Traffic Sign Localization and Orientation Classification for Automated Map Updating

Xianjing Han, Wenmiao Hu, Xuemeng Song, *Senior Member, IEEE*, Hannes Kruppa, See-Kiong Ng, Roger Zimmermann, *Senior Member, IEEE*

***Abstract*—High Definition (HD) maps, containing detailed road information, are essential for autonomous driving and many geo-related tasks. Recent developments in computer vision make it possible to automate the labor-intensive HD map maintenance work, such as localizing traffic signs within a road network. However, updating traffic signs to HD maps is non-trivial, as it not only requires precise geo-location but also requires confirming whether a sign belongs to a specific road. In our work, we develop an end-to-end automated traffic sign update system, termed AutoTS, which is capable of using an image sequence collected during vehicle operation to extract the geo-location of a traffic sign and determine whether it belongs to the road driven on, from its orientation. In AutoTS, we design a noise and sparsity adaptive localization module, which can filter noisy location points and derive a geo-location from sparse location points. To identify the orientation of traffic signs, we devise a position-aware orientation classification module, which uses the ROI feature and the position-aware SIFT feature to explore the orientation characteristic and understand the road context. To facilitate the evaluation of the proposed method, we construct a traffic sign localization and orientation classification benchmark, KITTI-TS. Our AutoTS achieves an MAE of 2.38 meters in traffic sign localization, while the accuracy in orientation classification reaches 88.89%.**

***Index Terms*—Automated map making, traffic sign localization, spectral clustering.**

I. Introduction

High Definition (HD) maps, rich in detailed road boundaries, lanes, traffic signs, and other semantically meaningful landmarks, are pivotal in many geo-related scenarios, such as autonomous driving [1], [2] and traffic route planning [3]. The efficacy of HD maps relies heavily on timely updates, especially regarding traffic signs (*e.g.*, speed limit signs), which are crucial for tasks like estimated time of arrival predictions [4], [5] and traffic condition prediction [6]. However, in practice, maintaining HD maps needs professionals to drive automotive data collection vehicles equipped with high-grade sensors to gather updated information about traffic signs for all streets, which is time-consuming and inefficient.

Fortunately, the widespread use of devices that are equipped with cameras and Global Navigation Satellite Systems (GNSS, *e.g.*, GPS), along with advances in computer vision, has facilitated more efficient update schemes. In particular, several studies [7]–[9] have focused on the task of automated traffic sign localization by harnessing extensive street images captured by the vehicle dashboard camera [10]. Typically, they extract the locations of traffic signs from multiple street images and cluster them to determine the precise geo-location, or apply Structure from Motion (SfM) [11] to extract the location of a traffic sign from multi-view images. However, in real-world applications, updating traffic signs to HD maps not only requires their geo-locations but also their orientations (*i.e.*, the direction in which the sign is intended to guide vehicles) to confirm whether the traffic sign belongs to the current road. As shown in Figure 1, each detected traffic sign exhibits a specific orientation. Accurate orientation classification helps reduce map update errors and contributes to improving downstream geo-related tasks.

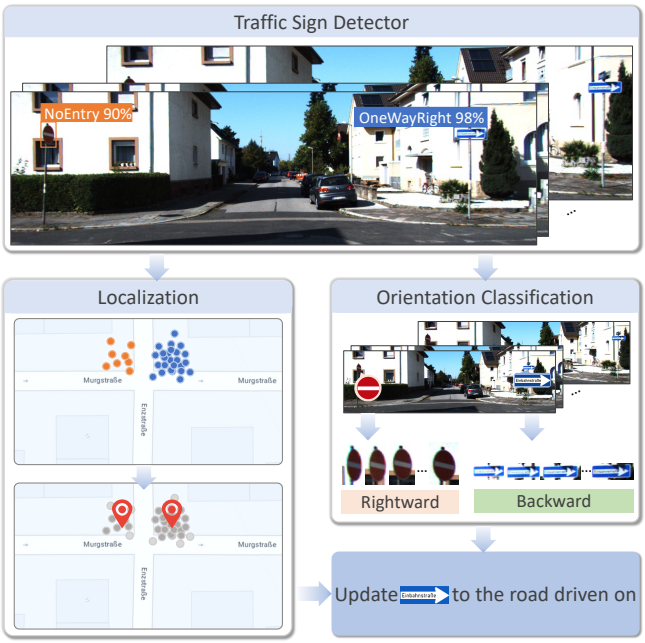


Fig. 1. The automated traffic sign update system extracts the geo-location and orientation from the detected traffic sign sequence. Detected traffic signs exhibit varying orientations. The "One Way for Right" sign is of backward orientation and belongs to the road driven on, while the "No Entry" sign is of rightward orientation and belongs to the road intersecting with the road driven on. Our system updates the "One Way for Right" sign to the current driving road on HD maps.

Therefore, to achieve comprehensive traffic sign updates, our work aims to simultaneously estimate the location and orientation of traffic signs captured in a sequence of street

Xianjing Han, Wenmiao Hu, See-Kiong Ng, and Roger Zimmermann are with Grab-NUS AI Lab, National University of Singapore, 117602, Singapore. E-mail: xianjing@nus.edu.sg, hu.wenmiao@u.nus.edu, seekiong@nus.edu.sg, rogerz@comp.nus.edu.sg.

Xuemeng Song is with the School of Computer Science and Technology, Shandong University, 266237, China. E-mail: sxmustc@gmail.com.

Hannes Kruppa is with the Grabtaxi Holdings Pte. Ltd., 018937, Singapore. E-mail: hannes.kruppa@grabtaxi.com.

(a) Normally placed yield sign (a) Badly placed yield sign

Fig. 2. Two "Yield" signs have different placements, but both of them belong to their current driving roads (*i.e.*, backward orientation).

images. However, this task is non-trivial due to the following challenges. **1)** Noisy and Sparse Location Points. To estimate the location of a traffic sign, we require both the GPS location of each street image and the depth between the camera and the traffic sign. However, errors in GPS signals and depth estimation can introduce noisy location points. Additionally, factors such as limited camera field of view (FOV) and poor lighting conditions may cause traffic signs to be undetected, resulting in sparse location data. Therefore, how to derive accurate geo-locations of traffic signs from such noisy and sparse location points constitutes a challenge. And **2)** Complex Orientation Characteristics in Real-World Scenarios. The orientation of traffic signs can be subtle and is easily affected by real-world placement conditions. Figure 2 illustrates two 'Yield' signs, one with standard placement and the other poorly placed. The improper placement makes it difficult to determine the correct orientation. However, by considering the non-intersecting road context, we can infer the backward orientation. Therefore, how to correctly identify such subtle orientation characteristics, while incorporating road context, poses another challenge.

To address these challenges, we devise AutoTS, an automated traffic sign update system designed for HD map updates, which jointly performs geo-localization and orientation classification of traffic signs. As shown in Figure 3, we first fine-tune a traffic sign detector to extract traffic signs from street images. For traffic sign localization, we estimate the depth between the camera and each detected traffic sign using an advanced self-supervised monocular depth estimation model By combining these depth estimates with the GPS locations of the image sequence, we obtain a set of location points for each traffic sign sequence. To derive the geo-location of the traffic sign from these noisy and sparse location points, we propose a noise and sparsity adaptive localization (NSAL) method. NSAL filters out the noisy location points based on the affinity matrix and tackles the sparsity among the remaining location points by enhancing their spatial density, which is achieved by assigning different weights to each location point. For traffic sign orientation identification, we design a position-aware orientation classification module, which constructs a position-aware SIFT feature along with the ROI feature to capture orientation characteristics of traffic signs from the image sequence. This module also incorporates road context information to assist in orientation classification. The main contributions are summarized as follows:

- We propose AutoTS, an automated traffic sign update system that jointly performs geo-localization and orientation classification of traffic signs, seeking to address a key challenge in HD map updates.
- We design a noise and sparsity adaptive localization method to calculate the geo-location of each traffic sign, which can filter noisy location points and is capable of deriving traffic sign geo-locations from sparse location points.
- To determine whether a detected traffic sign belongs to the road driven on, we propose a position-aware SIFT feature combined with the ROI feature to extract orientation characteristic from the traffic sign image sequence.
- We construct an open-source traffic sign localization and orientation classification benchmark, KITTI-TS, to facilitate the evaluation and comparison. Experiments conducted on KITTI-TS demonstrate the effectiveness of our AutoTS framework. The source code has been released to benefit the research community [1].

