A Highly Efficient and Effective Attribute Learning Framework for Road Graph from Aerial Imagery and GPS

Yang He
Amazon Last Mile
Bellevue, USA
yanhea@amazon.com

Emre Eftelioglu
Amazon Last Mile
Bellevue, USA
efteli@amazon.com

Mohamed Moustafa
Amazon Last Mile
Bellevue, USA
mmoustm@amazon.com

Amber Roy Chowdhury
Amazon Last Mile
Bellevue, USA
amberch@amazon.com

ABSTRACT

Road attributes play a pivotal role in digital maps, providing critical information for various routing and planning applications that aim to create a safe and efficient traffic environment. While some road attributes are available in existing map data such as OpenStreetMap [3], these sources may not cover all regions, meet high-quality standards, or include specific attributes required for specialized applications using these. To address these challenges, we propose a novel framework that leverages multi-task deep learning to learn road attributes from remote sensing imagery and GPS data. Our approach treats the task as a multi-task learning problem and incorporates convolutional and graph neural networks into an end-to-end learning framework. This enables efficient prediction of multiple road attributes for a set of input roads. To evaluate our system, we collect annotations and develop our model using public map sources. Our results demonstrate promising performance in predicting road type, road median, lane number, road directionality, and width in meters. By exploring different road attributes compared to previous works, our efforts open up new possibilities for novel applications in this domain. Overall, our research contributes to advancing the understanding and prediction of road attributes, enhancing the quality and completeness of digital maps, and enabling the development of innovative solutions for various applications.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; **Supervised learning**; **Neural networks**; • **Applied computing** → **Transportation**; • **Information systems** → **Geographic information systems**; **Global positioning systems**.

KEYWORDS

road attributes, satellite imagery, multi-task deep learning, graph neural networks

BigSpatial '23, November 13, 2023, Hamburg, Germany

ACM ISBN 979-8-4007-0345-4/23/11.
https://doi.org/10.1145/3615833.3628594

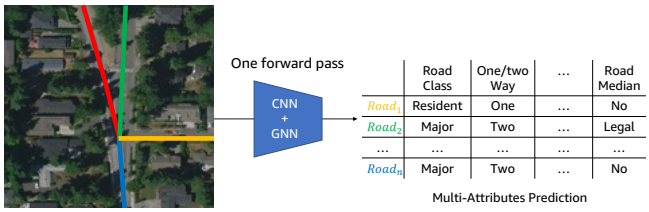

Figure 1: The proposed method outputs multiple road attributes from aerial imagery [1]for every input road geometry with one forward operation. Our model consists of CNN and GNN components for efficient and high-performance road attribute prediction.

ACM Reference Format:
Yang He, Emre Eftelioglu, Mohamed Moustafa, and Amber Roy Chowdhury. 2023. A Highly Efficient and Effective Attribute Learning Framework for Road Graph from Aerial Imagery and GPS. In *11th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data (BigSpatial '23), November 13, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3615833.3628594

1 INTRODUCTION

Digital maps are indispensable tools in today's navigation systems, serving a multitude of purposes, such as optimizing logistics costs and enabling efficient and safe route selection. As a result, the academic community and industry partners have placed significant emphasis on addressing map learning problems to ensure the accuracy and timeliness required for maintaining high-quality maps. Apart from road extraction techniques [13, 7, 33, 27, 9], which focus on inferring the missing roads, the acquisition of road attributes (e.g., one-/two-way roads, speed limit, number of lanes) also plays a vital role in constructing rich and valuable map data, for better next-step applications.

Modern approaches have made great efforts in leveraging deep neural network architectures [11] for visual recognition to predict road attributes from satellite imagery, by converting the road attribute learning problem into image classification where the image center is aligned onto a target road [12, 31]. Beyond the satellite

[1]All figures use randomly selected images for illustration and are not indicative of Amazon customer or delivery locations.

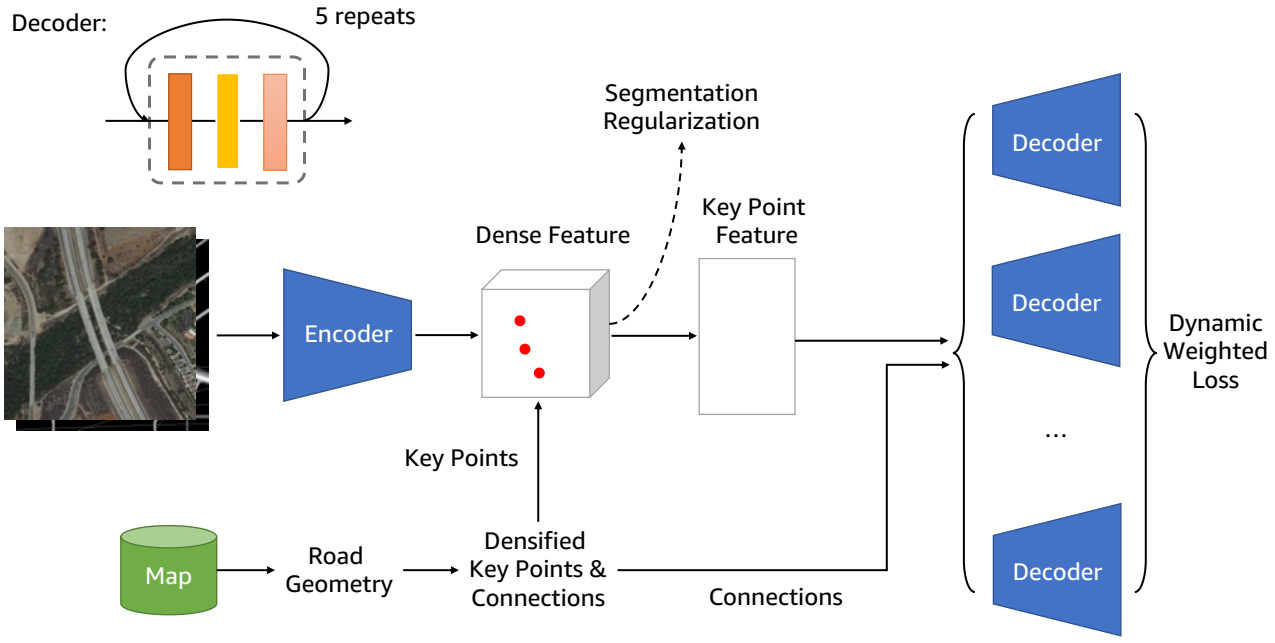

Figure 2: Diagram of the proposed model architecture for multi-task road attribute learning. It consists of an encoder for dense feature extraction and multiple GNN-based decoders for multi-task road attribute prediction for individual input road geometries. The decoder is composed of GraphSAGE operation, LayerNorm, and ReLU.

