Multimodal Fusion of Satellite Images and Crowdsourced GPS Traces for Robust Road Attribute Detection

Yifang Yin
Grab-NUS AI Lab, NUS
idsyin@nus.edu.sg

An Tran
GrabTaxi Holdings
an.tran@grab.com

Ying Zhang
Grab-NUS AI Lab, NUS
dcszyi@nus.edu.sg

Wenmiao Hu
GrabTaxi Holdings
wenmiao.hu@grab.com

Guanfeng Wang
GrabTaxi Holdings
guanfeng.wang@grab.com

Jagannadan Varadarajan
GrabTaxi Holdings
vjagan@gmail.com

Roger Zimmermann
Grab-NUS AI Lab, NUS
dcsrz@nus.edu.sg

See-Kiong Ng
Grab-NUS AI Lab, NUS
seekiong@nus.edu.sg

ABSTRACT

Automatic inference of missing road attributes (*e.g.*, road type and speed limit) for enriching digital maps has attracted significant research attention in recent years. A number of machine learning based approaches have been proposed to detect road attributes from GPS traces, dash-cam videos, or satellite images. However, existing solutions mostly focus on a single modality without modeling the correlations among multiple data sources. To bridge the gap, we present a multimodal road attribute detection method, which improves the robustness by performing pixel-level fusion of crowdsourced GPS traces and satellite images. A GPS trace is usually given by a sequence of location, bearing, and speed. To align it with satellite imagery in the spatial domain, we render GPS traces into a sequence of multi-channel images that simultaneously capture the global distribution of the GPS points, the local distribution of vehicles' moving directions and speeds, and their temporal changes over time, at each pixel. Unlike previous GPS based road feature extraction methods, our proposed GPS rendering does not require map matching in the data preprocessing step. Moreover, our multimodal solution addresses single-modal challenges such as occlusions in satellite images and data sparsity in GPS traces by learning the pixel-wise correspondences among different data sources. Extensive experiments have been conducted on two real-world datasets in Singapore and Jakarta. Compared with previous work, our method is able to improve the detection accuracy on road attributes by a large margin.

CCS CONCEPTS

• **Information systems → Data mining**; • **Computing methodologies → Neural networks**.

KEYWORDS

Road attributes, satellite images, digital maps, GPS trajectories

ACM Reference Format:
Yifang Yin, An Tran, Ying Zhang, Wenmiao Hu, Guanfeng Wang, Jagannadan Varadarajan, Roger Zimmermann, and See-Kiong Ng. 2021. Multimodal Fusion of Satellite Images and Crowdsourced GPS Traces for Robust Road Attribute Detection. In *29th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '21), November 2–5, 2021, Beijing, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3474717.3483917

1 INTRODUCTION

To have an accurate and up-to-date digital map is crucial for today's ride-hailing providers such as Didi [1] and Grab [15] to maintain the high quality of their taxi and car services. Incomplete map data such as a missing road or even a missing road attribute can lead to misleading routing decisions or inaccurate estimation of a driver's arrival time. Unfortunately, enriching digital maps with road attributes is tedious and labor-intensive as it requires heavy manual input from human annotators. This results in the missing of road attribute labels in a significant number of roads in both commercial and crowdsourced digital maps. Using OpenStreetMap (OSM) [10] as an example, its data completeness and accuracy vary significantly among different cities around the world. While about 23% of the roads in downtown Singapore are annotated with the speed limit label, this number drops significantly to only a few in central Jakarta. Therefore, it is crucial to develop a robust road attribute detection method to reduce the manual cost for crowdsourced map updating. For a target road attribute, the method should be able to filter out a large number of true negatives and only return a small set of detected candidates to human annotators for further verification.

Existing road attribute detection methods mostly model this task as a multi-class classification problem. Based on the data sources they adopted, existing methods can be roughly divided into two categories, GPS based and image (*e.g.*, dash-cam videos or satellite images) based methods. GPS based methods extract road features from large-scale crowdsourced GPS traces of vehicles. As the raw GPS traces are simply a sequence of latitude and longitude pairs that does not contain any information about the true route the vehicle

SIGSPATIAL '21, November 2–5, 2021, Beijing, China

ACM ISBN 978-1-4503-8664-7/21/11.
https://doi.org/10.1145/3474717.3483917

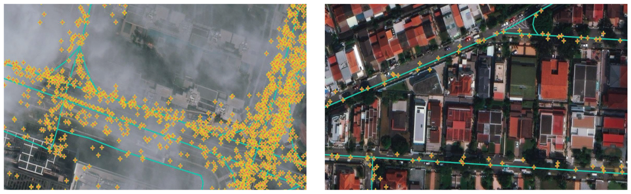


Figure 1: Illustration of the challenges for road attribute detection from satellite images or GPS traces alone. Left: a satellite image with heavy clouds. Right: sparse GPS samples over a narrow road.

traveled, map matching algorithms [19, 20] must be applied first to map GPS traces to road segments based on estimated probabilities. This results in limitations of existing GPS based methods as the effectiveness of map matching algorithms can be degraded by noisy GPS traces and incomplete map data. Inspired by the great success of Convolutional Neural Networks (CNN) on image classification, satellite imagery based road attribute detection methods have been proposed recently, which extract road descriptors from high resolution satellite images [6, 12]. Though promising results have been reported on the detection of a few road attributes, the generality of this method is limited as many road attributes such as one-way road and speed limit are barely visible from satellite images.

It is challenging to detect road attributes from a single data source as every data source has its own limitations. Figure 1 shows an example where little information about the road can be extracted from a cloudy satellite image or from a few sparse GPS traces. However, prior work on multimodal fusion of satellite images and GPS traces for road attribute detection is very limited, as the two data sources are in significantly different formats, which makes the fusion non-trivial. To address this issue, we propose to render GPS traces into a sequence of multi-channel images. The advantages of our proposed solution are twofold. First, the generated multi-channel images from GPS traces are spatially aligned with satellite images on the pixel level. Thus, the two data sources can be easily fused at the input layer by concatenation where pixel-wise correspondences can be learnt. Second, our proposed method does not rely on map matching so it is more efficient compared with traditional GPS based road attribute detection methods. Previous work on GPS rendering focused on the 2D modeling of latitude and longitude locations only [22]. Comparatively, our work focuses on the rendering of bearing and speed as they capture not only the spatial distribution of the GPS points, but also the local distribution (*i.e.*, at each pixel) of vehicles' moving directions and speeds, which are more important for the road attribute detection. We further extend the GPS rendering from 2D to 3D to capture the temporal changes of GPS traces on the road. The key contributions of this paper are summarized as follows:

- To the best of our knowledge, we present the first spatial-temporal multimodal fusion framework that learns the pixel-wise correspondences from the aligned satellite images and GPS traces for robust road attribute detection.
- We propose to extract informative road features from GPS traces by rendering location, bearing, and speed into a sequence of multi-channel images, which is essentially different to and more effective than the traditional GPS based road attribute detection methods.
- Our multimodal fusion solution can be integrated with existing network architectures. We demonstrate its effectiveness by integrating with AlexNet, MobileNet, and DenseNet where significant performance gain can be obtained when comparing with single-modal models.
- We have conducted extensive experiments on two real-world datasets in Singapore and Jakarta. Our method obtains the state-of-the-art classification accuracy of 91.4%, 76.3%, 86.2%, and 83.8% on the detection of one-way road, number of lanes, speed limit, and road type, respectively.