II. Related Work

Traffic Sign Detection. Since traffic signs convey crucial road information for driving assistance, many researchers in computer vision have studied the task of traffic sign detection [12]. Traffic sign detection methods can primarily be categorized into traditional [13]–[15] and deep learning-based methods [16]–[18]. Traditional methods typically leverage the physical characteristics of traffic signs, such as color and shape, for recognition. For example, Le *et al.* [13] conducted color detection and segmentation based on a Support Vector Machine (SVM) to retrieve candidate regions of traffic signs in videos. With the success of deep learning in computer vision, several deep learning-based traffic sign detection and recognition methods have been developed based on mainstream object detection models (e.g., Faster R-CNN [19], YOLO [20], and SSD [21]). For example, Wang *et al.* [16] proposed an improved feature pyramid model to boost traffic sign detection by modifying the feature pyramid network in YOLO. In our work, we fine-tune a pre-trained faster R-CNN to detect traffic signs in the image, and use the detection results to estimate the location and orientation of traffic signs.

Traffic Sign Localization. Accurate traffic sign localization is significant to autonomous driving and HD mapping. Given that GPS locations are usually available for the captured images, the core challenge of traffic sign localization lies in estimating the depth of the traffic sign relative to the camera. Early traffic sign localization efforts mainly use LiDAR [22] and stereo cameras [23] to obtain the depth between the traffic sign and the collection devices. For example, Doval *et al.* [23] detected traffic signs from RGB images and estimated their depth through stereo vision. Although these methods yield accurate results, the required sensors are expensive and impractical for large-scale deployment. To reduce hardware costs and improve scalability, some works focus on using monocular camera images to handle traffic sign localization [7], [9], [24]–[26]. For instance, Musa *et al.* [9] utilized SfM-based 3D reconstruction to estimate the depth between the traffic sign and the vehicle. Eisemann *et al.* [27] proposed a NeRF-based approach that reconstructs directional traffic signs from

[1] https://github.com/hanxjing/AutoTS.

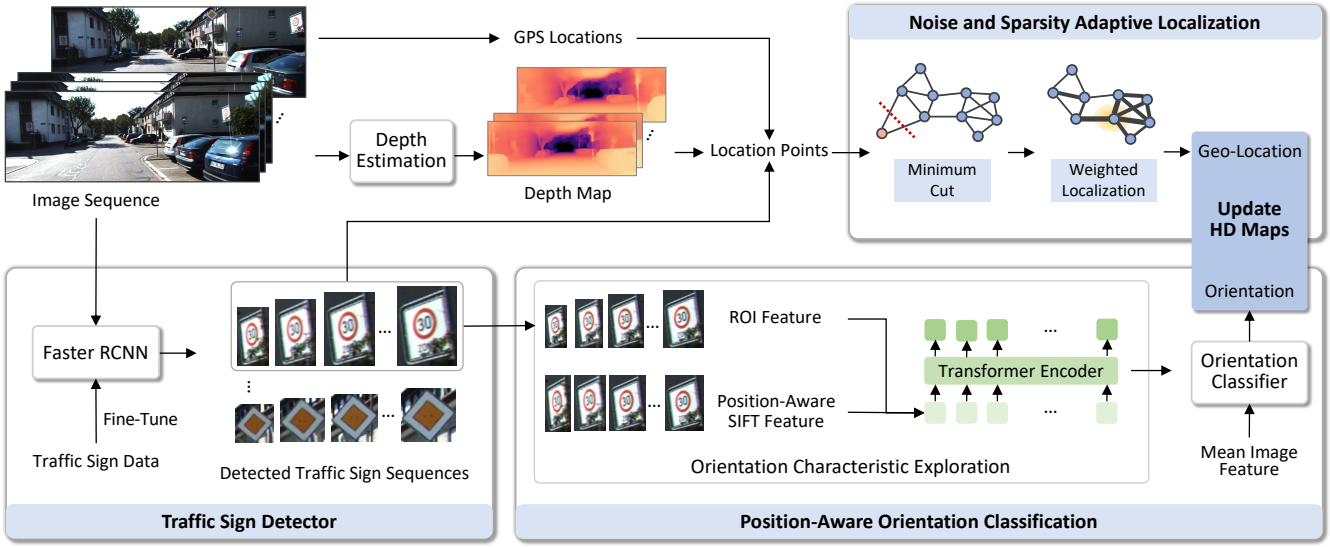


Fig. 3. The framework of our AutoTS comprises three components: 1) traffic sign detector, 2) noise and sparsity adaptive localization, and 3) position-aware orientation classification. We first detect multiple traffic sign sequences from an image sequence. For each detected traffic sign sequence, we extract the depth of the traffic sign in the image using a depth estimation model and obtain a set of location points. We then design a noise and sparsity adaptive localization method to derive the final traffic sign geo-location from these location points. We also extract ROI features and position-aware SIFT features to model the orientation characteristic, and combine the image features to for orientation classification. Based on the location and orientation, we update the traffic sign to HD maps.

monocular images, achieving accurate 3D localization and geometry estimation suitable for HD map updates. However, these approaches incur high computational costs. In addition, Pedersen *et al.* [7] adopted a thin-lens model [28] together with the real traffic sign height to calculate the depth between the traffic sign and the vehicle. Despite its promising performance, this method has limited application scenarios, since real traffic sign heights always vary on different types of roads and are not available for some regions. Yang *et al.* [26] proposed a vision-based pipeline that detects and localizes traffic signs using YOLOv4 and a depth estimation model, demonstrating accurate localization from a single image. In our work, we adopt an advanced self-supervised monocular depth estimation tool [29] to obtain the depth and update traffic signs to maps with not only the geo-location but also the orientation of traffic signs to determine whether it belongs to the road driven on. Note that although Pedersen *et al.* [7] defined the direction of a traffic sign as the azimuth angle of the road segment it faces, their goal was to use this direction to improve traffic sign localization, which differs from the orientation (*i.e.*, functional direction) considered in our work.

Graph-Based Clustering Methods. Clustering methods aim to group similar instances into cohesive clusters. Mainstream clustering methods include hierarchical methods [30], density-based methods [31], partitional methods [32], model-based methods [33], and graph-based methods [34], [35]. Graph-based clustering methods like spectral clustering [34] use an affinity matrix to link all data points, effectively handling density variations. Spectral clustering constructs a graph and forms clusters through a graph-cut algorithm. The goal is to have low-weight edges between different groups and high-weight edges within groups [36]. When the cluster sizes in spectral clustering are unconstrained, it may result in unbalanced minimum cuts, which can be exploited to identify the least relevant data points. Building on this, we develop a noise and sparsity adaptive localization method for traffic sign localization, using the minimum cut to filter noisy points and assign different weights to location points to reinforce the density relation among sparse location points. The final geo-location of the traffic sign is then extracted from the refined set of location points.

III. Methodology

In this section, we first formally define the research task, and then detail our AutoTS approach shown in Figure 3.