imagery, providing a straight observation, GPS data is a complementary and important clue to reflect driving behavior. Therefore, some related work combined GPS data for enhancing attribute prediction [31, 32]. However, treating road attribute learning as image classification is naive, owing to the lack of consideration in road geometry and overall layout. To overcome this, graph neural networks (GNNs) [6] are adopted to propagate intermediate results through road connections and then refine predicted attributes.

Despite the success of previous research, we point out several aspects in leveraging powerful deep neural networks for attribute learning. First, in general image classification tasks, the target objects typically constitute the majority of images, that convolution operations [21] or linear projections [8] can effectively extract visual features and classify the images. In contrast, roads have a distinct characteristic of being long and narrow, resulting in them occupying fewer pixels compared to other contextual elements, which potentially limits the performance of attribute learning. Second, performing image classification is an in-efficient way for different location of a road, as the images aligned from neighboring points have large overlapped regions. The redundancy makes practical deployment difficult and expensive in the worldwide level, which motivates us to seek for a more efficient framework and network architecture. Third, separate models are trained in previous work for individual attributes, which further restricts efficiency of model deployment. However, there are some correlation between different attributes (e.g., road class and number of lanes, road class and directionality). We believe this makes it possible to share features across multiple attributes and even improve individual tasks, since richer annotations are provided to a network.

In this paper, we propose a novel network architecture for multi-task road attribute learning from aerial (or satellite) imagery and GPS statistics, based on the combination between CNN and GNN. Instead of cropping images into a fixed size for image classification network, our model is able to flexibly take images at arbitrary sizes during training and inference as inputs.

We first leverage CNN to compute a dense feature map for a very big image, which can correspond to a $600 \times 600 m^2$ area and contain hundreds of roads. Then, we accurately extract features for the key points sampled from road geometry with bi-linear interpolation, that we do not need to align roads into image centers. Besides, we believe our model's success stems from its location-level feature extraction capability, helping the network focus on roads and nearby areas. Finally, GNN layers are utilized to propagate information over the key points and jointly trained with CNN layers end-to-end. Last but not least, to make full use of networks' capability of encoding visual representations, we provide building outline masks to our model as a regularization term, to encourage the network to learn more contextual dependencies and correlations between road attributes and context.

Below is a list of our contributions:

- We propose an efficient and effective road attribute learning network architecture. Our model extracts local features for individual key points from target roads and propagate information over the entire road graph, and then predicts multiple road attributes for all the roads in a region.
- We provide decent performance in several tasks simultaneously, including road median, which most previous work did not demonstrate, but it holds great importance (for U-turns, etc.).

- We demonstrate that building outlines serve as a valuable cue for road attribute learning. We apply them to encode additional semantic contexts, which offers new insights for future research in road attribute learning. This approach enables a better capture of context and facilitates the learning of geo-spatial representations for road attributes.

2 RELATED WORK

2.1 Road attribute learning

In the last decades, computer vision techniques became a hot topic that were applied to a variety of problems. Among these is to learn/infer roads and their attributes. [12] builds a CNN image classifier on satellite images to predict road attributes, where road segments are aligned at the center of images. Satellite images provide a direct observation for many road attributes, such as number of lanes and road type. Similarly, [31, 15, 26, 18] leverage satellite images to reach a similar goal, where they are used to extract visual representation to infer local road attributes [15] or conditions [26]. In addition to satellite images, GPS traces are also a crucial resource, especially to learn driving related attributes (e.g., road directions, speed limits, number of lanes, etc.). Early research on road attribute learning focused on hand-crafted features from GPS data, which can only used for certain attributes, such as road type [1], road boundary [30], etc. Besides, in order to extract GPS features for each road segment, map matching is required for many methods [5] which introduces further computational cost and increases the design complexity.

Recently, [31] renders location, bearing, and speed into a GPS image sequence, and combines these with satellite images using a deep neural network to predict road attributes, and achieves decent performance. Finally, graph representation of road networks is important to understand the road attributes and their interaction with each other. Therefore, several works use graph representations for the road structures to predict road type [14] and speed limit [16].

2.2 Graph neural networks for geospatial problems

Graph neural networks (GNNs) emerge as a powerful tool to model graph structures using deep neural network, similar to popular convolutional neural networks for array structures. It shows many useful applications in intelligent transportation systems [24] and geospatial problems [17, 28], because geo-objects and their connections are naturally represented as a graph structure, such as roads. It has shown several important mobility related applications, including traffic forecasting [4], spatial data analysis [20], disaster management [22]. In particular, GNN has been adopted in several maps learning problems, such as road inference [29, 13, 2] and road attribute learning [12]. The above-mentioned methods combine CNN and GNN, where CNN is used to learn local visual patterns and GNN helps to propagate the information through graph structures. The GNN component allows a model to focus on road topologies, rather than irrelevant image areas. However, those approaches usually crop a whole image into several overlapped small image patches, and feed each small patch into CNN. Therefore, they have to process redundant information over and over, limiting their efficiency at training and inference.