The rest of the paper is organized as follows. First, we report the important related work in Section 2. Next, we formulate the problem and introduce our proposed road feature extraction from multimodal data sources in Section 3. We present our multimodal fusion network for robust road attribute detection in Section 4 and evaluate the effectiveness of our proposed approach in Section 5. Finally, we conclude and suggest future work in Section 6.

2 RELATED WORK

Early research on road attribute detection has focused extensively on the feature extraction from GPS traces and probabilistic modeling for individual road attributes such as road type [3], road boundary [24], and lane detection [2]. For instance, Chen and Krumm presented a probabilistic model to derive the number of traffic lanes from GPS traces [8]. They proposed to use a Gaussian mixture model to model the distribution of GPS traces across multiple traffic lanes. Li *et al.* adopted the Support Vector Machine (SVM) as the classifier to detect the road class and road name from a combination of movement trajectories and geotagged social media data [18]. Van *et al.* proposed to extract different features from GPS traces, based on which a decision tree was built for each of the road attributes to be detected [23]. Due to the intrinsic noise in GPS data, map matching algorithms [19, 20, 25, 26] are mostly applied in the preprocessing to assign each GPS point to the road. Then per-road features are extracted from the corresponding GPS points that are matched to it. Finally, road attribute detection is conducted based on the per-road features extracted in the previous step. The effectiveness of such GPS-based road attribute detection methods purely rely on the quality of the GPS traces, where the detection accuracy can be degraded by inaccurate map matching results and GPS sparsity.

In addition to GPS traces, recent efforts have been made on investigating the use of other data sources such as satellite images [12], dash-cam videos [16], and crowdsourced map data [27] in the detection of missing road attributes. From the perspective of data coverage and quality, satellite imagery is considered to be one of the most promising data sources. High quality satellite images have long been utilized for automatic road network extraction, which includes deriving the road network geometry and topology [4, 28]. Model architectures such as U-net [21] or Deeplab [7] are usually adopted to segment an entire satellite image into semantic regions.

However, compared to the problem of road network extraction, road attribute detection is far more challenging, given that attributes like the speed limit are barely visible in satellite images. Recently, He *et al.* [12] presented an image-based road attribute detection approach using Graph Neural Networks (GCN). However, they evaluated their method on two road attributes (*i.e.*, road type and number of lanes) only. To extend the scope of the existing methods, joint analysis of multiple data sources can be an effective way to deal with more challenging road attributes. Unfortunately, the research on multimodal feature extraction and fusion from different data sources is still quite limited in the field of automatic road attribute detection.

3 ROAD FEATURE MODELING

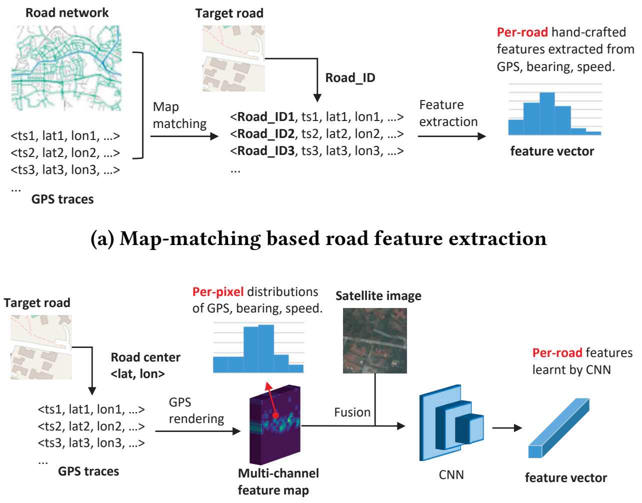
We propose to extract image-based road features from satellite images and crowdsourced GPS traces, and model the detection of each road attribute as an image classification problem. A GPS trace is mostly given as a sequence of sensor data (including latitude, longitude, bearing, speed, *etc.*) ranked by timestamps. Due to the difference in data format, it remains unclear what is the most effective way to fuse GPS data with satellite images for robust road attribute detection. Targeting at this problem, we introduce how we extract and fuse the image-based road features from both satellite images and GPS traces, followed by discussions on the key parameter settings in our proposed framework.

3.1 Satellite Images

To extract road features from satellite images, we first extract road networks from OpenStreetMap which provides a free and user-generated map of the world [10]. Next, we crop a satellite image at the center of each road segment as the road feature to be passed to the classifier. We set the image resolution corresponding to the zoom level 18 of the OpenStreetMap, where each pixel approximately represents 0.596 m on the Equator. Yin *et al.* [27] proposed to use the images of the local map data for road attribute detection. We are interested in utilizing this feature as an additional input in the future. Thus, we set the resolution of the satellite images to be consistent with the zoom levels of the map data to facilitate the feature fusion. He *et al.* [12] proposed to use a higher resolution at 12.5 cm/pixel to capture details on the road such as lane markings. However, the coverage and availability of such high resolution satellite images are very low. We set the size of each cropped image to 224 × 224, corresponding to a 134 × 134 meter tile on each road segment. Such an image contains not only the road, but also the surrounding environment around the road, which can be helpful for the missing road attribute detection as well [27].

3.2 Crowdsourced GPS Traces

A GPS trace is defined to be a sequence of records associated with timestamps. Each record consists of location, bearing, and speed returned by sensors. The location of a GPS record is usually represented by the latitude and longitude pair. The bearing is the clock-wise angle of the device’s moving direction with respect to the earth’s true north direction. A raw GPS trace is noisy and does not contain the information of the true route the vehicle travelled. Therefore, traditional GPS-based road attribute detection methods mostly perform map matching algorithms [20] to find the group of traces that are associated with each road segment in the preprocessing phase. Figure 2 illustrates the different strategies of (a) the conventional and (b) our proposed road feature extraction methods. As shown in Figure 2 (a), conventional road feature extraction methods first predict the road ID associated with each GPS point based on map matching. Then, given the ID of a target road, statistics such as distributions are extracted from the corresponding GPS records. Thereby roads are represented by hand-crafted feature vectors. Comparatively, we propose to directly render GPS traces into a multi-channel feature map, at the center coordinate of the target road. The multi-channel feature map is generated by extracting per-pixel distributions for the number of GPS traces, bearing, and speed, which can be easily fused with satellite images as they are spatially aligned at the pixel level. CNNs are next adopted to learn the final road representations, which are more robust and informative compared to the hand-crafted road features used in conventional methods. Moreover, as map matching is not applied, our method is more efficient when dealing with large amount of GPS data. It is also less sensitive to the quality of the map data. This is because map matching is conducted on the global road network whereby its performance can be degraded by missing roads. Our method, on the other hand, only utilizes the local information (*e.g.*, location and orientation) of the target road when generating features for it.



