}}(t_i, t_i^*) + \mathcal{L}_{\mathrm{cls}}(p_i, p_i^*), \quad (4)$$

where N represents the total number of anchors in the batch, and i indexes each anchor. t_i and t_i^* represent the



(a) Annotation example (b) Category settings (c) Sample statistics

Fig. 4. (a) annotation example, (b) category settings and (c) sample statistics for the OpenInter benchmark.

coordinates of the predicted and ground truth bounding boxes, respectively, while p_i and p_i^* denote the probabilities of belonging to the foreground intersection and the ground truth categories. We employ focal loss [32] as the classification loss $\mathcal{L}_{\text{cls}}$, and carefully design the regression loss $\mathcal{L}_{\text{reg}}$, which includes five parameters $\{x, y, w, h, \theta\}$. Introducing an additional angle parameter θ allows for bounding box regression in arbitrary orientations. As illustrated in fig. 3, θ is the acute angle with the horizontal axis. For $\{x, y, w, h\}$, we directly utilize prediction offset regression:

$$\begin{aligned} t_x &= (x - x_a)/w_a, t_y = (y - y_a)/h_a, \\ t_x^* &= (x^* - x_a)/w_a, t_y^* = (y^* - y_a)/h_a, \end{aligned} \tag{5}$$

where x, x_a, x^* denote the predicted, anchor, and ground truth coordinates, respectively, while other parameters follow a similar notation. Since the edges are interchangeable and angles are periodic, direct regression for θ can result in slow convergence. To refine orientation perception, we introduce dual-vector matching based on trigonometric functions:

$$\begin{aligned} t_{\sin\theta} &= \sin(\theta \cdot \pi/180), t_{\cos\theta} = \cos(\theta \cdot \pi/180), \\ t_{\sin\theta}^* &= \sin(\theta^* \cdot \pi/180), t_{\cos\theta}^* = \cos(\theta^* \cdot \pi/180). \end{aligned} \tag{6}$$

which refers to similar designs in previous work [30]. To ensure that $t_{\sin\theta}^{*2} + t_{\cos\theta}^{*2} = 1$, trigonometric normalization [29] is applied. Furthermore, considering the substantial variation in scales among different road intersections within the same scene, we introduce a spatially adaptive factor to dynamically adjust intersection regression weights:

$$\gamma_i = e^{-k_i} + e^{-l_i} + 1, \tag{7}$$

Here, k_i represents the area of the i-th generated bounding box, and l_i represents the Intersection-over-Union (IoU) with the corresponding ground label. The decreasing nature of the negative exponential function e^{-l_i} ensures greater weights for smaller intersections, thus directing the model to generate bounding boxes with high IoU to improve the boundary confusion caused by angular periodicity. Inspired by [27], [33], the trigonometry-induced adaptive regression function is set as:

$$\mathcal{L}_{\text{reg}}(t_i, t_i^*) = \sum_{j \in \{x,y,w,h,\theta\}} \gamma_i \frac{\mathcal{L}_{\text{smooth}}(t_{i,j}, t_{i,j}^*)}{|\mathcal{L}_{\text{smooth}}(t_{i,j}, t_{i,j}^*)|}, \tag{8}$$

The parameter θ includes dual-objective learning for $\{\sin\theta, \cos\theta\}$, and $\mathcal{L}_{\text{smooth}}$ indicates smoothing $\mathcal{L}_1$ loss [34]:

$$\begin{aligned} \mathcal{L}_{\text{smooth}}(t_{i,j}, t_{i,j}^*) &= \text{smooth}_{l_1}(||t_{i,j} - t_{i,j}^*||), \\ \text{smooth}_{l_1}(x) &= \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \end{aligned} \tag{9}$$

D. Application in HD Map Road Generation

The primary focus of our research is the automatic generation of High-Definition maps for unstructured roads. In this context, we utilize identified intersections to infer road topology and to construct road geometry. Specifically, the boundaries of these intersections intersect with the vehicle's mapping trajectories. We first aggregate these trajectory breakpoints to establish the "nodes" of the road network, in a manner akin to the approach described in [35]. Subsequently, we rasterize the semantic point cloud and employ Hybrid A* algorithm [36] to identify global driving paths, which serve as reference lines for the road network.