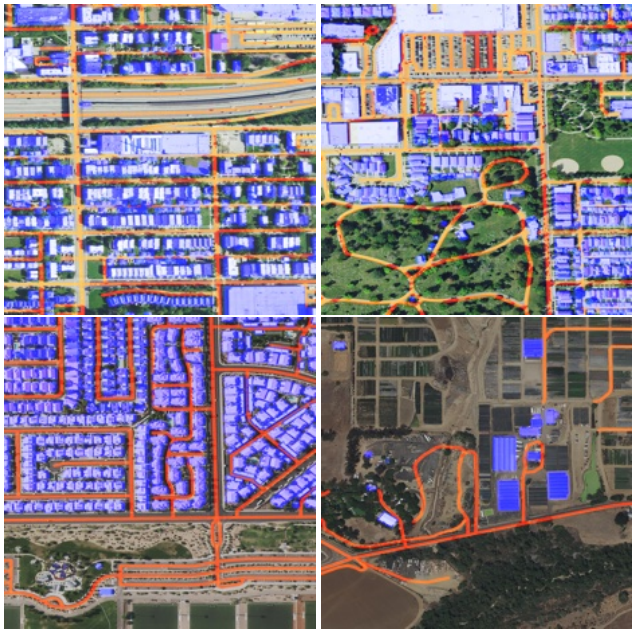

Figure 3: Examples of aerial images and semantic information used in this work. Those images are at 2048×2048 resolution and each pixel corresponds to $30 \times 30cm^2$. The building outline and road masks are drawn with blue and red colors. We encourage our network to segment building outlines and road masks to learn semantic intermediate representations.

Different from previous work [12, 31], we do not have to align a road as the image center to adopt image classification networks on fixed image sizes. We feed a larger image and multiple road locations into the network to learn and predict road attributes in a more efficient way, where we flexibly sample features from a dense feature map at arbitrary locations. Besides, the proposed framework does not require complicated pre-processing (e.g., image cropping), and sample local features at multiple locations in one forward operation. Therefore, our model is not only more efficient, but also provides wider context to propagate information. Eventually, we combine graph neural networks on a road graph to propagate image and GPS information, and train all the components end-to-end.

3 PROPOSED METHOD

We present our framework for road attribute learning from aerial imagery and GPS data in this section. Our model is based on dense prediction on individual key points provided by the road geometry from existing map data, for example OpenStreetMap (OSM) [3] used in this work. It consists of a CNN-based encoder to learn visual features for individual key points, as well as a GNN-based decoder to propagate the information through the graph of road geometries. In the combination between CNN and GNN, our model can be trained in an end-to-end manner, and predict road attributes for each input road segment efficiently. In detail, our model has multiple decoder branches, to produce various road attributes for all the input road geometry in one forward pass, which helps to deploy large-scale road attribute learning in practice. Figure 2 plots the overall architecture of the proposed model. It has two input

Figure 4: Road geometry densification example. We insert key points equidistantly (bottom) over the original road geometry (top). We feed original key points (blue circles), inserted key points (red dash circles), their connections, and corresponding aerial images into the proposed network.

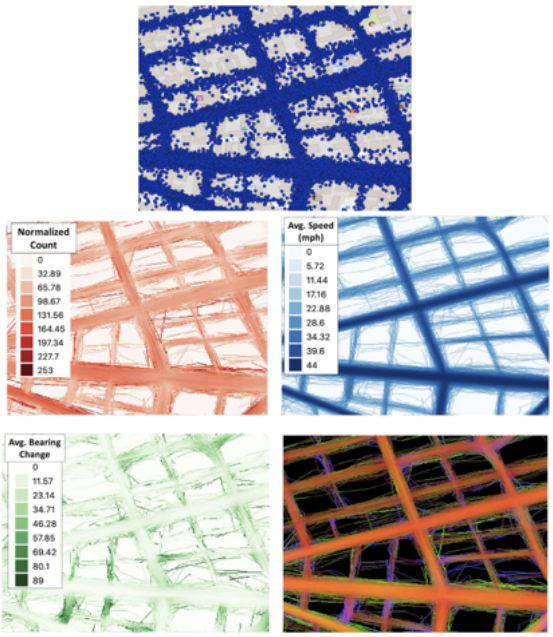

Figure 5: An example of GPS raster image created from raw GPS data points shown in the top. We visualize three channels of normalized count, average speed, average bearing change, and all 12-bands raster images.

branches. The first branch feeds image/GPS pairs into the network and learn visual features. The second branch takes road graphs as the inputs, indicating key point locations and their connections.

3.1 Aerial imagery and road geometry

Aerial (or satellite) imagery is used as the core data source in our work to learn road attributes. The images provide a direct observation of roads and nearby environments from bird's eye views. They typically capture many useful visual signals for learning road attributes, for example lanes, road markers, and some obstacles. Besides, road geometry indicates the location of roads, which not only provides the target for prediction, but also illustrates the overall road structure and layout for a region. Therefore, we leverage a data pair consisting of a road graph composed of multiple roads and the corresponding image captured from a vast area to acquire knowledge about road attributes. The utilization of a large area enables us to gather comprehensive information about a group of roads, thereby facilitating efficient attribute prediction and enhancing overall performance. For instance, as depicted in Figure 4, the top image exhibits three roads with different colors. By considering the surrounding topography and landform, we can infer various characteristics of these roads solely based on the aerial images, such as road class and speed limits. At the end, apart from images and roads, we also query additional information and operate road graph for input data preparation.

Input data pre-processing We apply aerial images, GPS raster images, and corresponding road geometries as the inputs of our network. To receive input data at proper forms, we operate original images and map data. Specifically, we crop aerial images into 2048×2048 in spatial resolution, which refers to $614 \times 614 m^2$ area in real world. Besides, as the aerial images provide the geo-coordinates (i.e., longitude and latitude) for the top-left and bottom-right corners, we can crop the corresponding GPS data and road geometry from map data to align with aerial images. Last, the geo-coordinates for the key points of each road can be converted into image-coordinates, which are not necessary to be integers in our work. Since some roads are partitioned into multiple different image crops, we also keep the index of road geometry in addition to image-coordinates to predict attributes for entire roads during inference time, where a road can be very long winding one.