(a) Map-matching based road feature extraction

(b) Proposed GPS rendering based road feature extraction

Figure 2: Comparison of (b) our proposed GPS rendering to (a) conventional map-matching based road feature extraction methods.

Formally, let $P^i = \{p_1^i, p_2^i, ..., p_m^i\}$ denote the set of GPS points that fall into the nearby region of road segment r_i, and where $p_j^i = (lat_j^i, lon_j^i, bearing_j^i, speed_j^i)$ is a 4-tuple that contains the readings of latitude, longitude, bearing, and speed. We render the GPS points P^i into a 224 × 224 multi-channel image at the same resolution of the satellite images as illustrated in Algorithm 1. Let GL^i, GB^i, and GS^i represent the corresponding channels generated

Algorithm 1: 2D GPS Traces Rendering

```
Input: A set of GPS points $P^i$ in the nearby region of a road
       segment $r_i$
Output: A multi-channel image $G^i$ as the feature extracted
        from $P^i$ for road segment $r_i$
3D Array $GL^i$, $GB^i$, $GS^i$;
/* $GL^i$,$GB^i$, and $GS^i$ are the image channels
   generated based on location, bearing, and
   speed, respectively.                              */
for each point $p^i_j$ in $P^i$ do
    /* $p^i_j$ = ($lat^i_j$, $lon^i_j$, $bearing^i_j$, $speed^i_j$) is a
       4-tuple that contains the readings of
       latitude, longitude, bearing, and speed. */
    $x$, $y$=locate_pixel($lat^i_j$, $lon^i_j$, $r_i$);
    update_location_channel($GL^i$, $x$, $y$) //based on Eq. 1;
    update_bearing_channel($GB^i$, $x$, $y$, $bearing^i_j$) //based on
      Eq. 2;
    update_speed_channel($GS^i$, $x$, $y$, $speed^i_j$) //based on
      Eq. 3;
$G^i$=Concat(($GL^i$, $GB^i$, $GS^i$),axis=-1);
kernel_smoothing($G^i$) //based on moving average;
normalization($G^i$) //based on Eq. 4;
return $G^i$;
```

based on location, bearing, and speed, respectively. For each GPS point $p^i_j \in P^i$, we first transform it to the pixel location, given as (x, y), in the image, and then update GL^i, GB^i, and GS^i as below.

Update Location Channel. GL^i is defined to be a single-channel image. It counts the number of GPS points that are projected onto each pixel. Let $GL^i(x, y)$ denote the element at (x, y) in image GL^i. Then, GL^i is updated by

$$GL^i(x, y) = GL^i(x, y) + 1 \tag{1}$$

Update Bearing Channel. GB^i is defined to be a M_b-channel image where M_b is the number of bins we adopted to quantize the bearing values in degree into a histogram at each pixel. Let Bin_b denote the bin size to generate the bearing histogram, we update GB^i given a GPS point p^i_j with $bearing^i_j$ at (x, y) in the image as

$$\begin{aligned} &GB^i(x, y, int(bearing^i_j/Bin_b)) \\ &\quad = GB^i(x, y, int(bearing^i_j/Bin_b)) + 1 \end{aligned} \tag{2}$$

Update Speed Channel. GS^i is defined to be a M_s-channel image where M_s is the number of bins we adopted to quantize the speed values in m/s into a histogram at each pixel. Let Bin_s denote the bin size to generate the speed histogram, we update GS^i given a GPS point p^i_j with $speed^i_j$ at (x, y) in the image as

$$\begin{aligned} &GS^i(x, y, int(speed^i_j/Bin_s)) \\ &\quad = GS^i(x, y, int(speed^i_j/Bin_s)) + 1 \end{aligned} \tag{3}$$

Next, we concatenate GL^i, GB^i, and GS^i to form a $(1 + M_b + M_s)$-channel image as the image-based feature extracted from GPS traces for road segment r_i. Finally, to reduce the impact of the intrinsic noise and the uneven distribution of GPS traces on the rendering result, we apply kernel smoothing and normalization as described below to obtain the final image-based feature G^i extracted from GPS.

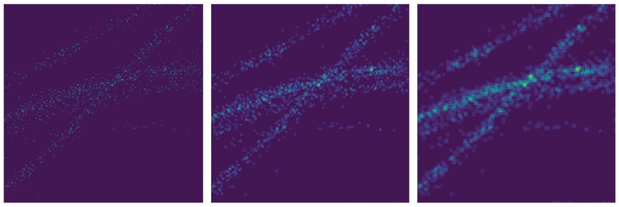


Figure 3: GPS rendering results after applying kernel smoothing. From left to right: rendering of the (a) original GPS traces, (b) smoothed GPS traces with kernel=3, and (c) smoothed GPS traces with kernel=5.

Kernel Smoothing. As shown in Figure 3, with a high rendering resolution at 0.6 m/pixel, the projection of the original GPS points around a road segment r_i can be noisy and sparse. To address this issue, we smooth each channel of G^i by computing the moving average over a square kernel with size K. Figure 3 shows the moving average rendering of the GPS points with $K = 3$ and $K = 5$. Alternative weighting functions such as 2D Gaussian kernel [11] can be adopted, but the parameters need to be tuned based on the characteristics of the GPS data.

Normalization. The distribution of GPS traces on roads can be unbalanced due to different road types or locations [5]. To reduce the impact of GPS disparity, we normalize GL^i, GB^i, and GS^i by

$$\begin{aligned} GL^i(x, y) &= GL^i(x, y)/\max_{\{x',y'\}} GL^i(x', y') \\ GB^i(x, y) &= GB^i(x, y)/\textstyle\sum_{c'=0}^{M_b-1} GB^i(x, y, c') \\ GS^i(x, y) &= GS^i(x, y)/\textstyle\sum_{c'=0}^{M_s-1} GS^i(x, y, c') \end{aligned} \tag{4}$$

to obtain the final GPS rendering result G^i. The location channel is normalized based on the max value over all the pixels, while the bearing and speed channels are normalized based on the sum over all the respective channels at each pixel. In other words, we normalize the location at the image level and normalize the bearing and speed at the pixel level to make them complementary to each other.