A. Problem Formulation

In our work, we aim to address the problem of traffic sign localization and orientation classification. Formally, suppose we have a set of traffic signs $\mathcal{S}$. For each traffic sign $s \in \mathcal{S}$, we have an image sequence $\mathcal{I}_s = \{I_i\}_{i=1}^{N_s}$, where N_s is the total number of the images. Each image sequence $\mathcal{I}_s$ is paired with a corresponding series of GPS locations $\mathcal{G}_s = \{g_i\}_{i=1}^{N_s}$. Besides, we annotate the ground truth geo-location l_s (*i.e.*, latitude and longitude), orientation $o_s \in \mathbf{O}$, and category c_s (*e.g.*, "Yield" and "Speed Limit") for each traffic sign s, where $\mathbf{O}$ is the discrete orientation category set. Thus, our goal is to learn a mapping function $\mathcal{F}$ defined as follows:

$$\mathcal{F} : (\mathcal{I}_s, \mathcal{G}_s) \rightarrow (l_s, o_s, c_s). \tag{1}$$

We can estimate the geo-location, orientation and category of traffic signs through $\mathcal{F}$, enabling us to update the traffic sign to HD maps.

Algorithm 1 Noise and Sparsity Adaptive Localization

Input: Location points $\mathcal{P}_s = \{p_1, p_2, \cdots, p_i\}_{i=1}^{N_s}$, threshold α
Output: Geo-location $\tilde{l}_s$

```
Calculate the affinity matrix W by Gaussian kernel
for all P' ⊆ P_s do
    if min Cut(P', P'') ≤ α then
        Remove smaller P' from P_s
    end if
end for
Get selected location point set P_j^*
for all p_i ∈ P_j^* do
    Calculate D[i,i] = Σ_{j∈N_s^*} W_ij
end for
Obtaining the geo-location l̃_s = Σ_{i=1}^{N_e} D̂[i,i] p_i
```

B. Traffic Sign Detection

We employ the pre-trained Faster R-CNN [19] to detect traffic signs in the image. To adapt the detector to our task, we fine-tune the Faster R-CNN on our constructed traffic sign dataset. Each can be represented with the category, bounding box b_i, and ROI feature v_i, which represent the visual content enclosed by b_i. These features are subsequently used for traffic sign localization and orientation classification.

It is worth noting that a single image may contain multiple traffic signs. When different traffic signs are detected within an image sequence, we identify each individual traffic sign based on the IoU (intersection over union) of bounding box b_i and the consistency of the traffic sign category in the image sequence. We then process each detected traffic sign sequence separately.

C. Image-level Location Point Extraction

Following previous work [7], we estimate the depth of the detected traffic sign in each image and integrate the image's GPS location to derive the location point of the traffic sign. In our work, considering the limited application scenarios of the thin-lens model and the high computational cost of the SfM method adopted in existing works [7], [9], we extract the depth by an advanced self-supervised monocular depth estimation model PlaneDepth [29], which provides accurate depth estimation.

For each image I_i in the sequence $\mathcal{I}_s$, we extract the depth d_i of the detected traffic sign by the self-supervised monocular depth estimation model PlaneDepth [29]. In particular, we first derive the depth map for the input image. We then extract the depth of traffic sign from the center of its bounding box b_i. The extracted depth is the perpendicular distance between the detected traffic sign and the vehicle that captured the image. Based on the detected traffic sign depth d_i and the GPS location g_i of the image I_i, we derive the location point p_i of the detected traffic sign as follows,

$$\begin{cases} \theta_i^t = \dfrac{FOV}{2} b_i^*, \\ p_i = \text{Geo}(\theta_i^t, \theta_i^c, d_i, g_i), \end{cases} \tag{2}$$

where θ_i^t is the yaw angle (azimuth) of the traffic sign to the camera, calculated by the camera horizontal FOV and the center deviation b_i^* of the bounding box. θ_i^c is the camera yaw angle in the world coordinate system. $\text{Geo}(\cdot)$ is used to calculate the location in the world coordinate system based on the trigonometric function. Finally, we obtain a set of location points $\mathcal{P}_s$ for the image sequence $\mathcal{I}_s$.

D. Noise and Sparsity Adaptive Localization

Based on the location point set $\mathcal{P}_s$, we can derive the final geo-location of the traffic sign. To mitigate the impact of noisy location points caused by depth estimation errors, and to address the sparsity of certain traffic sign samples caused by limited FOVs and poor lighting conditions, we develop the NSAL. Inspired by the classic spectral clustering [34], which employs an affinity matrix to establish relationships among the data and hence can handle sparse data, we also construct an affinity matrix to represent the location point set $\mathcal{P}_s$ as a graph. Different from the typical spectral clustering, which aims to partition data into comparable clusters, our NSAL is designed to discard noisy location points via a minimum cut and assign adaptive weights to the remaining points to enhance the spatial density relationships among the sparse samples.

Minimum Cut-based Noisy Point Removal. Following spectral clustering, we first employ the Gaussian kernel to build the affinity matrix for the location points of the detected traffic sign sequence as follows,

$$W_{ij} = \exp\left(-\frac{\|p_i - p_j\|^2}{2\sigma^2}\right), \tag{3}$$

where p_i and $p_j \in \mathcal{P}_s$. σ is the smoothing hyperparameter of the Gaussian kernel. The affinity matrix connects all location points as a graph. As the minimum cut in spectral clustering usually results in uneven clustering by finding the most irrelevant points [36] (as shown in Figure 3), we use this characteristic to remove the noisy points from $\mathcal{P}_s$. The graph cut in spectral clustering is defined as follows,

$$Cut(\mathcal{P}', \mathcal{P}'') = \sum_{i \in \mathcal{P}', j \in \mathcal{P}''} W_{ij}, \tag{4}$$

where $\mathcal{P}'$ and $\mathcal{P}''$ are two complementary subgraphs of $\mathcal{P}_s$. The value of $Cut(\cdot)$ is directly proportional to the correlation between subgraphs. To identify the noisy location points, we calculate the minimum cut, *i.e.*, the smallest $Cut(\cdot)$, among all the location points. We consider the smaller subset from the minimum cut to be the set of noisy location points. Since the number of noisy location points is unknown, we traverse all possible cuts and discard the smaller subsets where $Cut(\cdot) \leq \alpha$, with α being a threshold hyperparameter. This process results in the final selected location point set $\mathcal{P}_s^*$.

Weighted Geo-location Extraction. Based on the selected location points $\mathcal{P}_s^*$, we then extract the final geo-location of the traffic sign. Since the location points can be sparse in some cases, directly computing their geometric center may lead to unreliable results. To address this, we enhance the contribution of location points that are more closely connected to their neighbors, as they are generally more reliable in sparse

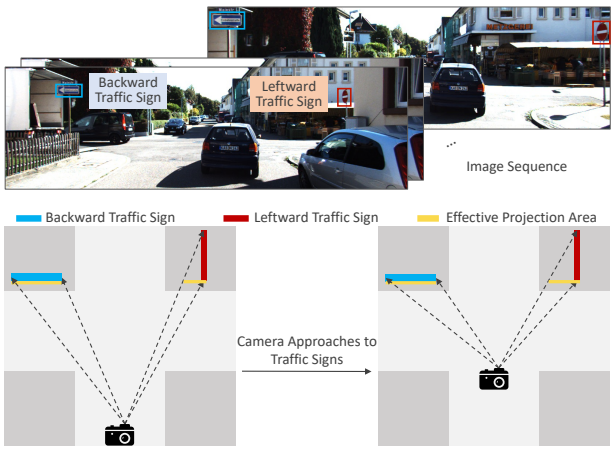


Fig. 4. As the camera approaches the traffic signs, the effective projection area of the leftward traffic sign changes, whereas that of the backward traffic sign remains fixed.

scenarios. Specifically, we assign different weights to each location point based on the affinity matrix, which reinforces the spatial density relationships among the sparse points. We first compute the degree matrix of the location points as follows:

$$D[i,i] = \sum_{j=1}^{N_s^*} W_{ij}, \tag{5}$$

where N_s^* is the total number of the selected location points. The degree matrix reflects the connection of each point to other points. We deem that compared to points with smaller degree values, a point with a larger degree value should have more surrounding points when the data is dense. Therefore, we weight the points with their degrees and derive the final traffic sign geo-location $\tilde{l}_s$, which is defined as follows,