IV. OpenInter: Road Intersection Benchmark

To quantitatively evaluate the performance, we build the first multi-modal benchmark for complex low-speed scenes tailored for road intersection identification, named OpenInter.

A. Benchmark Overview

We utilize parking space vector maps, point cloud occupancy grid maps, and vehicle trajectory information to produce data with high-quality and dense intersection coordinates and category annotations. To increase data diversity and remove domain bias, we collect data on 271 parking lots carefully selected by mapping experts from multiple sensors and platforms with diverse resolutions. Specifically, our OpenInter consists of 759 annotated frames containing 6694 oriented bounding boxes of road intersections, each of which is classified to one of 8 foreground categories to describe semantic properties. As shown in fig. 4, the semantic categories include *quarter, uTurn, tJunction, crossing, multiway, crossing+uTurn, quarter+uTurn and tJunction+uTurn*. The category division largely depends on the driving trajectory and whether an intersection is common to judge, which aims to provide meaningful drivable intersection locations. And some other areas like intersections may be impassable, *i.e.*, they have no practical application value for mapping in real-world autonomous driving scenarios, so we remove them.

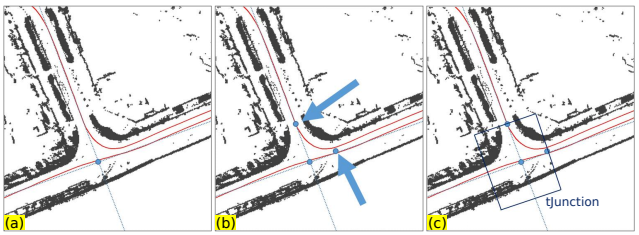


Fig. 5. The labeling process for our road intersection data. (a) Find the center of each intersection. (b) Find the critical point where the trajectory changes from curved to straight and from divergent to convergent. (c) According to the principle of symmetry and approximate rectangle, complete the oriented bounding box and specify the category.

TABLE I

PERFORMANCE COMPARISON WITH REPRESENTATIVE ORIENTED OBJECT DETECTORS ON OPENINTER.

Method	Backb.	Per-class AP(%)								mAP(%)
		Q	T	C	U	M	Q+U	T+U	C+U	
CFA [37]	R50	88.5	81.2	75.9	31.0	8.1	50.9	49.6	18.3	50.4
S^2A-Net [33]	R50	87.4	75.0	70.6	32.9	**16.3**	52.5	38.0	42.6	51.9
RoI Trans [29]	R50	88.1	69.2	73.5	41.5	11.0	48.1	29.5	61.4	52.8
Oriented RCNN [27]	R50	86.9	82.5	58.9	47.1	5.7	55.4	36.7	50.8	53.0
KFIoU [30]	R50	87.5	71.3	68.4	29.3	4.2	71.9	65.2	29.7	53.4
Ours	R50	88.9	67.9	72.1	37.8	3.5	32.2	**71.4**	55.3	53.6
Ours	R101	**90.3**	**88.2**	**77.0**	13.3	1.4	68.6	34.3	**65.8**	54.9
Ours	Swin-B	88.9	86.8	37.8	**72.8**	12.6	**75.6**	55.3	22.4	**56.5**

†Q: quarter, T: tJunction, C: crossing, U: uTurn, M: multiway, Q+U: quarter+uTurn, T+U: tJunction+uTurn, C+U: crossing+uTurn.

Obviously, the potential category imbalance in OpenInter brings challenges to fine-grained intersection recognition, especially few-shot category (*e.g.*, *multiway*) puts higher requirements on algorithm performance. To fairly test the generalization capability, we choose the data of 200 parking lots as the training set, and the data of the remaining 71 parking lots as the test set, where the corresponding number of frames are 553 and 206 respectively.

B. Annotation Protocol

To facilitate unstructured high-precision road network mapping, we consider various annotation protocols to better label road intersections. Inspired by the bounding box annotations commonly used in computer vision to describe visual concepts (*e.g.*, objects, regions, and attributes), we chose the protocol of any quadrilateral. The intersection is expressed as $\{(x_i, y_i), i = 1, 2, 3, 4\}$ to compactly and accurately draw the outline, where (x_i, y_i) has the position of each vertex of the bounding box coordinate. As shown in fig. 5, to control the annotation quality, we strictly regulate the annotation process. Compared with the usual positioning, the fine-grained semantic annotation in OpenInter provides more possibilities for downstream application expansion.