3.3 Calibration based on Road Direction

To reduce the impact of road directions on the extraction of road features, we calibrate the road features in the following two aspects. First, we rotate both the satellite images and the GPS based multi-channel images to ensure that the road direction is always horizontal in the image [12]. Second, instead of using the absolute bearing values in the GPS traces, we compute the angle distance between the moving direction of the vehicle and the direction of road segment r_i to calculate GB^i [27]. This is based on the observation that some road attributes such as the one-way/two-way road can be more correlated with the relative angle rather than the absolute bearing values.

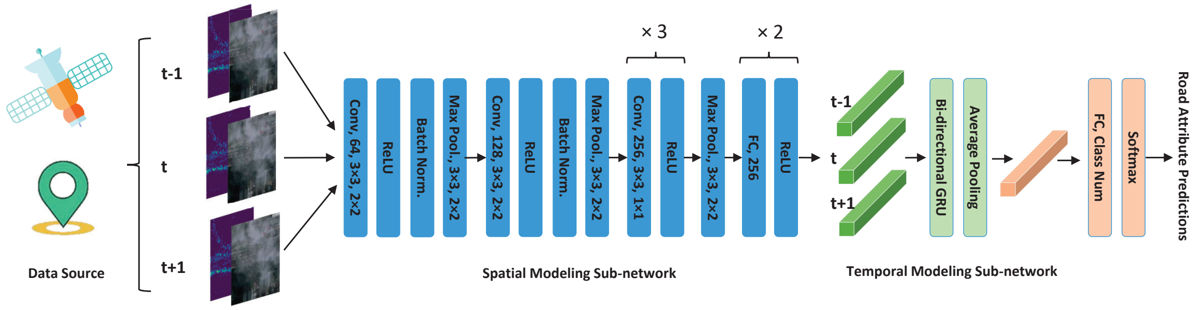


Figure 4: Network architecture of our proposed multimodal fusion network for robust road attribute detection.

Both calibration methods strengthen the features around the roads while weaken the features of the surrounding environments. This is especially helpful for the detection of road attributes such as the number of lanes as the road width can be more easily recognized after calibration. However, the detection of some other road attributes such as the speed limit and road type may also rely on the features of the surrounding environments, *e.g.*, residential roads are always within residential areas. It may weaken the feature consistency of the surrounding environments too much by applying both calibration methods, resulting in less satisfactory detection rates of certain road attributes.

Table 1: Categories of the four road attributes to be detected.

Road attributes	Category
One/Two Way	one-way road, two-way road
No. of Lanes	1, 2, 3, 4, 5, 6
Speed Limit (km/h)	40, 50, 60, 70, 80, 90
Road Type	residential, primary, secondary, tertiary, service, footway

3.4 From 2D to 3D GPS Rendering

We have introduced how to render GPS traces into a multi-channel image in Section 3.2. When sufficient GPS traces are available, it is beneficial to extend the GPS rendering from 2D to 3D by introducing a third time axis. For example, assume that we have one day's GPS trajectories of all the taxi drivers in a city. It is unnecessary to perform the 2D rendering using the whole day's GPS data as oversampling will only introduce significant computational cost. One way to make use of the data is to split a 24-hour day into time slots and extract multiple road features from each of the time slots. The interval of the time slot should be chosen based on the data. Basically, the GPS traces in each time slot should have a reasonable dense coverage over the area of interest.

Formally, let us divide the time axis into T time bins with equal duration. In each time bin t, we generate a multi-channel image G_t^i using only the GPS traces that are within time bin t. Subsequently, the 3D GPS rendering result can be written as $G_{3D}^i = \{G_0^i, G_1^i, ..., G_T^i\}$. The advantages of the 3D GPS rendering are twofold. First, it scatters GPS noise into different time bins and thus reduces its impact on the overall representation. Second, it effectively captures the temporal changes of GPS traces on the road. The patterns of GPS data are different during peak and non-peak hours, which can be an important feature for the detection of some road attributes such as the road type classification.

4 MULTIMODAL FUSION FOR ROAD ATTRIBUTE DETECTION

For each road with a missing road attribute (*e.g.*, road type and speed limit) that we would like to detect, we extract the image-based road features from both satellite images and crowdsourced GPS traces, and model the detection of the missing road attribute as an image classification problem. As shown in Figure 4, we adopt a network architecture comprised of different modules, including a spatial modeling sub-network with five convolutional layers and two fully-connected layers, a temporal modeling sub-network with one bi-directional GRU layer followed by average pooling, and one classification layer. We adopt a kernel size of 3 and set the number of filters to 64, 128, 256, 256, and 256, respectively. A stride of 1 is adopted in the last three convolutional layers, while a stride of 2 is adopted in the rest convolutional layers and the max pooling layers. In this paper, we target at four road attributes, namely one-way/two-way road, number of lanes, speed limit, and road type. The categories for each road attribute are shown in Table 1. Taking road type as an example, the classifier will output probability scores over the six categories and the one with the highest probability score will be selected as the final prediction. The cross-entropy loss is adopted for model optimization.

By rendering GPS traces into multi-channel images, multimodal fusion can be directly conducted at the input layer. For 3D GPS rendering, we concatenate the satellite RGB image to the multi-channel GPS image in each time bin and use the network shown in Figure 4 to perform road attribute detection. For 2D GPS rendering, we simply concatenate the RGB channels of the satellite image and the location, bearing, and speed channels generated from the GPS traces as the input. In this case, the input does not contain any temporal features, so we remove the temporal modeling sub-network to conduct road attribute detection. This fusion strategy has the advantage of being able to learn filters from the multimodal features directly as the satellite images and the GPS traces are spatially aligned at the same rendering resolution.

Table 2: Numbers of samples in the training/testing datasets for the four road attributes.

Dataset	One/Two Way	No. of Lanes	Speed Limit	Road Type
Singapore	15049/3763	10413/2667	7553/1923	13556/3388
Jakarta	5398/1350	3404/881	-	4171/1017

Table 3: Road attribute classification accuracy comparison based on satellite images, GPS traces, and their fusion.

Classifier	Singapore				Jakarta		
	One/Two Way	No. of Lanes	Speed Limit	Road Type	One/Two Way	No. of Lanes	Road Type
Satellite	0.7778	0.6430	0.7374	0.6942	0.8200	0.6288	0.6411
GPS - 2D	0.8198	0.6678	0.7722	0.6671	0.8089	0.5970	0.6332
GPS - 3D	0.8297	0.6865	0.8040	0.7075	0.8267	0.6311	0.6588
Fusion - 2D	0.8488	0.6967	0.8045	0.7624	0.8289	0.6470	0.7178
Fusion - 3D	**0.8546**	**0.7132**	**0.8242**	**0.7904**	**0.8363**	**0.6606**	**0.7355**

5 EVALUATION

We first introduce the experimental setup in Section 5.1, and then proceed with the evaluation of our proposed road attribute detection framework. We compare the effectiveness of the satellite images, GPS trajectories, and their fusion in the detection of four road attributes, namely one-way/two-way road, number of lanes, speed limit, and road type. Next, we perform an ablation analysis on the settings of bin number and kernel size for GPS feature generation to verify the design of our proposed model. Finally, we compare our proposed multimodal model to seven state-of-the-art methods to demonstrate its effectiveness in road attribute detection.