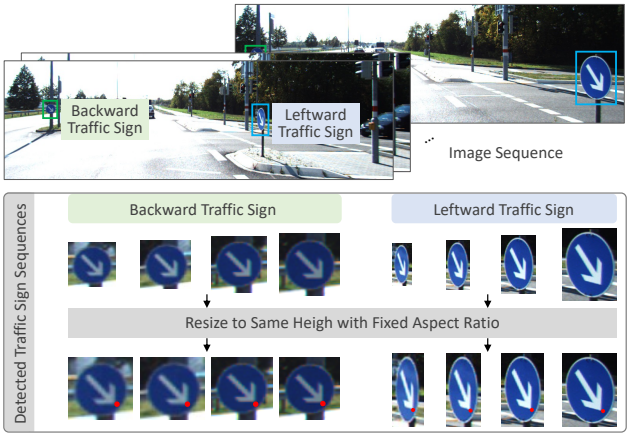
$$\tilde{l}_s = \sum_{i=1}^{N_e} \hat{D}[i,i] p_i, \tag{6}$$

where $\hat{D}_i$ is value of the normalized degree matrix and $||\hat{D}_i|| = 1$. The overall workflow of our noise and sparsity adaptive localization method is shown in Algorithm 1.

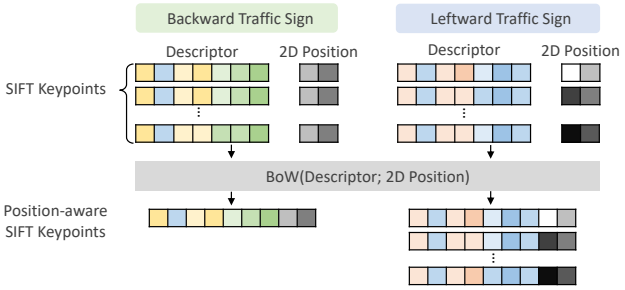
E. Position-Aware Orientation Classification

To identify the orientation of a traffic sign, we explore its orientation characteristics in the image sequence and utilize the ROI feature as well as the designed position-aware SIFT feature to represent these characteristics. As the orientation of a traffic sign can be affected by the road context (*e.g.*, if the road driven on is intersecting), we also incorporate the road context to ensure a comprehensive orientation classification.

Orientation Characteristic Exploration. We observe that the orientation characteristics of traffic signs with different orientations can be effectively captured from the detected traffic sign sequence. As shown in Figure 4, when the vehicle (*i.e.*, camera) approaches a traffic sign, the effective projected area of the leftward or rightward traffic sign changes significantly, while that of the backward traffic sign is fixed. As illustrated in Figure 5 (a), changes in the projected area lead



(a) Shape change in leftward traffic sign sequence is more obvious than that of backward traffic sign sequence.



(b) 2D positions diversify same descriptors of leftward traffic sign.

Fig. 5. Shape changes and position-aware SIFT keypoints in detected traffic sign sequences.

to shape changes in leftward and rightward traffic signs across the image sequence. In particular, the shape changes of the leftward traffic signs are obvious and become more apparent after resizing the detected signs to a uniform height while maintaining a fixed aspect ratio. Therefore, we treat the shape change in the traffic sign sequence as a key orientation cue. This characteristic can be represented by the ROI feature v_i extracted by the traffic sign detector.

Position-Aware SIFT Feature. In addition to the ROI feature, we also use SIFT keypoints [37] to reinforce the orientation characteristic. As shown in Figure 5 (a), we highlight the same keypoints with red dots in the detected traffic sign sequence. The same keypoints in the detected backward traffic signs have similar 2D positions in the detected area, while those in the detected leftward traffic signs have different 2D positions in the detected area. This positional variation serves as a valuable cue for distinguishing orientations. To exploit this characteristic, we first resize each detected traffic sign to a fixed height while maintaining its aspect ratio. We then extract a SIFT keypoint set $\mathcal{K}_i = \{des_1, des_2, \cdots, des_{N_i}\}$ from the resized traffic sign, where des_{N_i} is the descriptor of each keypoint. We concatenate each descriptor with its 2D position (x, y) in the detected area to obtain a set of position-aware SIFT keypoints $\tilde{\mathcal{K}}_i$ as follows,

$$\tilde{\mathcal{K}}_i = \{\tilde{des} | [des; x; y]\}, \tag{7}$$

where $\tilde{des}$ denotes the position-aware SIFT keypoint. In leftward/rightward traffic sign sequences, the same keypoints may

have differing 2D positions. As shown in Figure 5 (b), to represent this characteristic, we aggregate all position-aware SIFT keypoints across the sequence and remove redundant ones (*i.e.*, same keypoints with same 2D positions) by conducting the Bag of Word (BoW) [38] method as follows,

$$\mathcal{E} = \text{BoW}(\tilde{\mathcal{K}}_1, \tilde{\mathcal{K}}_2, \cdots, \tilde{\mathcal{K}}_i), i = 1, \ldots, N_s, \tag{8}$$

where $\mathcal{E} = \{\tilde{des}_i\}_i^{N_k}$ is a unique set of total position-aware SIFT keypoints of a detected traffic sign sequence and N_k is the number of unique position-aware SIFT keypoints. Based on $\mathcal{E}$, we represent each detected traffic sign in the sequence with a binary position-aware SIFT feature $e_i \in \mathcal{R}^{N_k}$ as follows,

$$e_i = f_e(\mathcal{E}, \tilde{\mathcal{K}}_i), i = 1, \ldots, N_s, \tag{9}$$

where $f_e(\cdot)$ finds the most similar keypoint in $\mathcal{E}$ for each keypoint in $\tilde{\mathcal{K}}_i$. e_i indicates the presence of each keypoint in $\mathcal{E}$ in $\tilde{\mathcal{K}}_i$. Since the 2D position helps diversify keypoints in leftward/rightward cases, the backward traffic sign can be classified based on the similarity of e_i among the traffic sign sequence.

Comprehensive Orientation Classification. Due to the capability of the self-attention mechanism of Transformer [39] models in capturing the context information in a sequence, we adopt a Transformer as our encoder to model the orientation characteristics among the sequence with both the ROI feature v_i and position-aware SIFT feature e_i. To take the road semantics into account, we also employ the mean image feature as part of the orientation classification, which is defined as follows,

$$\begin{cases} h_s = \text{Transf}(v_i; e_i), \\ \tilde{o}_s = f_o(h_s; x_s), \end{cases} \tag{10}$$

where $\text{Transf}(\cdot)$ denotes the Transformer encoder and h_s is the output orientation hidden vector. x_s is the mean image feature obtained by the mean of all the image features in the sequence. $f_o(\cdot)$ is the orientation decoder, which consists of a fully connected layer. $\tilde{o}_s$ is the predicted orientation.

We optimize the orientation classification with a cross-entropy loss function as follows,

$$\mathcal{L}_{wce} = \sum_{i=1}^{|\mathcal{S}|} w_{o_s} \cdot \log\left(\frac{\exp(\tilde{o}_s^*)}{\sum \exp(\tilde{o}_s)}\right), \tag{11}$$

where $\tilde{o}_s^*$ the predicted logit corresponding to the ground-truth orientation of traffic sign s. w_{o_s} is the weight assigned to the ground-truth orientation class of traffic sign s, which is set inversely proportional to the class frequency to mitigate the impact of class imbalance.