Building outline Road attributes are usually related to surrounding environment. For example, there are less building but more trees for high-way compared with the roads to a university or living areas. Therefore, we believe building outlines can provide useful clues to help network to encode more semantic contexts. Providing building outlines and road masks to our model, we aim to teach the network to recognize buildings and roads, in order to capture the correlation between road attributes for a specific road and other objects from the same area. Figure 3 draws some examples used in our work, that we can clearly see the distributions of buildings and roads are varying a lot for different regions and environments.

Road graph densification Map data providers offer the road geometry, indicating the location of individual roads on the earth. A road geometry is represented as a sequence of key points, that we leverage those key points to predict road attributes. Furthermore, instead of using the original key points from a map only, we also add new key points between the original points, which benefits for sampling more features of a road and representing a road better. For

example, in Figure 4, three roads are shown in the top image, where the map only constructs a road geometry by using the key points related to direction changes or intersections. We equidistantly insert some key points (i.e., K_1 to K_8) between the original key points, and apply our newly constructed road graph to learn attributes for individual roads defined in the original map (e.g., white, orange, and blue roads). In principle, we predict attributes for a complete road by predicting each small section (e.g., AK_3, AB, K_1K_2) and summarizing them. More details will be discussed in the rest of this section.

3.2 GPS rasterization

GPS trajectories collected from individual users provide an important aspect of the underlying road attributes. Since these are affecting the people's driving behavior, their usage in a road attribute inference model becomes fundamental. In our work, we are not interested in the individual behaviors but aim to infer their collective behavior over a road segment. Therefore, we aggregate all trajectories across a time span (e.g. a couple of months) into a raster image where each band represents a key motion behavior (speed, acceleration, turns, etc.). Figure 5 shows an example GPS raster where its bands are separately illustrated. The details of GPS raster generation is omitted from this paper since it is not our main contribution. However, [9] provides a detailed description about the topic.

Using GPS raster images provides us an opportunity to leverage GPS data with the same architecture as the one we have for remote sensing imagery. Thus, they provide statistics for trip counts over a road, average speed and acceleration, bearing changes, and trip direction in eight cardinal directions (N, NE, E, SE, S, SW, W, NW). These are concatenated with remote sensing imagery, forming images with 15 channels (bands), as the input of our road attribute learning model which support physical (e.g., road median, number of lanes) as well as motion related attributes.

3.3 CNN-based encoder

To learn features for individual roads from aerial images and GPS raster images, we concatenate them together to form images $X \in \mathbf{R}^{H\times W\times 15}$ with 15 channels and feed into a CNN backbone (i.e., ResNet-34 in our work) to extract multi-scale dense visual features at n stages $\{F_i\}_{i=1}^{n}$, where each feature is at different spatial resolution $F_i \in \mathbf{R}^{C_i\times H_i\times W_i}$ and downsampled by factor s_i compared with original image resolution.

Given dense feature maps, our next goal is to acquire representations for all input road segments. Our solution is based on feature extraction from dense maps for each key points, as shown in Figure 4. As we know the geo-coordinates (i.e., LAT, LON) for an entire image and every key point, we are able to convert the geo-coordinate of key points into image-coordinates (x, y), which provides location to sample features on $\{F_i\}_{i=1}^{n}$. In particular, F_i is spatially downsampled from X by a factor s_i, therefore, we extract features at i^{-th} stage at the location (x_i, y_i), where $x_i = \frac{x}{s_i}$ and $y_i = \frac{y}{s_i}$. Because of the downsampling and conversion from geo-coordinates to image-coordinates, (x_i, y_i) cannot guarantee to be integer. Therefore, we need to handle feature extraction at fractional locations. In detail, we apply bilinear interpolation to reach this goal, which is differentiable and allows for training the whole model end-to-end. Formally, the i^{-th} stage feature for key point at image location (x, y) can be computed by

$$\begin{aligned} feat_i(x,y) = {} & F_i(:, \lfloor x_i \rfloor, \lfloor y_i \rfloor) * (1 + \lfloor x_i \rfloor - x_i) * (1 + \lfloor y_i \rfloor - y_i) \\ & + F_i(:, 1 + \lfloor x_i \rfloor, \lfloor y_i \rfloor) * (x_i - \lfloor x_i \rfloor) * (1 + \lfloor y_i \rfloor - y_i) \\ & + F_i(:, \lfloor x_i \rfloor, 1 + \lfloor y_i \rfloor) * (1 + \lfloor x_i \rfloor - x_i) * (y_i - \lfloor y_i \rfloor) \\ & + F_i(:, 1 + \lfloor x_i \rfloor, 1 + \lfloor y_i \rfloor) * (x_i - \lfloor x_i \rfloor) * (y_i - \lfloor y_i \rfloor), \end{aligned} \tag{1}$$

where $\lfloor * \rfloor$ is a round-down operation. At the end, the feature from the CNN i^{-th} stage for a key point is a vector at the C_i dimension.

Because a key point might be shared with different road segments with possibly different GT, predicting attributes for individual key points would cause to ambiguous. To better predict road attributes for input road segments, where they can share key points in the map data, in particular, on the intersection areas (e.g., point "A" in the bottom plot of Figure 4). To overcome this issue, we represent a small road section by neighboring densified key points, based on key point features. Instead of predicting attributes for each key points, we produce prediction for each road section using the concatenation of neighboring key point features, such as K_1K_2, AK_4 in Figure 4. At the end, we also leverage multi-stage features from a CNN backbone, capturing different levels of representations. A simple way to produce attribute prediction is to learn a linear projection layer with all the feature concatenation. However, this solution lacks the consideration of road graphs and thus fails to learn the dependencies between different roads. In the following, we present a predominant decoder based on graph neural networks.