5.1 Experimental Setup

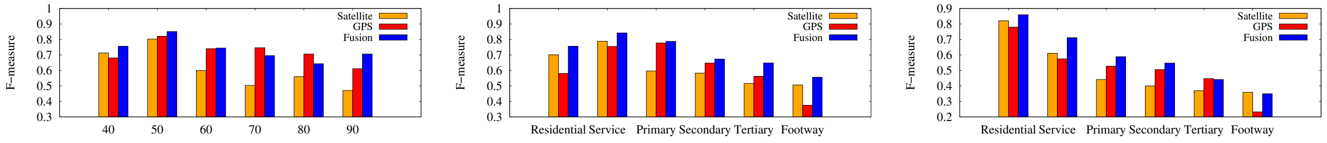
We evaluated our proposed methods based on two large-scale real-world datasets in Singapore and Jakarta, respectively. To prepare the datasets, we derive the ground-truth labels of four road attributes, namely one-way/two-way road, number of lanes, speed limit, and road type, from the OpenStreetMap data. We remove the road segments without ground-truth labels and randomly divide the remaining into 80%-20% splits for training and testing. The number of training and testing samples for each road attribute is illustrated in Table 2. We are unable to perform speed limit detection on the Jakarta dataset as only a few roads in Jakarta are annotated with the speed limit label. For feature extraction, we use satellite imagery from DigitalGlobe and three-hour (*i.e.*, 4:00 pm - 7:00 pm) real-world GPS traces of in-transit Grab drivers in Singapore and Jakarta [15]. We divide the GPS traces into three one-hour time bins for 3D GPS rendering. The sampling rate of the GPS traces are mostly 1 Hz in our experiments. For optimization, we use the Adam optimizer with a learning rate of 0.001 and a batch size of 32. For comparison, we report the overall classification accuracy and the per-class F-measure as the evaluation metrics.

5.2 Model Justification

We first compared the classification accuracy obtained based on satellite images only, GPS traces only, and their fusion. The results are reported in Table 3 with the **best result** highlighted in each column. In this experiment, the number of bins to generate the bearing and speed channels from GPS traces were set to $M_b = 3$ and $M_s = 3$, respectively. The kernel size for GPS smoothing was set to 9×9. The tuning of these parameters will be evaluated and discussed in the next section.

The classifiers trained on satellite images or GPS traces have their own limitations. For example, the visibility of roads may not always be good due to occlusions caused by trees, buildings, or even heavy clouds in a satellite image. The crowdsourced GPS traces, on the other hand, contain intrinsic noise resulting in incorrectly placed GPS points off the road. Generally speaking, the classifiers trained in Singapore performed better than those trained in Jakarta. From Table 2 we can see that the quality of the Singapore dataset is better as the number of roads with ground-truth attribute labels is much higher than that in Jakarta. Next, we compare the performance with input features generated from either 2D or 3D GPS rendering. On both datasets, *GPS-3D* and *Fusion-3D* outperformed *GPS-2D* and *Fusion-2D* in all the cases. This is because the input feature generated from 3D GPS rendering is more descriptive as it captures not only the spatial distributions but also the temporal changes of the GPS traces on the road. *Fusion-3D* significantly improved the classification accuracy compared to the individual classifiers trained on satellite images and GPS traces separately. On the road type detection, the classification accuracy has been improved by 11.6% ~ 14.7%, which demonstrates the effectiveness of our proposed approach.

The per-class F-measure comparison of the 2D classifiers trained based on satellite images, GPS traces, and their fusion is shown in Figure 5. F-measure is computed as $2 \cdot \frac{precision \cdot recall}{precision + recall}$, which considers both precision and recall. For speed limit detection, the GPS based classifier outperformed the satellite image based classifier in most of the cases except the class of 40 km/h. This is because GPS traces contain vehicle's bearing and speed information, which is missing from the satellite images. Such sensor data provides additional information for road attribute detection. However, one issue is that such sensor data can be noisy especially when the vehicle's speed is low. For road type detection, the satellite image based

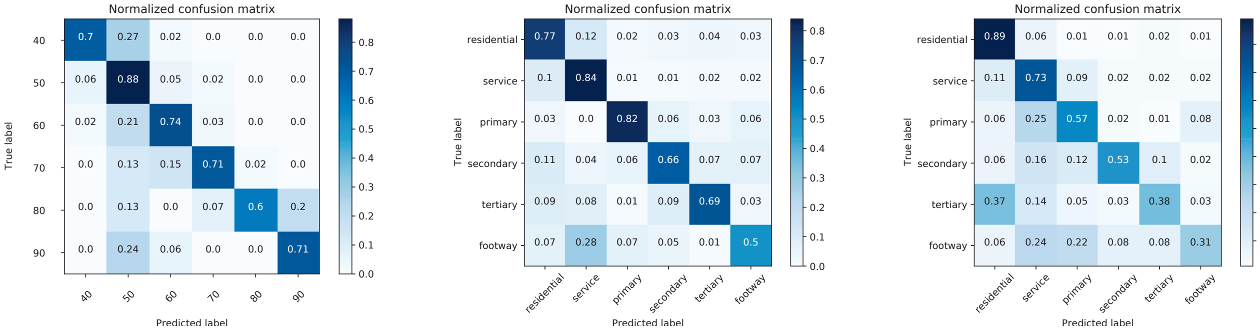


(a) Singapore, speed limit

(b) Singapore, road type

(c) Jakarta, road type

Figure 5: Per-class F1 measure of road attribute detection based on satellite images, GPS traces, and their fusion on the detection of speed limit and road type.



(a) Singapore, speed limit

(b) Singapore, road type

(c) Jakarta, road type

Figure 6: Confusion matrix of our proposed multimodal road attribute detection method on the detection of speed limit and road type.

classifier performed better on classes of “Residential”, “Service”, and “Footway”, while the GPS based classifier performed better on classes of “Primary”, “Secondary”, and “Tertiary”. The same trend shows on both the Singapore dataset and the Jakarta dataset. This is because the contextual information in a satellite image shows the surrounding environment around a road. Thus, it helps with the detection of certain road types such as residential roads that are around residential areas and service roads that are for access to parking, driveways, and alleys. On the other hand, classes of “Primary”, “Secondary”, and “Tertiary” define the most, second most, and third most important roads in a country's road network system. Both the number of vehicles travelled on these roads and the quality of the corresponding GPS traces tend to be high. So the GPS based classifier tends to recognize such road types more easily.