IV. KITTI-TS Dataset

To facilitate the evaluation and comparison of the proposed method, we construct the KITTI-TS benchmark based on the widely used KITTI dataset [40], which serves as a standard benchmark for depth estimation [41]–[43] and monocular 3D object detection [44], [45]. The KITTI dataset includes a variety of urban scenarios that reflect real-world complexities in autonomous driving, including data captured from various road types with a wide range of traffic signs. Furthermore, the collection vehicle of the KITTI dataset is equipped with a camera, GNNS, and a laser scanner, providing the necessary multi-modal information to support the development and validation of traffic sign localization methods.

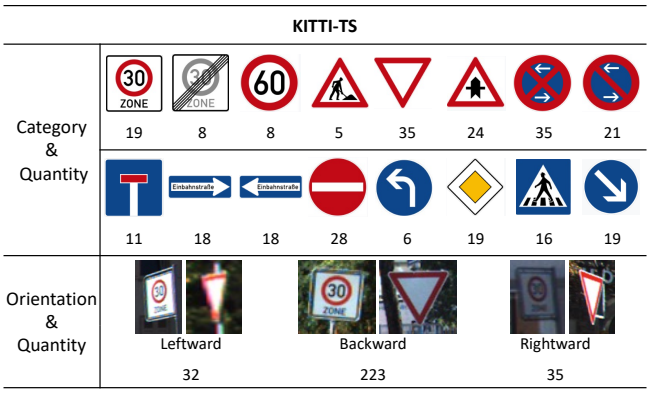


Fig. 6. The traffic sign categories, orientations and their quantities in the KITTI-TS dataset.

TABLE I
PERFORMANCE OF TRAFFIC SIGN DETECTOR IN TERMS OF COCO METRICS (%) ON THE KITTI-TS DATASET.

Method	AP	AP50	AP75	APs	APm
Traffic sign detector	60.08	84.27	72.78	51.91	67.61

To construct the KITTI-TS benchmark, we first annotate traffic signs within the raw KITTI dataset. Traffic sign categories containing fewer than five instances are discarded to ensure sufficient representation. Next, we extract the depth of each traffic sign using the LiDAR point cloud data provided by the laser scanner onboard the KITTI data collection vehicle. By combining this depth information with the camera's orientation and GPS data, we compute the ground-truth geo-location of each traffic sign. Following this, we annotate the orientation of each traffic sign. In the real world, traffic signs may face arbitrary directions. However, as we aim to determine whether the traffic sign belongs to the road driven on, we simplify the orientation by categorizing it into discrete values (*i.e.*, leftward, backward, rightward). We define backward traffic signs as belonging to the road driven on.

Finally, we obtain 290 individual traffic signs from 16 different categories, with 3,062 bounding boxes and location points among 2,387 images. The detailed traffic sign categories, orientations, and their quantities are shown in Figure 6.

V. Experiment

A. Experiment Settings

Datasets. We present experimental results on the **KITTI-TS** dataset, which we partition into a training set (80%) and a testing set (20%). To further evaluate the effectiveness of our noise- and sparsity-adaptive localization (NSAL) method in the traffic sign localization task, we conduct additional experiments on the **Aalborg** dataset [7], [8], which contains 277 individual traffic signs and 7,828 associated location points. As the Aalborg dataset [7] is designed to evaluate

clustering-based localization approaches, it provides only the estimated location points and the ground-truth geo-location of each traffic sign, without the corresponding image-level detections. Therefore, we use only the location points in this dataset to validate our NSAL method. Additionally, Aalborg dataset includes the direction from each traffic sign to the road segment (*i.e.*, location of image-collecting vehicle), which can be leveraged to improve the localization task. We follow the same dataset settings as used in prior work [7].

Evaluation Metrics. For the traffic sign localization, following previous work [7], we adopt the MAE (mean absolute error) and RMSE (Root mean squared error) to evaluate the localization results. Besides, we also utilize the Recall@K (R@K) to evaluate the proportion of traffic signs with localization error of less than K meters, where $K = \{1, 2\}$. For orientation classification, we adopt two evaluation metrics: **1)** Accuracy, which measures the proportion of correctly classified traffic signs among all samples, and **2)** mean Recall (mRecall), which computes the average recall across all orientation classes. Due to the imbalanced distribution of orientation labels, the model tends to be biased toward backward orientations. Therefore, mRecall is used to more comprehensively assess both the class-wise balance and overall effectiveness of the orientation classification. We also list the recall of each orientation class to gain more detailed classification performance.

Implementation Details. In the traffic sign detector, we adopt a Faster R-CNN [19] with ResNeXt-101-FPN pre-trained on the COCO dataset [46]. We fine-tune the Faster R-CNN on our KITTI-TS dataset with batch size 12 and 50K iterations to obtain the traffic sign detector. The detection results in terms of the COCO evaluation metrics are shown in Table I. The COCO evaluation metrics are provided in Detectron2 [47] to evaluate the detector, including Average Precision (AP), Average Precision at different IoU thresholds (AP50 and AP75), Average Precision for small objects (APs) and Average Precision for median objects (APm). Average Precision (AP) is calculated as the area under the precision-recall curve, averaged over IoU thresholds from 0.50 to 0.95 (with a step of 0.05), reflecting the trade-off between precision and recall. Note that the ground-truth bounding boxes are manually annotated and may contain slight inaccuracies, which can limit the maximum achievable AP. Note that as one image may contain multiple traffic signs, in the inference phase, we consider that bounding boxes with same categories and IoU less than 0.2 in adjacent images belong to the same traffic sign.

In the noise and sparsity adaptive localization, we utilize the provided checkpoint of PlaneDepth [29] to estimate the depth between the camera and the traffic sign, which achieves promising estimation results on KITTI without depth supervision. The smoothing hyperparameter σ in NSAL is set to 2.5. The threshold α in NSAL is set to 0.1.

In the position-aware orientation classification, due to the unbalanced orientation data distribution, where the backward traffic signs account for 77.4% of all samples, the predicted orientation biases to the backward category. We hence adopt the weighted cross-entropy loss to balance the biased prediction. The weight of each category is assigned according to the inverse class frequency. We employ the Adam optimizer, and each training lasts for 300 steps. We set the batch size and initial learning rate to 8 and 0.0008, respectively.

TABLE II
PERFORMANCE COMPARISON OF TRAFFIC SIGN LOCALIZATION REGARDING MAE (M), RMSE (M), R@1M (%) AND R@2M(%) ON THE KITTI-TS AND AALBORG DATASETS. THE BEST RESULTS IN EACH DATASET ARE HIGHLIGHTED IN BOLD.

Methods	MAE ↓	RMSE ↓	R@1m ↑	R@3m ↑
KITTI-TS Dataset				
GeoLocating [7]	3.98	6.27	11.86	22.03
GeoLocating+NSAL	3.80	5.92	15.25	27.12
AutoTS (ours)	**2.38**	**3.42**	**30.51**	**54.23**
Aalborg Dataset				
GeoLocating [7]	5.88	7.07	3.57	8.93
GeoLocating+NSAL	**5.34**	**6.49**	**5.36**	**10.71**
GeoLocating-D [7]	5.12	5.92	7.14	12.50
GeoLocating-D+NSAL	**4.90**	**5.65**	**10.71**	**16.07**

B. Comparison on Localization

In this section, we first compare our AutoTS method with the existing traffic sign localization method [7] to evaluate the ability of geo-locating traffic signs. We then conduct an ablation study of our NSAL to evaluate the effectiveness of its different components. Note that the experimental results in localization are derived from all the samples, rather than just the testing set.

Comparison with Existing Works. Given that GeoLocating [7] shares a similar application scenario with our method and provides publicly available code and dataset, we adopt it as the baseline for comparison on the KITTI-TS and Aalborg datasets. GeoLocating employs the thin-lens model [28] to estimate the depth from the camera to each traffic sign, generating a set of location points. It then clusters these points for each traffic sign category using the DBSCAN algorithm and iteratively searches for optimal hyperparameters. For a fair comparison, we implement GeoLocating on the KITTI-TS dataset and search for the optimal hyperparameters to obtain geolocations of traffic signs.