3.4 GNN-based decoder

After extracting multi-stage feature vectors for all the key points, we learn graph neural network (GNN) blocks to perform message passing through a road network structure. In this work, we adopt GraphSAGE [10] as the basic graph learning layer to construct our GNN block. Assuming the feature vectors with size of $C \times K$, where there are K key points in total, the GNN block outputs features with size of $C^* \times K$. The first goal of GNN is to propagate context information through the road graph. The second goal of applying GNN is to reduce the feature dimension from C to C^*. Alternatively, we can also learn a linear projection layer to reduce the feature dimension, however, we observe GNN achieves better accuracy. With the visual embeddings on key points, we perform prediction for individual road sections, which is a subsection of a road segment. Finally, we can learn a classifier or regressor for the prediction for each road section.

3.5 Model training

Dynamic weighting loss Given n tasks, our model calculates loss functions from n decoders, resulting in n loss functions $\mathcal{L}_1$, $\mathcal{L}_2$, ..., $\mathcal{L}_n$. We can apply loss weights for individual tasks, and the multi-task loss function can be written as

$$\mathcal{L}_{multi} = \sum_{i=1}^{n} w_i \cdot \mathcal{L}_i, \tag{2}$$

Table 1: Statistics of dataset and annotation in this work. The number of labeled road geometry and the number of pre-defined categories for each task are listed. We regard the road width prediction is a regression task, that we output 1-d scalars in meters.

Split	Location	# of Images	Type (6)	Median (3)	# of Lanes (8)	Directionality (2)	Speed Limit (14)	Width (1)
Train	San Diego	2,869	116,852	116,546	116,557	116,852	39,333	116,546
	Las Vegas	876	62,598	13,417	2,575	13,441	5,053	13,417
	Chicago	3,215	247,328	50,993	9,076	51,341	16,838	50,993
Test	Seattle	524	4,410	4,393	4,399	4,410	3,776	4,393
	Phoenix	746	18,932	18,932	18,932	18,932	2,423	18,932
	Miami	1,526	13,137	13,137	13,137	13,137	645	13,137

where w_1, w_2, ..., w_n are the loss weights for individual tasks. $\mathcal{L}_i$ is the loss for each task, which can be either the cross-entropy loss for classification or the Huber loss for regression in this work. We usually perform prediction using the feature concatenation of adjacent key points for the classification of each road section (e.g., AK_3, AK_4, AB, K_1K_2 in Figure 4). Those tasks may need significantly different loss weights for better overall performance, since the importance and ease of learning varies across tasks.

To avoid complicated hyperparameter tuning for different task combinations, we apply an uncertainty-based method [19] to learn the loss weights automatically and dynamically, which gives us the flexibility and performance guarantee to add new tasks or remove existing tasks. Rather than setting fixed weights for individual tasks, learning dynamic loss weights can help us to balance different tasks better, and our loss function can be formulated as

$$\mathcal{L}_{dynamic} = \sum_{i=1}^{n} e^{-2\sigma_i} \cdot \mathcal{L}_i + \sigma_i, \tag{3}$$

where σ_i is a learnable parameter for i^{-th} task. In particular, σ_i is a real number and initialized as 0 at the beginning of training. They are updated from back-propagation [25] with all the other model parameters together.

Regularization using semantic segmentation To enforce the network to encode more useful intermediate representations, we learn an additional semantic segmentation branch to output road and building masks. As discussed in Sec. 3.1, we provide building outline masks to our networks. Besides, road geometry can also be converted into binary road masks. As a result, we are able to learn semantic segmentation from the dense feature maps, which are shared with our attribute prediction decoder. Finally, the segmentation supervision can affect attribute prediction results. In this way, the network is able to learn more contextual dependencies and provide more information to determine road attributes. From our experiments, we observe learning a semantic segmentation as a regularization helps improve most attributes. In the following, the loss function to learn our final model is

$$\mathcal{L}_{final} = \sum_{i=1}^{n} (e^{-2\sigma_i} \cdot \mathcal{L}_i + \sigma_i) + w_{seg} \cdot \mathcal{L}_{seg}, \tag{4}$$

where w_{seg} is a fixed loss weight during training, and $\mathcal{L}_{seg}$ is the loss to classify each pixel as building, road, or background.

3.6 Model inference: segment-level prediction

Our model produces tabular outputs of road segment-level prediction, by partitioning aerial images into several small patches. As mentioned in Sec. 3.1, a road geometry may be partitioned into several different image patches, therefore, we predict attributes for input road geometries by running model inference on each patch and aggregating them. During pre-processing, we record the road index and average the predictions of all the sections with a same index as the final prediction. Particularly, we also augment the section prediction by reversing the concatenation order between adjacent key points, i.e., $(Pred(A, B) + Pred(B, A))/2$. It will not increase much time, as we only need to extract the dense feature for input images one time, which is the majority computation cost of our model.

4 DATASET PREPARATION

To build the model architecture and train our proposed neural network, we used the data based on OpenStreetMap (OSM) [3].

In order to reduce the need for hand-annotating data for model training purposes, we leverage the existing road attributes available at OSM. Thus, we get the road attributes for road segments from the US if they are available at OSM (most of those are not completely available). The primary road attributes we are interested in are road type, road median, number of lanes, road directionality, and speed limit. More specifically, we use road types in 6 categories including residential, primary, secondary, tertiary, service, footway. For road median, we have 3 categories, i.e. no median, legal median, and physical median. The lane numbers can be $1 \sim 8$. The directionality may be one-way or two-way. The speed limit can be from 5mph to 70mph, with a step of 5mph. Finally, we treat the road width prediction as a regression task, in that we aim to estimate each road in meters.

As OSM provides road type for most road geometries, we have the largest number of labeled data for road types, while other attributes have fewer labeled data. In order to increase the coverage of other attributes, we also annotate some road segments by observing street view images and aerial images. In particular, we collect all the annotations for road median since their completeness is critical and OSM has a very little coverage over these.