Finally, we show the confusion matrix obtained by our proposed 2D multimodal road attribute detection method in Figure 6. For speed limit detection, the number of test samples for the six classes are 415, 1075, 299, 102, 15, and 17, respectively. We can see that most confusions are between common and rare classes. Due to the class imbalance, the detection of samples from rare classes is more challenging as the classifier tends to favor the majority class. For road type detection, the classifiers trained in both Singapore and Jakarta can recognize “Residential” and “Service” roads easily. “Primary”, “Secondary”, and “Tertiary” roads can be recognized more easily in Singapore than in Jakarta. One possible reason is that the number of training samples for those categories is not sufficient in the Jakarta dataset due to the lack of the ground-truth road type labels in the OpenStreetMap data.

5.3 Ablation Study on GPS Rendering

To better understand our proposed method, we performed an ablation analysis on the key parameter settings in the GPS rendering. We first study the impact of the number of bins/channels M_b and M_s that we used to render the sensor data into images. The *GPS-2D* classifier is used in this ablation study and the results are reported in Table 4. The first row reports the classification accuracy obtained by training with GPS locations only, which serves as a baseline when other sensor data is not available in the GPS traces. Recall that we render the GPS locations into a single-channel image GL that captures the spatial distribution of the GPS points by counting the number of GPS points that are projected onto each pixel. Comparatively, the M_b-channel image GB and the M_s-channel image GS rendered from bearing and speed tend to be more descriptive as they capture not only the spatial distribution of the GPS points, but also the local distribution of vehicles' moving directions and speeds at each pixel. By comparing the results shown in 2-4 rows, we can see that improved classification accuracy can be obtained by increasing the number of channels M_b and M_s especially for the detection of speed limit and road type. The last row in the table shows the classification accuracy obtained by training based on the fusion of GL, GB, and GS, where the input $G = Concat((GL, GB, GS), axis = -1)$. As features extracted from GPS location, bearing, and speed capture complementary information, the classifier trained on 7-channel G outperformed the classifier trained on 7-channel GB or 7-channel GS in the six out of the seven cases.

Next, we study the impact of the kernel size we adopted for GPS smoothing. We set the kernel size to 1×1, 3×3, 5×5, 7×7, 9×9, and

Table 4: Road attribute detection based on GPS traces with different sensor data and varying bins.

Sensor Data	Bin No.	Singapore				Jakarta		
		One/Two Way	No. of Lanes	Speed Limit	Road Type	One/Two Way	No. of Lanes	Road Type
GPS	1	0.7736	0.6393	0.6828	0.6012	0.7933	0.5800	0.6185
Bearing	3	0.8129	0.6423	0.7431	0.6281	0.8081	0.5823	0.6214
Bearing	7	0.8134	0.6562	0.7546	0.6420	0.8059	0.5845	**0.6352**
Speed	3	0.7962	0.6475	0.7249	0.6434	0.7844	0.5664	0.6155
Speed	7	0.7914	0.6490	0.7421	0.6600	0.7956	0.5834	0.6185
GPS+B.+S.	1+3+3	**0.8198**	**0.6678**	**0.7722**	**0.6671**	**0.8089**	**0.5970**	0.6332

Table 5: Road attribute detection based on GPS traces smoothed using kernels of different sizes.

Kernel Size	Singapore				Jakarta		
	One/Two Way	No. of Lanes	Speed Limit	Road Type	One/Two Way	No. of Lanes	Road Type
1×1	0.7853	0.6123	0.6417	0.5962	0.7837	0.5619	0.5929
3×3	0.8023	0.6457	0.7067	0.6346	0.7941	0.5709	0.6028
5×5	0.8087	0.6494	0.7343	0.6479	0.7948	0.5811	0.6205
7×7	0.8177	0.6530	0.7540	0.6588	0.8059	0.5800	0.6224
9×9	0.8198	**0.6678**	**0.7722**	0.6671	0.8089	**0.5970**	**0.6332**
11×11	**0.8222**	0.6599	0.7618	**0.6741**	**0.8096**	0.5959	0.6293

11×11 and report the results in Table 5. The advantages of applying kernel smoothing to GPS traces are twofold. First, it helps reduce the negative impact caused by sensor noise. Second, crowdsourced GPS traces can be sparse in some regions where pixels on a road can have no GPS points projected to it, especially when we render at a high resolution. As Table 5 shows, classifiers trained with smoothed GPS traces outperformed the classifiers trained with the original GPS traces by a large margin. We consider a kernel size of 9×9 to be a good choice based on the experimental results on our datasets, and use this as the optimal setting in the rest of the experiments.

5.4 Comparison with the State-of-the-arts

We compare our proposed method to two types of state-of-the-art approaches. First, our proposed multimodal fusion solution can be easily integrated with any existing network architectures to train an end-to-end classifier for road attribute detection. For verification, we compared our proposed method to a number of state-of-the-art image classification architectures including **AlexNet [17]**, **MobileNet [13]**, and **DenseNet [9]**. Table 6 reports the classification accuracy based on different combinations of model architectures and calibration methods on the Singapore dataset. The size of the AlexNet is similar to our baseline network, both of which consist of five convolutional layers followed by ReLU activations. AlexNet adopted a kernel size of 11×11 and 5×5 respectively in the first two convolutional layers, which is the biggest among the four models. Comparatively, MobileNet and DenseNet went much deeper. To improve a model's efficiency, MobileNet factorized a standard 2D convolution into a depthwise convolution and a 1×1 pointwise convolution. DenseNet divided the network into multiple densely connected dense blocks composed of narrow layers, which are connected by transition layers that perform 1×1 convolution and 2×2 average pooling. DenseNet have been utilized for satellite image processing such as land use classification in previous work where promising results have been obtained [9]. From the results we can see that our proposed 2D method performed competitively well with MobileNet and DenseNet, while our proposed 3D method obtained the best result on the detection of all the four road attributes.

Next, we analyze the effectiveness of the road direction based calibration methods described in Section 3.3. Table 6 shows the classification accuracy together with the performance gain *w.r.t* the classifiers trained without applying any calibrations (*i.e.*, the results reported in the first row). On one hand, by applying both image rotation and bearing adjustment calibrations, our model significantly improved the detection accuracy of the one-way/two-way road and the number of lanes by 6.9% and 7.0%, respectively. On the other hand, for speed limit and road type, the best detection rates were obtained by applying the bearing adjustment calibration only. One possible reason is that the detection of these road attributes partially relies on the features of the surrounding environments. For example, highways in the same downtown area may have similar speed limit and residential roads are always within the residential area. However, both image rotation and bearing adjustment tend to weaken the feature consistency of the surrounding environments as the former changes the pixel location and the latter changes the pixel feature. The best classification accuracy obtained by our model is 91.4%, 76.3%, 86.2%, and 83.8% for one-way/two-way road, number of lanes, speed limit, and road type detection, respectively.