V. EXPERIMENTS AND DISCUSSIONS

A. Implementation Details

In the experiment, the orientation-aware learning part of the proposed method is built in the classic oriented detector framework [27] by default, and the ResNet-50/101 (R50/101) [38] and Swin Transformer Base (Swin-B) [39] pre-trained on ImageNet [40] are selected as the backbone network. The AdamW [41] optimizer with an initial learning rate of 0.0004 and a weight decay of 0.05 is employed to train the model for 36 epochs. We use 8 NVIDIA A100 GPUs with a batch size of 10 for model training, and a single A100 GPU for testing. In the implementation, all data are cropped into 1024×1024 size patches with a 512 pixel overlap. Following [42], we employ mean Average Precision (mAP) as the accuracy metric for quantitative evaluation.

TABLE II

PERFORMANCE COMPARISON ON HD MAPPING.

Method	NRoad↑	R2T↑	En2R↑	En2P↑	P2Ex↑
xDeepFM+DBSCAN [14]	64.6	78.8	90.2	96.3	89.4
Multi-Scale Graph [24]	72.0	82.2	91.7	96.9	88.7
Ours (Swin-B)	77.2	91.3	95.9	98.6	90.2

B. Results and Discussions

1) Results on OpenInter: To verify the performance of the proposed method, we compare it with other classical oriented detectors on the OpenInter benchmark. As shown in table I, our method outperforms advanced KFIoU [30] by a slight advantage of 0.2% when using ResNet-50 as the backbone network. Especially for the two composite categories of *tJunction+uTurn* and *crossing+uTurn*, it outperforms KFIoU by 6.2% and 25.6%. We believe that the main reason is that the trajectories of these categories have complex geometric topologies, and the proposed trigonometry-induced adaptive regression can be more adaptable to such irregular intersections by refining the orientation perception. Overall, the advantages of the proposed method are obvious. When Swin-B is used as the backbone to extract features, the overall performance gain is increased by as much as 3.1%. This also proves that it has good scalability for different visual backbones, whether CNN or Transformer.

2) Application: We further verify the impact of intersection identification performance on HD mapping. As shown in table II, the proposed method has obvious advantages over other classic competitors. NRoad indicates the number of roads. R2T indicates the coverage completeness of the road network relative to the collected trajectories. En2R, En2P and P2Ex respectively indicate the routing success rate from the entrance to other roads, from the entrance to points of interest (*i.e.*, charging pile), and from points of interest to exits. Excellent performance on NRoad and R2T show that a more intact road network can be obtained using our method, while En2R, En2P and P2Ex reflect superior topological connectivity. As shown in fig. 6, thanks to the high recall of intersection identification, the road network in the HD map with our method is more complete and reliable.

C. Ablation Study

As shown in tables III to V, we conduct a series of ablation experiments on OpenInter. All models employ ResNet-50 as the backbone network unless otherwise specified.

1) Effect of Multiple Modalities: To explore the impact of multi-modal processing on intersection identification, we evaluate the performance of unimodal and multi-modal inputs. As shown in table III, the method based on trajectory as input is superior to other unimodal processing, which is