Table 1 lists the statistics in terms of availability of labels and the split used in this paper, that we train our attribute prediction model for US. The training data is collected from 3 cities: San Diego,

Table 2: Comparison performance of the proposed model against popular methods in our US dataset. The best and second-best performance are highlighted with the bold and underline fonts, respectively.

Location	Method	Type (%, ↑)	Median (%, ↑)	# of Lanes (%, ↑)	Directionality (%, ↑)	Speed Limit (%, ↑)	Speed Limit (mph, ↓)	Road Width ($meters$, ↓)
Seattle	ResNet-50 [11]	6.5	85.1	92.1	<u>94.4</u>	<u>26.9</u>	<u>5.05</u>	1.88
	Swin-B [23]	9.6	**87.5**	**93.5**	<u>94.4</u>	21.1	6.45	<u>1.77</u>
	RoadTagger [12]	<u>65.4</u>	83.5	88.5	90.2	26.5	5.09	1.85
	Ours (ResNet-34)	**78.6**	<u>86.5</u>	<u>92.7</u>	**94.6**	**27.6**	**4.94**	**1.71**
Phoenix	ResNet-50 [11]	80.3	91.5	<u>93.4</u>	93.8	66.1	2.75	2.79
	Swin-B [23]	<u>81.3</u>	**94.0**	**94.5**	<u>95.1</u>	**68.2**	<u>2.74</u>	<u>1.99</u>
	RoadTagger [12]	77.8	85.2	90.8	91.0	38.8	5.88	2.36
	Ours (ResNet-34)	**81.5**	<u>93.3</u>	92.8	**95.4**	<u>66.7</u>	**2.73**	**1.51**
Miami	ResNet-50 [11]	76.0	81.8	<u>96.5</u>	95.2	17.0	7.97	2.65
	Swin-B [23]	<u>77.5</u>	**84.1**	**97.3**	<u>95.5</u>	<u>25.4</u>	<u>7.51</u>	<u>2.84</u>
	RoadTagger [12]	69.5	79.9	91.4	90.4	19.5	7.72	2.43
	Ours (ResNet-34)	**79.2**	<u>83.6</u>	96.0	**95.7**	**26.5**	**6.17**	**1.60**

Las Vegas, and Chicago. We test our model on Seattle, Phoenix, and Miami.

As mentioned in Sec. 3.1, we apply cropped aerial images at size of 2048×2048, where the numbers of images for individual regions are also presented in Table 1. Last, we construct the GPS raster images in our work, which are aligned with aerial images and also at size of 2048×2048. For the image patches without any GPS data points, we create empty GPS raster images with a value of 0.

5 EXPERIMENTAL RESULTS

We depict the experimental results of our approach and provide the discussion in this section. We first conduct experiments using our collected dataset and split for the attributes listed in Table 1. We highlight that our approach predicts multiple road attributes with only one network. Besides, we also follow [31] to conduct experiments on Singapore using satellite images from DigitalGlobe.

5.1 Implementation details

We implement our multi-task deep neural network using PyTorch and PyG packages. We use 8 Tesla V100 GPU cards to train our model. The batch size on each GPU is 1, and we synchronize the distribution of normalization layers in the network. Therefore, it is equal to train the network with batch size 8. To build our model, we apply ImageNet-1K pretrained ResNet-34 [11] to create our encoder, and set 5 GNN blocks to build the decoder for message passing, where each block consists of a GraphSAGE operation [10], a normalization (LayerNorm) layer, and a ReLU layer. We apply AdamW optimizer to train our models with betas $(0.9, 0.999)$. For the decoders, we set the learning rate as $10\times$ compared with the encoder. At the end, we train our models for our collected dataset in US and Singapore dataset with 100,000 and 30,000 iterations respectively. We apply data augmentation to train our model, including vertical/horizontal flipping, and random image distortions in brightness, contrast, and saturation.

Comparison methods & evaluation protocol To demonstrate the effectiveness of our method, we set up several baselines and show comparisons with previous work [31, 12]. We train single-task image classification models as our baselines using different popular network backbones based on ImageNet-1K pretrained representation, including ResNet-50 [11] and Swin Transformer (Swin-B) [23].

Regarding the evaluation protocol, we report the segment-level classification accuracy over all the road segments with GT, i.e., $acc = \frac{\#\ of\ correct\ prediction}{\#\ of\ road\ segments}$, for road type, road median, number of lanes, road directionality, speed limit. For speed limit, we report the mean absolute errors (MAE) in mph, as this error complements to classification accuracy. Last, we also report the MAE in meters for road width prediction, as road width prediction is a regression task.

5.2 Results

Results on aerial imagery from US In Table 2, we show our experimental results on multi-task attribute prediction for our collected dataset and compare with other models in Table 2, where evaluation is conducted on three areas from Seattle, Phoenix, and Miami. For comparison methods [11, 23, 12], we train separate models for individual tasks with public source codes, that they treat attribute prediction as an image classification problem. For ResNet-50, it is a stronger backbone from the same model family as ours (i.e., ResNet-34). For Swin-B, it is a state-of-the-art network from the popular Vision Transformer (ViT) family, which figures out the cutting-edge image classification performance. For RoadTagger, it is similar to our method in terms of the utilization of CNN and GNN, but their model is trained under image classification framework with more resource costs and less context modeling capability. To have a fair comparison, we apply the image patches at 224×224 and align the key points used in our model as the image centers. Eventually, the number of training images is much larger than the number of road segments. As a result, we use ~1.5 million image/GPS pairs to train separate image classification models for each attribute.