Since the Aalborg dataset provides only estimated location points, we only apply our NSAL method to these points to evaluate its effectiveness. Specifically, based on the GeoLocating-derived clusters of these location points, we further apply our NSAL method to generate more accurate geolocations from the location points. This pipeline is denoted as GeoLocating+NSAL, combining GeoLocating's depth estimation with our NSAL refinement. Moreover, since the Aalborg dataset also includes annotated direction of each location point to the image-collecting vehicle, we further incorporate the direction information to refine the geolocation of traffic signs. This variant is denoted as GeoLocating-D.

The results are presented in Table II, based on traffic signs detected by our traffic sign detector. We can observe that: **1)** On the KITTI-TS dataset, our AutoTS clearly outperforms GeoLocating, demonstrating the superiority of our method in geo-locating traffic signs. **2)** On the KITTI-TS dataset, our AutoTS has better performance than GeoLocating+NSAL, highlighting the effectiveness of our selected self-supervised

TABLE III
ABLATION STUDY OF TRAFFIC SIGN LOCALIZATION REGARDING MAE (M), RMSE (M), R@1M (%) AND R@2M(%) ON THE KITTI-TS DATASET. THE * SIGN REPRESENTS THAT THE EXPERIMENTS CONDUCT WITH SPARSE DATA. THE BEST RESULTS IN ALL THE DATA ARE HIGHLIGHTED IN BOLD.

Methods	MAE↓	RMSE↓	Recall-1m↑	Recall-2m↑
AutoTS w/o MinCut	2.51	3.74	27.11	50.87
AutoTS w/o Weight	2.49	3.72	28.81	49.15
AutoTS-K-Means	2.56	3.82	27.11	47.46
AutoTS-DBSCAN	2.48	3.79	28.81	52.54
AutoTS (ours)	**2.38**	**3.42**	**30.51**	**54.23**
AutoTS-K-Means*	2.46	3.52	28.81	52.54
AutoTS-DBSCAN*	2.42	3.48	30.51	52.54
AutoTS*	2.36	3.37	32.20	54.23

TABLE IV
ABLATION STUDY OF POSITION-AWARE ORIENTATION CLASSIFICATION IN TERMS OF ACCURACY (%) AND MRECALL (%) ON THE KITTI-TS DATASET. † DENOTES THE RESULT WITH THE GROUND TRUTH TRAFFIC SIGN BOUNDING BOXES. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Methods	Accuracy	Recall of Each Category			mRecall
		Left	Back	Right	
AutoTS w/ ROI	71.43	12.50	91.30	22.22	42.01
AutoTS w/ ROIS	73.02	37.50	84.78	44.44	55.57
AutoTS w/o SIFT	76.19	37.50	89.13	44.44	62.38
AutoTS w/o MImg	80.95	50.00	89.13	**66.67**	68.60
AutoTS w/ BiLSTMs	82.54	62.50	91.30	55.56	69.79
AutoTS w/ LSTMs	82.54	50.00	93.48	55.56	67.07
AutoTS (ours)	85.71	50.00	**95.65**	**66.67**	70.77
AutoTS†	**88.89**	**75.00**	**95.65**	**66.67**	**79.11**

monocular depth estimation model over the thin-lens model for depth extraction in this task. The high-quality location points further enhance traffic sign localization performance. **3)** On both the KITTI-TS and Aalborg datasets, GeoLocating+NSAL achieves better performance than GeoLocating. This demonstrate our NSAL method can assist in geo-locating traffic signs. Although the improvement brought by NSAL is relatively modest, it is specifically designed to handle challenges such as noise and sparsity. The performance of NSAL may also be limited by the quality of the location points. And **4)** With the inclusion of direction information, both GeoLocating-D and GeoLocating-D+NSAL outperform GeoLocating and GeoLocating+NSAL, respectively, demonstrating the effectiveness of direction information in the Aalborg dataset.

Ablation Study of NSAL Method. To thoroughly investigate our NSAL, we introduce the following baseline methods:

- AutoTS w/o MinCut: We disable the minimum cut from our NSAL to analyze its effect.
- AutoTS w/o Weight: To investigate the impact of the weighted cluster center extracting, we replace it with the K-Means clustering method to derive the final geo-location of traffic signs.
- AutoTS-K-Means: We use the intrinsic cluster center calculation method to derive the geo-location from the location points to verify the effectiveness of our NSAL method.
- AutoTS-DBSCAN: To compare our NSAL with DBSCAN, a clustering method capable of handling noise points, we directly utilize the DBSCAN clustering method to derive the geo-location.

We further conducted experiments on samples with sparse location points to validate our method's ability to handle sparse data. Specifically, we regard traffic signs with fewer than 10 location points as sparse samples. The results are shown in Table III, from which we observe that: **1)** Our AutoTS outperforms both AutoTS w/o MinCut and AutoTS w/o Weight, indicating the effectiveness of these two components in our NSAL. The minimum cut helps eliminate noisy or inconsistent location points by preserving the most spatially coherent cluster. The weighted clustering further refines the estimated location by emphasizing high-confidence points, leading to more robust localization results. **2)** Our AutoTS outperforms both AutoTS-K-Means and AutoTS-DBSCAN, demonstrating the advantage of our NSAL in handling noisy and sparse location points compared to K-Means and DBSCAN. Finally, **3)** when dealing with sparse data, our AutoTS* performs better than both AutoTS-K-Means* and AutoTS-DBSCAN*. This shows the improved sparse data processing ability of our NSAL.

C. Comparison on Orientation Classification

Since there are no existing efforts focusing on the task of traffic sign orientation classification, we explore the effectiveness of different features in our position-aware orientation classification framework and investigate the performance of different networks in capturing orientation characteristics. To this end, we introduce the following baseline methods:

- AutoTS w/ ROI: To examine the effect of the image sequence in orientation classification, we solely utilize the ROI feature of a single traffic sign for orientation classification.
- AutoTS w/ ROIS: To compare with AutoTS w/ ROI, we employ the ROI feature of the detected traffic sign sequence to classify its orientation.
- AutoTS w/o SIFT: To verify the effect of the position-aware SIFT feature, we remove it from the input of orientation classification.
- AutoTS w/o MImg: To investigate the effect of the mean image feature, we remove it from our AutoTS to classify the orientation.
- AutoTS w/ BiLSTMs and AutoTS w/ LSTMs: Given that both BiLSTMs [48] and LSTMs [49] are advanced sequence information processing methods, we substitute the Transformer encoder in our AutoTS with BiLSTMs [48] and AutoTS w/ LSTMs [49] to model the orientation characteristic.

The results are shown in Table IV. The results of AutoTS are based on traffic signs detected by our traffic sign detector. We find that: **1)** AutoTS w/ ROIS achieves a better accuracy and mRecall results than AutoTS w/ ROI, demonstrating the necessity of incorporating image sequences for effective orientation classification. Furthermore, AutoTS w/ ROI, which relies on a single traffic sign, performs relatively poorly on the leftward and rightward categories due to its limited capacity to capture orientation-specific features. In contrast, AutoTS w/ ROIS, which leverages a sequence of detected traffic signs, is better able to extract orientation cues, resulting in more

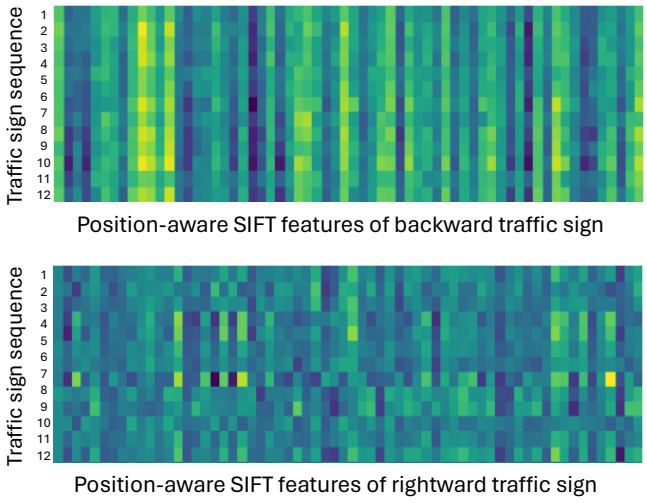


Fig. 7. Position-aware SIFT features of traffic sign with different orientations.