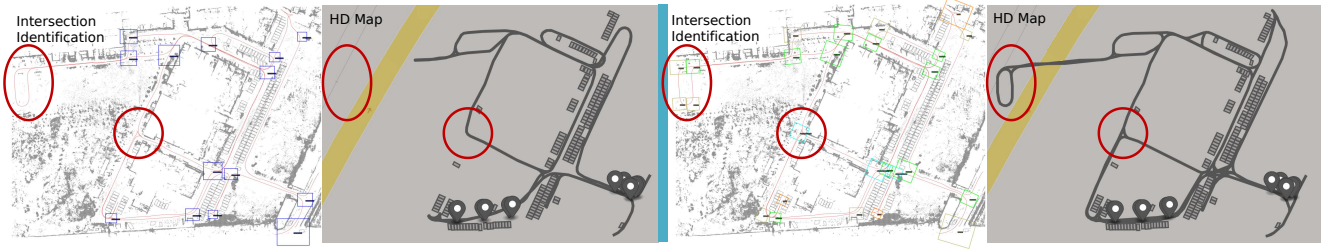


Fig. 6. Qualitative visual comparison between [24] (left) and our method (right). The color of the bounding boxes in the identification results is used to distinguish different categories. Obviously, our method can provide fine-grained semantics and orientation information of intersections. Some intersections that were missed during identification resulted in the inability to generate a complete road network during HD mapping, which are marked by red circles.

TABLE III

ABLATIONS FOR DIFFERENT MODALITY DATA.

OGM	VM	VT	Per-class AP(%) Q	T	C	U	M	mAP(%)
✓	✗	✗	48.3	32.9	44.6	13.3	1.8	31.2
✗	✓	✗	72.5	47.4	59.3	18.7	2.2	44.8
✗	✗	✓	83.1	52.5	65.6	26.5	**7.1**	49.6
✓	✓	✓	**88.9**	**67.9**	**72.1**	**37.8**	3.5	**53.6**

†VT: vehicle trajectory, VM: vector map, OGM: occupancy grid map (*i.e.*, point cloud).

TABLE IV

ABLATIONS FOR DUAL-OBJECTIVE MATCHING.

$t_{\sin\theta}$	$t_{\cos\theta}$	Per-class AP(%) Q	T	C	U	M	mAP(%)
✗	✗	84.2	63.8	67.5	31.6	**4.2**	49.8
✓	✗	86.7	65.2	70.1	32.8	2.7	51.7
✗	✓	87.1	64.7	68.5	34.5	3.1	52.4
✓	✓	**88.9**	**67.9**	**72.1**	**37.8**	3.5	**53.6**

TABLE V

ABLATIONS FOR SPATIALLY ADAPTIVE FACTOR.

e^{-k_i}	e^{-l_i}	Per-class AP(%) Q	T	C	U	M	mAP(%)
✗	✗	87.3	67.1	71.3	35.4	2.1	52.2
✓	✗	87.9	67.8	71.4	36.9	3.2	52.9
✗	✓	88.2	67.3	71.6	37.2	2.9	53.1
✓	✓	**88.9**	**67.9**	**72.1**	**37.8**	**3.5**	**53.6**

consistent with the prior that intersection labeling mainly relies on trajectory judgment. The performance of the multi-modal method shows impressive improvements, especially a 22.4% enhancement compared to only using OGM. Notably, employing the proposed framework to process multi-modal information hardly increases additional computational costs, but brings significant advantages in identification accuracy.

2) Effect of Dual-Objective Matching: We experimentally explore the impact of trigonometry-induced optimization objectives on identification. Specifically, we apply direct regression similar to eq. (5) to θ as the baseline, *i.e.*, we train θ in the following manner:

$$t_\theta = (\theta - \theta_a) \cdot \pi/180, t^*_\theta = (\theta^* - \theta_a) \cdot \pi/180. \quad (10)$$

As shown in table IV, whether $t_{\sin\theta}$ or $t_{\cos\theta}$ is used as the objective alone for regression, the performance can be stably improved, which proves that the introduction of trigonometric function coding can effectively assist θ optimization. Consistently, the integration of $t_{\sin\theta}$ and $t_{\cos\theta}$ achieves the best results, because the uniqueness of dual-vector helps alleviate the coordinate ambiguity.

3) Effect of Spatially Adaptive Factor: As shown in table V, introducing e^{-k_i} effectively improves the mAP, which drives the model to pay more attention to small-scale objects in training through the monotonous decrease of the negative exponential function. In addition, e^{-l_i} alleviates boundary confusion by prompting the model to produce bounding boxes with high IoU and thereby boosts overall accuracy. The crystallization of the two brings the best results, proving that injecting the spatially adaptive factor into the regression loss is a good training strategy for localization.