From this table, we observe that our model outperforms other methods in all the locations, which clearly demonstrate the effectiveness of our approach based on dense feature extraction and

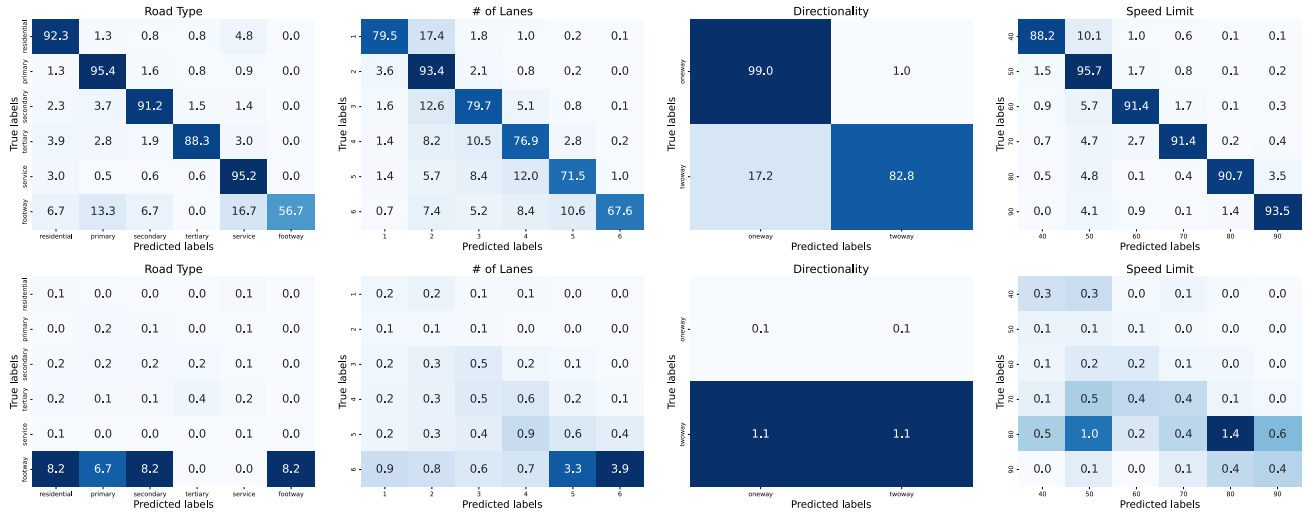

Figure 6: Confusion matrices of our Image+GPS model for Singapore. We report the mean and the stand deviation (both in %) of 5 runs, at the top and bottom rows.

Table 3: Comparison results with previous multi-modal road attribute prediction method [31] on Singapore. The best and second-best performance are highlighted with the bold and underline fonts respectively.

Method	Image	GPS	Type	# of Lanes	One/Two Way	Speed Limit
[31]	✓		69.42	64.30	77.78	73.74
		✓	70.75	68.65	82.97	80.84
	✓	✓	79.04	71.32	85.46	82.42
Ours	✓		87.93	84.20	<u>98.19</u>	89.27
		✓	<u>93.51</u>	<u>86.45</u>	98.14	**93.24**
	✓	✓	**93.73**	**87.27**	**98.20**	**93.42**

sampling on key points. Modeling wide contexts, our model is able to predict various attributes at more than 90% accuracy, except for speed limit. We observe the speed limit accuracies for Seattle and Miami are less than 30%. To explain this, we believe the data distribution gap between training areas (i.e., San Diego, Las Vegas, Chicago) and Miami might be very different. In contrast, looking at the speed limit performance in Phoenix, we can see ~65% is reached, where a possible reason is that the landform similarity between Phoenix and Las Vegas. In addition to classification accuracy, we can observe that the MAEs are 4.94mph, 2.73mph, and 6.17mph for the three testing areas. We would like to emphasize that the minimum difference in speed limit is 5mph in our dataset, therefore, the prediction still shows some reasonable and promising results, as the MAEs on Seattle and Phoenix are less than 5mph.

Comparing our model with other methods, we can observe the proposed model achieves the best results for most attributes among the all, even though our model leverages a smaller backbone and does a more complex task to predict multiple attributes, which clearly show the effectiveness of our framework based on road section prediction and key point-level feature extraction. Furthermore, we highlight the following observations. (1) Our model is more stable to all the attributes. For example, ResNet-50 and Swin-B fails to recognize road type in Seattle, even they also obtain competing results in another two stations. We interpret this as a necessity of wider context modeling. Only focusing on local patches may degenerate during inference time. In contrast, wider context can provide richer information for more robust and precise prediction. (2) For the attributes that images can provide strong clues, image classification solution can also reach to state-of-the-art performance, even though road structures and contextual information are ignored. For example, we observe ResNet-50 and Swin-B achieves very high accuracy in road median and the number of lanes. Besides, they even reach higher performance than our model in those two attributes and Swin-B achieve the best results in road median and lane number prediction. These results demonstrate a stronger network backbone for visual inference is quite important, as we learned from the past research. On the contrary, our framework is still comparative and a little better than ResNet-50 by summarizing the road median and lane number accuracies in the three areas. (3) Our model achieves much better results than others in speed limit and road width, in spite of the multi-attributes capability from our model, that we conclude our method as a highly efficient and effective framework.

Results on satellite imagery from Singapore We show the results using DigitalGlobe satellite imagery and GPS in Table 3. For a comprehensive analysis, we train several versions using different inputs of satellite imagery, GPS rasters, and their combination. Following [31], we conduct road attribute prediction on Singapore, including road type, number of lanes, road directionality (i.e., one-way or two-way), and speed limit. We apply the same label space to [31], and train our multi-task model, which produces all the attributes with only one network. We query the OSM in the version of 210412 with a similar number of road segments to [31], which was published in 2021, and randomly partition into training and testing set with the 80%/20% split. Because the split used in [31] is not public, we generate 5 random splits to train our models, where

Table 4: Ablation study w.r.t regularization (+R), dynamic weighting loss (+D), and graph neural networks (+G). The best and second-best performance are highlighted with the bold and underline fonts respectively.