Finally, we compare our proposed approach to four state-of-the-art road attribute detection methods. **Van *et al.* [23]** utilized decision trees as the classifiers to detect road attributes from handcrafted GPS features extracted with map matching. **He *et al.* [12]** utilized CNNs or GCNs as the classifiers to detect road attributes

Table 6: Classification accuracy comparison to the state-of-the-art CNN architectures and evaluation on the calibrations based on road directions.

Calibration	Method	One/Two Way	Gain	No. of Lanes	Gain	Speed Limit	Gain	Road Type	Gain
None	AlexNet [17]	0.8299	-	0.6610	-	0.7473	-	0.7308	-
	MobileNet [13]	0.8307	-	0.6742	-	0.8008	-	0.7651	-
	DenseNet [14]	0.8366	-	0.6775	-	0.7982	-	0.7659	-
	Ours - 2D	0.8488	-	0.6967	-	0.8045	-	0.7624	-
	Ours - 3D	0.8546	-	0.7132	-	0.8242	-	0.7904	-
- image rotation	AlexNet [17]	0.8456	1.9%	0.7019	6.2%	0.7509	0.5%	0.7373	0.9%
	MobileNet [13]	0.8666	4.3%	0.7229	7.2%	0.7826	-	0.7972	4.2%
	DenseNet [14]	0.8538	2.1%	0.7289	7.6%	0.7956	-	0.7952	3.8%
	Ours - 2D	0.8669	2.1%	0.7225	3.7%	0.7972	-	0.7881	3.4%
	Ours - 3D	0.8724	2.1%	0.7439	4.3%	0.8107	-	0.8132	2.9%
- bearing adjustment	AlexNet [17]	0.8411	1.3%	0.6337	-	0.6984	-	0.7470	2.2%
	MobileNet [13]	0.8469	2.0%	0.6625	-	0.7754	-	0.8076	5.6%
	DenseNet [14]	0.8576	2.5%	0.6757	-	0.8060	1.0%	0.8200	7.1%
	Ours - 2D	0.8738	2.9%	0.7019	0.7%	0.8300	3.2%	0.8070	5.8%
	Ours - 3D	0.8742	2.3%	0.7263	1.8%	**0.8617**	4.5%	**0.8383**	6.1%
- image rotation & bearing adjustment	AlexNet [17]	0.8836	6.5%	0.7278	10.1%	0.7353	-	0.7211	-
	MobileNet [13]	0.8937	7.6%	0.7402	9.8%	0.7852	-	0.8117	6.1%
	DenseNet [14]	0.9091	8.7%	0.7413	9.4%	0.7894	-	0.7869	2.7%
	Ours - 2D	0.9059	6.7%	0.7447	6.9%	0.7795	-	0.7834	2.8%
	Ours - 3D	**0.9136**	6.9%	**0.7630**	7.0%	0.8159	-	0.8070	2.1%

Table 7: Comparison to the state-of-the-art road attribute detection methods in terms of the classification accuracy.

Method	GPS	Satellite image	Map matching	One/Two Way	No. of Lanes	Speed Limit	Road Type
Van *et al.* [23]	✓	✗	✓	0.8359	-	-	-
He *et al.* [12]	✗	✓	✗	0.7826	0.6419	0.7592	0.6992
Yin *et al.* [27]	✓	✓	✓	0.8828	0.6869	0.7629	0.7287
Sun *et al.* [22]	✓	✓	✗	0.8121	0.6633	0.7587	0.7409
Ours - 2D	✓	✓	✗	0.9059	0.7447	0.8300	0.8070
Ours - 3D	✓	✓	✗	**0.9136**	**0.7630**	**0.8617**	**0.8383**

from satellite images. **Yin *et al.* [27]** extracted hand-crafted GPS features with map matching, processed GPS and images with two separate sub-networks, and fused the features before the classification layer. **Sun *et al.* [22]** performed GPS rendering without map matching, and fused GPS and images at the input layer. CNNs are adopted as the classifiers.

More specifically, Van *et al.* [23] proposed a GPS-based method that compares the heading of GPS points and the heading of the road. It clustered the points into three categories: "similar", "opposite", and "outliers" based on a threshold of 20 degrees. Points in the "outliers" cluster were removed, and a road was considered to be one-way if the percentage of the number of points in the "similar" cluster is larger than 0.9. The drawback of this method is that it failed to provide a solution for the number of lanes and road type detection. The speed limit decision tree was also performed poorly on our dataset possibly due to the difference between countries and geographic regions. He *et al.* [12] proposed an image-based road attribute detection method. We compared to their CNN baseline as we filtered out a significant number of unlabeled OSM roads, which creates difficulties in modeling the remaining roads in our datasets as a graph. In their original paper, they used satellite images with a very high resolution at 12.5 cm/pixel to capture the details on the road such as lane markings. However, this method performs less satisfactorily when the image resolution decreases. Yin *et al.* [27] extracted 1D hand-crafted features from map-matched GPS traces, which were next fused with visual features based on the two-stream fusion strategy. In their original paper, the visual features were extracted from the images of the local map data. Here we replaced the map visualization with the satellite image as the visual input to make it a fair comparison. One drawback of this method is that it relies on the results of map matching, which can be sensitive to the quality of both GPS traces and map data. Moreover, the 1D GPS features and the 2D satellite images can only be combined based on late fusion, where pixel-wise correspondences between the two data sources cannot be learnt. Our method, on the other hand, directly renders GPS traces into 2D images. Thereby

GPS features and satellite images can be effectively fused at the input layer, based on which more dense and robust features can be automatically learnt. A similar idea has been presented by Sun *et al.* [22] that rendered GPS traces into a 3-channel image to fuse with satellite imagery. However, their rendering strategy is less effective than our proposed method as we simultaneously modeled the global distribution of the GPS points, the local distribution of vehicles' moving directions and speeds at each pixel, and their temporal changes. By applying the calibration method, we are able to obtain the best road attribute detection accuracy of 0.9136, 0.7630, 0.8617, and 0.8383 on the four road attributes, respectively.

6 CONCLUSION AND FUTURE WORK

We present a multimodal fusion framework that learns from both satellite images and crowdsourced GPS traces for robust road attribute detection. In order to learn multimodal pixel-wise correspondences, we propose to render GPS traces into a sequence of multi-channel images that align with satellite imagery in the spatial domain. Moreover, our proposed GPS rendering method does not require map matching to preprocess the raw GPS traces. Thus, our method is more efficient and less sensitive to the quality of the map data compared to traditional map matching based road feature extraction methods. For evaluation, we collected two real-word datasets in Singapore and Jakarta, respectively. We performed an ablation study on the key parameters of GPS rendering, and evaluated the effectiveness of our proposed multimodal fusion approach on four road attributes, namely one-way or two-way road, number of lanes, speed limit, and road type.