balanced and accurate classification across all three orientation categories. **2)** AutoTS outperforms AutoTS w/o SIFT in both Recall and mRecall, highlighting the effectiveness of the position-aware SIFT feature in capturing orientation-relevant information. To gain deeper insight, we visualize the position-aware SIFT features from the test set in Fig. 7. The vertical stripes in backward traffic sign sequence are more obvious than those in rightward traffic sign sequence, indicating a stronger similarity among the backward traffic sign sequence. This is instrumental in distinguishing the orientation in traffic sign sequences. **3)** AutoTS achieves superior Recall and mRecall over AutoTS w/o MImg, suggesting that road semantics can facilitate traffic sign orientation classification. **4)** AutoTS with the Transformer orientation encoder has better performance over AutoTS w/ BiLSTMs and AutoTS w/ LSTMs in terms of Recall and mRecall. This underscores the Transformer's superior capability in capturing and modeling sequence information. And **5)** AutoTS† outperforms AutoTS, demonstrating that errors in the traffic sign detector have a slight impact on the downstream orientation classification task.

Orientation classification is crucial for HD map updates and other downstream geo-related tasks. Existing HD map traffic sign update methods [50] primarily rely on object detection and localization. Assuming our KITTI-TS dataset is used in a map update scenario, our traffic sign detector identifies 63 distinct traffic signs in the test set, among which 46 are of backward orientation (*i.e.*, belong to the road driven on). If all detected traffic signs were naively added to the map without considering their orientations, the theoretical map update accuracy would be 73.01%. However, with the orientation classification module, certain leftward and rightward traffic signs can be correctly filtered out, increasing the theoretical map update accuracy to 89.58%, as calculated based on the results of AutoTS† in Table IV. This result highlights the effectiveness and necessity of accurate orientation classification in real-world HD mapping systems.

D. Qualitative Results

We visualize the qualitative results of AutoTS w/o MinCut, AutoTS w/o Weight, and AutoTS (ours) on selected examples in Figure 8. The observations are as follows: **1)** AutoTS outperforms AutoTS w/o MinCut, as the incorporation of the minimum cut effectively suppresses the influence of noisy location points, leading to improved localization accuracy. However, in the second example, both AutoTS and AutoTS w/o MinCut yield similar localization results. One possible reason is that the lower vegetation coverage in this scenario likely leads to more stable GPS signals, thus even without the minimum cut, AutoTS w/o MinCut still achieves satisfactory localization results. These examples highlight the benefit of the minimum cut in scenarios with degraded GPS quality. **2)** In the third example, AutoTS demonstrates superior localization performance compared to AutoTS w/o Weight. Due to the sparse distribution of traffic sign location points, the weighted clustering method in AutoTS reinforces the spatial density relationships, leading to more accurate localization. In contrast, the fourth example features a denser set of location points, resulting in comparable performance between AutoTS and AutoTS w/o Weight. Finally, **3)** regarding orientation classification, our AutoTS correctly classifies the orientation in the first three examples, but fails in the last one. Although the last traffic sign is backward, it is placed along a road corner with an ambiguous orientation, making it challenging for AutoTS to infer the correct orientation. This suggests that our position-aware orientation classification method may face limitations in scenarios where traffic signs are not optimally positioned and road context cannot offer valuable information.

VI. Conclusion

In this work, we present AutoTS, an end-to-end automated traffic sign update system, capable of estimating the geo-location and classifying the orientation of traffic signs from the image sequence. To localize a traffic sign, we integrate a self-supervised monocular depth estimation model to estimate the depth of detected traffic signs and design the NSAL method to infer geo-locations from multiple location points. To determine whether a detected sign belongs to the current driving road, we introduce a position-aware orientation classification approach. This method not only captures orientation characteristic across the image sequence but also incorporates road context to enhance orientation classification accuracy. Moreover, we construct KITTI-TS dataset, an open-source benchmark for traffic sign localization and orientation classification, to facilitate the evaluation of the proposed AutoTS framework. Extensive experimental results validate the effectiveness of our approach. As future work, we plan to extend the system to include additional types of road information, moving toward a more comprehensive and fully automated HD map update pipeline.

Acknowledgments

This research is supported by Singapore Ministry of Education Academic Research Fund Tier 2 under MOE's official grant number T2EP20221-0023 and Grab-NUS AI Lab, a joint collaboration between GrabTaxi Holdings Pte. Ltd. and National University of Singapore.

Fig. 8. Examples of traffic sign localization orientation classification results generated by AutoTS w/o MinCut, AutoTS w/o Weight, and our AutoTS. We also display the traffic signs detected in the image sequence along with their ground truth geo-location and orientation. The points shown on the map correspond to the location points of the detected traffic sign sequence. Points depicted in gray represent those discarded through the minimum cut process.

REFERENCES

[1] S. Teng, X. Hu, P. Deng, B. Li, Y. Li, Y. Ai, D. Yang, L. Li, Z. Xuanyuan, F. Zhu *et al.*, "Motion planning for autonomous driving: The state of the art and future perspectives," *IEEE Transactions on Intelligent Vehicles*, 2023.

[2] Z. Ma, Z. Zheng, J. Wei, Y. Yang, and H. T. Shen, "Instance-dictionary learning for open-world object detection in autonomous driving scenarios," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[3] L. Li, M. A. Cheema, H. Lu, M. E. Ali, and A. N. Toosi, "Comparing alternative route planning techniques: A comparative user study on melbourne, dhaka and copenhagen road networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 11, pp. 5552–5557, 2021.

[4] J. Huang, Z. Huang, X. Fang, S. Feng, X. Chen, J. Liu, H. Yuan, and H. Wang, "Dueta: Traffic congestion propagation pattern modeling via efficient graph learning for eta prediction at baidu maps," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 3172–3181.

[5] J. Song, J. Son, D.-h. Seo, K. Han, N. Kim, and S.-W. Kim, "St-gat: A spatio-temporal graph attention network for accurate traffic speed prediction," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 4500–4504.

[6] D. A. Tedjopurnomo, Z. Bao, B. Zheng, F. M. Choudhury, and A. K. Qin, "A survey on modern deep neural network for traffic prediction: Trends, methods and challenges," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 4, pp. 1544–1561, 2020.

[7] K. F. Pedersen and K. Torp, "Geolocating traffic signs using large imagery datasets," in *17th International Symposium on Spatial and Temporal Databases*, 2021, pp. 34–43.

[8] Pedersen, Kasper F and Torp, Kristian, "Geolocating traffic signs using crowd-sourced imagery," in *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, 2020, pp. 199–202.

[9] A. Musa, "Multi-view traffic sign localization with high absolute accuracy in real-time at the edge," in *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, 2022, pp. 1–12.

[10] A. Mehrish, P. Singh, P. Jain, A. V. Subramanyam, and M. Kankanhalli, "Egocentric analysis of dash-cam videos for vehicle forensics," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 9, pp. 3000–3014, 2019.

[11] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113.

[12] M. Swathi and K. Suresh, "Automatic traffic sign detection and recognition: A review," in *2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET)*. IEEE, 2017, pp. 1–6.

[13] T. T. Le, S. T. Tran, S. Mita, and T. D. Nguyen, "Real time traffic sign detection using color and shape-based features." in *ACIIDS (2)*, 2010, pp. 268–278.