4) Feature Visualization: By visualizing the feature response maps, we qualitatively evaluate the intersection representation capability of the model. As shown in fig. 7, it can be seen that with the help of the spatially adaptive factor, the model can respond with more discriminative

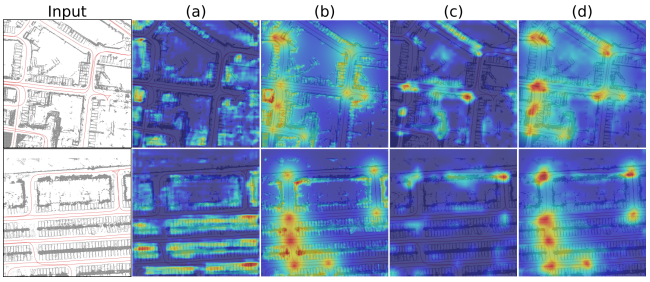


Fig. 7. Feature visualization with different learning objectives. (a) Direct regression (DR). (b) Trigonometry-induced dual-objective matching (TDM). (c) DR with spatially adaptive factor. (d) TDM with spatially adaptive factor.

power when facing small-area intersections. Employing dual-objective matching enables the identification of some irregular special cases, *e.g.*, *tJunction+uTurn* with interference clutter. Overall, the proposed trigonometry-induced adaptive regression helps the model distinguish intersections from the background, thereby achieving more robust positioning.

VI. CONCLUSION

This paper proposes a novel orientation-aware multi-modal learning paradigm for intersection identification and mapping. Specifically, a heterogeneous fusion technique is introduced to project different modality information into a unified feature space. A trigonometry-induced adaptive regression algorithm is designed to refine orientation perception, improving scale imbalance and boundary confusion through dual-objective matching with spatial adaptation. Importantly, our research advances the state-of-the-art by assembling a pioneering multi-modal benchmark, designed explicitly for intricate, low-speed operational scenarios. This dataset not only allows for robust evaluation but also paves the way for future research by offering a new gold standard. Comprehensive experimental analyses, inclusive of ablation studies, firmly substantiate the effectiveness of the proposed computational constructs and the overarching framework.

REFERENCES

[1] V. Sezer, T. Bandyopadhyay, D. Rus, E. Frazzoli, and D. Hsu, "Towards autonomous navigation of unsignalized intersections under uncertainty of human driver intent," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 3578–3585.

[2] A. Amini, G. Rosman, S. Karaman, and D. Rus, "Variational end-to-end navigation and localization," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 8958–8964.

[3] S. Khaitan and J. M. Dolan, "State dropout-based curriculum reinforcement learning for self-driving at unsignalized intersections," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 12 219–12 224.

[4] 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*, 2023, pp. 21 632–21 642.

[5] X. Yang, K. Stewart, L. Tang, Z. Xie, and Q. Li, "A review of gps trajectories classification based on transportation mode," *Sensors*, vol. 18, no. 11, p. 3741, 2018.

[6] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google earth engine: Planetary-scale geospatial analysis for everyone," *Remote sensing of Environment*, vol. 202, pp. 18–27, 2017.

[7] K. Nakamura and S. Bansal, "Online update of safety assurances using confidence-based predictions," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 12 765–12 771.

[8] M. Saeedimoghaddam and T. F. Stepinski, "Automatic extraction of road intersection points from usgs historical map series using deep convolutional neural networks," *International Journal of Geographical Information Science*, vol. 34, no. 5, pp. 947–968, 2020.

[9] Y. Wang *et al.*, "Ddu-net: Dual-decoder-u-net for road extraction using high-resolution remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.

[10] Z. Dong, F. Liang, B. Yang, Y. Xu, Y. Zang, J. Li, Y. Wang, W. Dai, H. Fan, J. Hyyppä *et al.*, "Registration of large-scale terrestrial laser scanner point clouds: A review and benchmark," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 163, pp. 327–342, 2020.

[11] X. Yang, L. Hou, M. Guo, Y. Cao, M. Yang, and L. Tang, "Road intersection identification from crowdsourced big trace data using mask-rcnn," *Transactions in GIS*, vol. 26, no. 1, pp. 278–296, 2022.