In the future, we plan to extract road features from more data sources such as local map data, street views, accelerometers, gyroscopes, *etc.*, to further improve the road attribute detection accuracy. We would also like to detect not only static but also dynamic road attributes, *e.g.*, the congestion level of a road in a day.

ACKNOWLEDGMENTS

This work was funded by the Grab-NUS AI Lab, a joint collaboration between GrabTaxi Holdings Pte. Ltd. and National University of Singapore, and the Industrial Postgraduate Program (Grant: S18-1198-IPP-II) funded by the Economic Development Board of Singapore. This work was also supported by Singapore Ministry of Education Academic Research Fund Tier 2 under MOE's official grant number MOE2018-T2-1-103.

REFERENCES

[1] 2019. Data source: DiDi Chuxing GAIA Open Dataset Initiative. https://outreach.didichuxing.com/research/opendata/en/

[2] Heba Aly, Anas Basalamah, and Moustafa Youssef. 2016. Robust and Ubiquitous Smartphone-based Lane Detection. *Pervasive and Mobile Computing* 26 (2016), 35–56.

[3] Heba Aly and Moustafa Youssef. 2015. semMatch: Road Semantics-based Accurate Map Matching for Challenging Positioning Data. In *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems.* 5:1–5:10.

[4] Favyen Bastani, Songtao He, Sofiane Abbar, Mohammad Alizadeh, Hari Balakrishnan, Sanjay Chawla, Sam Madden, and David DeWitt. 2018. Roadtracer: Automatic Extraction of Road Networks from Aerial Images. In *IEEE Conference on Computer Vision and Pattern Recognition.* 4720–4728.

[5] James Biagioni and Jakob Eriksson. 2012. Map Inference in the Face of Noise and Disparity. In *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems.* 79–88.

[6] Gabriel Cadamuro, Aggrey Muhebwa, and Jay Taneja. 2018. Assigning a Grade: Accurate Measurement of Road Quality using Satellite Imagery. *arXiv preprint arXiv:1812.01699* (2018).

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2017. Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE transactions on Pattern Analysis and Machine Intelligence* 40, 4 (2017), 834–848.

[8] Yihua Chen and John Krumm. 2010. Probabilistic Modeling of Traffic Lanes from GPS Traces. In *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems.* 81–88.

[9] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. 2018. Functional Map of the World. In *IEEE CVPR.* 6172–6180.

[10] M. Haklay and P. Weber. 2008. OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Computing* 7, 4 (2008), 12–18.

[11] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. Kernel Smoothing Methods. In *The elements of statistical learning.* 191–218.

[12] Songtao He, Favyen Bastani, Satvat Jagwani, Edward Park, Sofiane Abbar, Mohammad Alizadeh, Hari Balakrishnan, Sanjay Chawla, Samuel Madden, and Mohammad Amin Sadeghi. 2020. RoadTagger: Robust Road Attribute Inference with Graph Neural Networks. In *AAAI Conference on Artificial Intelligence*, Vol. 34. 10965–10972.

[13] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv preprint arXiv:1704.04861* (2017).

[14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely Connected Convolutional Networks. In *IEEE CVPR.* 4700–4708.

[15] Xiaocheng Huang, Yifang Yin, Simon Lim, Guanfeng Wang, Bo Hu, Jagannadan Varadarajan, Shaolin Zheng, Ajay Bulusu, and Roger Zimmermann. 2019. Grab-Posisi: An Extensive Real-Life GPS Trajectory Dataset in Southeast Asia. In *ACM SIGSPATIAL International Workshop on Prediction of Human Mobility.* 1–10.

[16] Z. Jan, B. Verma, J. Affum, S. Atabak, and L. Moir. 2018. A Convolutional Neural Network Based Deep Learning Technique for Identifying Road Attributes. In *International Conference on Image and Vision Computing New Zealand.* 1–6.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems.* 1097–1105.

[18] Jun Li, Qiming Qin, Jiawei Han, Lu-An Tang, and Kin Hou Lei. 2015. Mining Trajectory Data and Geotagged Data in Social Media for Road Map Inference. *Transactions in GIS* 19, 1 (2015), 1–18.

[19] Yin Lou, Chengyang Zhang, Xing Xie, Yu Zheng, Wei Wang, and Yan Huang. 2009. Map-Matching for Low-Sampling-Rate GPS Trajectories. In *ACM SIGSPATIAL Internetional Conference on Advances in Geographical Information Systems.*

[20] Paul Newson and John Krumm. 2009. Hidden Markov Map Matching Through Noise and Sparseness. In *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems.* 336–343.

[21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention.* 234–241.

[22] Tao Sun, Zonglin Di, Pengyu Che, Chun Liu, and Yin Wang. 2019. Leveraging Crowdsourced GPS Data for Road Extraction from Aerial Imagery. In *IEEE CVPR.* 7509–7518.

[23] Karl Van Winden, Filip Biljecki, and Stefan Van der Spek. 2016. Automatic Update of Road Attributes by Mining GPS Tracks. *Transactions in GIS* 20, 5 (2016), 664–683.

[24] Wei Yang, Tinghua Ai, and Wei Lu. 2018. A Method for Extracting Road Boundary Information from Crowdsourcing Vehicle GPS Trajectories. *Sensors* 18, 4 (2018), 1261.

[25] Yifang Yin, Rajiv Ratn Shah, Guanfeng Wang, and Roger Zimmermann. 2018. Feature-Based Map Matching for Low-Sampling-Rate GPS Trajectories. *ACM Transactions on Spatial Algorithms and Systems* 4, 2 (2018).

[26] Yifang Yin, Rajiv Ratn Shah, and Roger Zimmermann. 2016. A General Feature-based Map Matching Framework with Trajectory Simplification. In *ACM SIGSPATIAL International Workshop on GeoStreaming.* 7:1–7:10.

[27] Yifang Yin, Jagannadan Varadarajan, Guanfeng Wang, Xueou Wang, Dhruva Sahrawat, Roger Zimmermann, and See-Kiong Ng. 2020. A Multi-task Learning Framework for Road Attribute Updating via Joint Analysis of Map Data and GPS Traces. In *World Wide Web Conference.* 2662–2668.

[28] Lichen Zhou, Chuang Zhang, and Ming Wu. 2018. D-LinkNet: LinkNet With Pretrained Encoder and Dilated Convolution for High Resolution Satellite Imagery Road Extraction. In *Computer Vision and Pattern Recognition Workshops.* 182–186.