[14] G. Wang, G. Ren, Z. Wu, Y. Zhao, and L. Jiang, "A robust, coarse-to-fine traffic sign detection method," in *The 2013 international joint conference on neural networks (IJCNN)*. IEEE, 2013, pp. 1–5.

[15] W. Liu, S. Li, J. Lv, B. Yu, T. Zhou, H. Yuan, and H. Zhao, "Real-time traffic light recognition based on smartphone platforms," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 5, pp. 1118–1131, 2016.

[16] J. Wang, Y. Chen, Z. Dong, and M. Gao, "Improved yolov5 network for real-time multi-scale traffic sign detection," *Neural Computing and Applications*, pp. 1–13, 2022.

[17] C. Deng, M. Wang, L. Liu, Y. Liu, and Y. Jiang, "Extended feature pyramid network for small object detection," *IEEE Transactions on Multimedia*, vol. 24, pp. 1968–1979, 2021.

[18] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, "Traffic-sign detection and classification in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2110–2118.

[19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[21] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 21–37.

[22] F. Ghallabi, G. El-Haj-Shhade, M.-A. Mittet, and F. Nashashibi, "Lidar-based road signs detection for vehicle localization in an hd map," in *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019, pp. 1484–1490.

[23] G. N. Doval, A. Al-Kaff, J. Beltrán, F. G. Fernández, and G. F. López, "Traffic sign detection and 3d localization via deep convolutional neural networks and stereo vision," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 1411–1416.

[24] A. Stoven-Dubois, K. K. Miguel, A. Dziri, B. Leroy, and R. Chapuis, "A collaborative framework for high-definition mapping," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 1845–1850.

[25] P. Cheng, Y. Zhao, and W. Wang, "Detect arbitrary-shaped text via adaptive thresholding and localization quality estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 12, pp. 7480–7490, 2023.

[26] Z. Yang, H. Maeda, C. Zhao, G. Sato, and Y. Sekimoto, "Vision-based traffic sign detection and localization in tokyo metropolitan area," in *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2022, pp. 316–321.

[27] L. Eisemann and J. Maucher, "A nerf-based approach for monocular traffic sign reconstruction and localization," in *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2024, pp. 1664–1671.

[28] J. Fleck and J. R. Morris, "Equivalent thin lens model for thermal blooming compensation," *Applied Optics*, vol. 17, no. 16, pp. 2575–2579, 1978.

[29] R. Wang, Z. Yu, and S. Gao, "Planedepth: Self-supervised depth estimation via orthogonal planes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[30] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: an overview," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 1, pp. 86–97, 2012.

[31] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "Dbscan revisited, revisited: why and how you should (still) use dbscan," *ACM Transactions on Database Systems (TODS)*, vol. 42, no. 3, pp. 1–21, 2017.

[32] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the royal statistical society. series c (applied statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[33] M.-S. Yang, C.-Y. Lai, and C.-Y. Lin, "A robust em clustering algorithm for gaussian mixture models," *Pattern Recognition*, vol. 45, no. 11, pp. 3950–3961, 2012.

[34] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *Advances in neural information processing systems*, vol. 14, 2001.

[35] M. Maier, U. Luxburg, and M. Hein, "Influence of graph construction on graph-based clustering measures," *Advances in neural information processing systems*, vol. 21, 2008.

[36] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, pp. 395–416, 2007.

[37] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, pp. 91–110, 2004.

[38] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: a statistical framework," *International journal of machine learning and cybernetics*, vol. 1, pp. 43–52, 2010.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[40] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.

[41] Y. Cao, T. Zhao, K. Xian, C. Shen, Z. Cao, and S. Xu, "Monocular depth estimation with augmented ordinal depth relationships," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 8, pp. 2674–2682, 2019.

[42] M. Song, S. Lim, and W. Kim, "Monocular depth estimation using laplacian pyramid-based depth residuals," *IEEE transactions on circuits and systems for video technology*, vol. 31, no. 11, pp. 4381–4393, 2021.

[43] C. Feng, Z. Chen, C. Zhang, W. Hu, B. Li, and F. Lu, "Iterdepth: Iterative residual refinement for outdoor self-supervised multi-frame monocular depth estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 1, pp. 329–341, 2023.

[44] C. Tao, J. Cao, C. Wang, Z. Zhang, and Z. Gao, "Pseudo-mono for monocular 3d object detection in autonomous driving," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 8, pp. 3962–3975, 2023.

[45] L. Yang, X. Zhang, J. Li, L. Wang, M. Zhu, C. Zhang, and H. Liu, "Mix-teaching: A simple, unified and effective semi-supervised learning framework for monocular 3d object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 11, pp. 6832–6844, 2023.

[46] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[47] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[48] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[49] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[50] B. Wijaya, K. Jiang, M. Yang, T. Wen, Y. Wang, X. Tang, Z. Fu, T. Zhou, and D. Yang, "High definition map mapping and update: A general overview and future directions," *arXiv preprint arXiv:2409.09726*, 2024.

Xianjing Han received the B.E. degree from the Northeastern University, China, in 2017, and the Ph.D. degree from the School of Computer Science and Technology, Shandong University, China, in 2022. She is currently a research fellow with the School of Computing, National University of Singapore. She has published several papers in the top venues, such as ACM SIGIR, MM, and IEEE Transactions on Image Processing. Her research interests include multimedia computing and scene understanding. She has served as a reviewer for many top conferences and journals.

Wenmiao Hu received her B.E. degree from Nanyang Technological University, Singapore, in 2014, her M.S. degree from Technical University Munich, Germany, in 2016 and her Ph.D. degree from National University of Singapore. She is currently a Senior Data Scientist at Grabtaxi Holdings Pte. Ltd. She has published several papers in the top venues, including ACM MM, SIGSPATIAL and ICCV. Her research interests span location-based services, geo-localization and computer vision. She has also served as a reviewer for many top conferences and journals.

Xuemeng Song received the B.E. degree from the University of Science and Technology of China, in 2012, and the Ph.D. degree from the School of Computing, National University of Singapore, in 2016. She is currently an Associate Professor with Shandong University, China. She has published several papers in the top venues, such as ACM SIGIR, MM, and ACM Transactions on Information Systems. Her research interests include information retrieval and social network analysis. She has served as a reviewer for many top conferences and journals. She is also an AE of IEEE Transactions on Circuits and Systems for Video Technology and IET Image Processing.

Hannes Kruppa is currently the head of Engineering & Data Science at Grab. He received the Ph.D. degree from the Department of Computer Science, ETH Zurich, Switzerland. He has published several papers in the top venues, such as ACM MM, SIGSPATIAL, and BMVC.

See-Kiong NG is currently a Professor of practice with the School of Computing, National University of Singapore, Singapore, where he is also the Director of translational research with the Institute of Data Science. He received the B.S. degree in computer science from Carnegie Mellon University (CMU), Pittsburgh, PA, USA, the M.S. degree from the University of Pennsylvania, Philadelphia, PA, USA, and the Ph.D. degree in computer science from CMU. He has authored more than 130mpapers on diverse and crossdisciplinary research topics, from bioinformatics to smart cities based on data science and AI.

Roger Zimmermann received the M.S. and Ph.D. degrees from the University of Southern California (USC). He is currently a Professor with the Department of Computer Science, National University of Singapore (NUS). He was recently the Deputy Director of the Smart Systems Institute (SSI) and the Co-Director of the Centre of Social Media Innovations for Communities(CoSMIC) at NUS. He has coauthored a book, more than 350 conference publications, journal articles, and book chapters. He holds several patents. His research interests include streaming media architectures, media networking, applications of machine/deep learning, and spatio-temporal data management. He is an Associate Editor of the ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), the IEEE Transactions on Multimedia, the Springer Multimedia Tools and Applications (MTAP), and the IEEE Open Journal of the Communications Society (OJ-COMS). He is a Distinguished Member of the ACM.