[12] Y. Zhang, G. Tang, X. Fang, T. Chen, F. Zhou, and Y. Luo, "Hierarchical segmentation method for generating road intersections from crowdsourced trajectory data," *Applied Sciences*, vol. 12, no. 20, p. 10383, 2022.

[13] T. Alsahfi, M. Almotairi, R. Elmasri, and B. Alshemaimri, "Road map generation and feature extraction from gps trajectories data," in *Proceedings of the 12th ACM SIGSPATIAL International Workshop on Computational Transportation Science*, 2019, pp. 1–10.

[14] Y. Liu, R. Qing, Y. Zhao, and Z. Liao, "Road intersection recognition via combining classification model and clustering algorithm based on gps data," *ISPRS International Journal of Geo-Information*, vol. 11, no. 9, p. 487, 2022.

[15] H.-M. Cheng and D. Song, "Graph-based proprioceptive localization using a discrete heading-length feature sequence matching approach," *IEEE Transactions on Robotics*, vol. 37, no. 4, pp. 1268–1281, 2021.

[16] M. Haklay and P. Weber, "Openstreetmap: User-generated street maps," *IEEE Pervasive computing*, vol. 7, no. 4, pp. 12–18, 2008.

[17] M. Hashemi, "Automatic inference of road and pedestrian networks from spatial-temporal trajectories," *IEEE transactions on intelligent transportation systems*, vol. 20, no. 12, pp. 4604–4620, 2019.

[18] T. Ort, K. Murthy, R. Banerjee, S. K. Gottipati, D. Bhatt, I. Gilitschenski, L. Paull, and D. Rus, "Maplite: Autonomous intersection navigation without a detailed prior map," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 556–563, 2019.

[19] D. Feng, Y. Zhou, C. Xu, M. Tomizuka, and W. Zhan, "A simple and efficient multi-task network for 3d object detection and road understanding," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 7067–7074.

[20] Y. Wang, Y. Sun, J. Li, and M. Shi, "Cross-modal fusion-based prior correction for road detection in off-road environments," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 12 239–12 246.

[21] Y. Jung, M. Jeon, C. Kim, S.-W. Seo, and S.-W. Kim, "Uncertainty-aware fast curb detection using convolutional networks in point clouds," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 12 882–12 888.

[22] 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)*. IEEE, 2023, pp. 2774–2781.

[23] P. Sundaresan, R. Antonova, and J. Bohgl, "Diffcloud: Real-to-sim from point clouds with differentiable simulation and rendering of deformable objects," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 10 828–10 835.

[24] Y. Yin, A. Sunderrajan, X. Huang, J. Varadarajan, G. Wang, D. Sahrawat, Y. Zhang, R. Zimmermann, and S.-K. Ng, "Multi-scale graph convolutional network for intersection detection from gps trajectories," in *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, 2019, pp. 36–39.

[25] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 976–11 986.

[26] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[27] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented r-cnn for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3520–3529.

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

[29] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning roi transformer for oriented object detection in aerial images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2849–2858.

[30] X. Yang, Y. Zhou, G. Zhang, J. Yang, W. Wang, J. Yan, X. Zhang, and Q. Tian, "The kfiou loss for rotated object detection," *arXiv preprint arXiv:2201.12558*, 2022.

[31] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proceedings of the IEEE*, 2023.

[32] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[33] J. Han, J. Ding, J. Li, and G.-S. Xia, "Align deep features for oriented object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2021.

[34] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[35] X. Yang, L. Tang, L. Niu, X. Zhang, and Q. Li, "Generating lane-based intersection maps from crowdsourcing big trace data," *Transportation Research Part C: Emerging Technologies*, vol. 89, pp. 168–187, 2018.

[36] D. Dolgov, S. Thrun, M. Montemerlo, and J. Diebel, "Practical search techniques in path planning for autonomous driving," *Ann Arbor*, vol. 1001, no. 48105, pp. 18–80, 2008.

[37] Z. Guo, C. Liu, X. Zhang, J. Jiao, X. Ji, and Q. Ye, "Beyond bounding-box: Convex-hull feature adaptation for oriented and densely packed object detection," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2021, pp. 8792–8801.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[39] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[40] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 248–255.

[41] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[42] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, pp. 303–338, 2010.