Location	+R	+D	+G	Type (%, ↑)	Median (%, ↑)	# of Lanes (%, ↑)	Directionality (%, ↑)	Speed Limit (%, ↑)	Speed Limit (mph, ↓)	Road Width ($meters$, ↓)
Seattle				**79.8**	85.9	92.5	94.0	25.2	5.18	1.79
	✓			77.7	85.7	92.6	94.2	25.9	5.09	1.73
	✓	✓		78.5	85.9	**92.7**	94.3	26.1	5.00	1.73
			✓	78.2	86.6	92.3	94.0	25.2	5.20	1.77
	✓		✓	78.7	**86.8**	92.4	94.2	26.6	5.07	1.73
	✓	✓	✓	78.6	86.5	**92.7**	**94.6**	**27.6**	**4.94**	**1.71**
Phoenix				81.4	93.0	**92.8**	95.1	62.7	4.69	1.65
	✓			81.6	93.1	92.7	**95.4**	65.7	3.04	1.59
	✓	✓		81.5	93.1	92.7	**95.4**	65.5	2.98	1.58
			✓	81.4	92.9	**92.8**	95.3	65.9	2.90	**1.51**
	✓		✓	**81.9**	92.9	92.5	95.3	66.6	**2.72**	1.52
	✓	✓	✓	81.5	**93.3**	**92.8**	**95.4**	**66.7**	2.73	**1.51**
Miami				78.5	82.5	95.8	**95.7**	15.8	12.14	1.85
	✓			78.5	82.8	95.8	95.5	26.5	6.81	1.68
	✓	✓		78.8	83.0	95.8	95.6	**27.4**	**6.14**	1.65
			✓	78.1	82.4	**96.0**	95.6	20.5	6.35	1.63
	✓		✓	79.0	82.7	**96.0**	95.6	23.3	6.25	**1.59**
	✓	✓	✓	**79.2**	**83.6**	**96.0**	**95.7**	26.5	6.17	1.60

we report the average performance in 3. In particular, we do not apply the semantic segmentation regularization in the Singapore experiments, to have a fair comparison with previous work. We train three model variants with different inputs, including image-only model, GPS-only model, and image+GPS model.

In Table 3, we observe the proposed method achieves significantly better accuracy compared with [31] with all kinds of inputs. Specifically, our worst version, image-based model, is already much better than [31] using image+GPS inputs. Our performance for road type, directionality, and speed limit can reach to 90%+ accuracy. Besides, our accuracies of three models for number of lanes are also higher than 80%, which comparison method only achieves 64.30%, 68.65%, and 71.32% accuracy. Finally, we also report the confusion matrices for individual tasks in Figure 6. In this figure, we show the average accuracy and standard deviation from all 5 runs. From the standard deviations, we observe the performances for most categories are quite stable, even though the splits are randomly generated. The uncertainty of our model performance mainly comes from the rare classes, for example footway in road type, 6 lanes in number of lanes, and 80 km/h in speed limit, which is a high speed situation. At the end, we conclude the effectiveness of the proposed method, in obtaining strong attribute prediction performance, even though the efficiency is a contribution of our work.

5.3 Discussion

To understand the impact of different components of our model, we disable different components of the proposed model, and discuss ablation studies. Specifically, we remove the regularization branch of road and building masks, dynamic weighting loss in combining multiple tasks, and GNN blocks for message passing over key points on a road graph. For replacing GNN, we directly learn a linear projection layer to predict attributes. Table 4 compares all the combinations. Overall, we observe all the models achieves decent performance in most road attributes, which shows the effectiveness of our framework that leverage a large area and learn the relations between all the road segments. Besides, our final model achieves the most stable performance in all the attributes. Even though it is not always the best, the final model only performs slightly worse than the best case and is more balanced than our other versions. For example, our vanilla version achieves 79.8% accuracy in road type, but its performance of speed limit and road width is the worst among the all.

Regarding our ablation study, we would like to highlight several observations and insights. (1) GNN blocks help to propagate the information through the key points in a road graph, which shows improved results, especially for road median and road width. By leveraging neighboring road context, all the versions with GNN in those two attributes are better than their counterparts. For example, in Phoenix and Miami, the prediction errors of road width are clearly reduced. (2) Segmentation regularization with building/road masks benefits speed limit and road width. By learning image-level context with the regularization, speed limit and road width are always improved, which demonstrates the correlation between road attributes and surrounding environment. In particular, we can see that the baseline speed limit errors in Miami is 12.14mph, while it is reduced to 6.81 after applying the regularization. (3) Dynamic weighting loss is also helpful to achieve more stable overall performance. Especially, it provides a simple hyperparameter tuning strategy. We draw the learned loss weights of our final model in Figure 7. From this figure, we observe we need a larger weight for road directionality than other attributes. In contrast, the loss

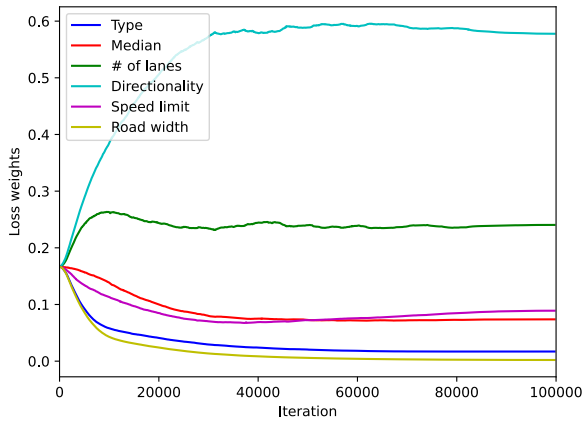

Figure 7: Learned dynamic weights for balancing different road attributes on our dataset.

weight for road width is supposed to be very small. Interestingly, a small weight for road width even leads to higher performance in predicting road width, which is opposite to the intuition. Those results clearly show the usefulness of our model and design.

6 CONCLUSION

We present a novel road attribute learning framework that utilizes multi-task deep learning to construct and enrich digital maps. Our proposed model combines the power of convolutional neural networks and graph neural networks, leveraging satellite imagery and GPS data as input for analyzing road structures. We point out several key insights in this work. Our approach enables simultaneous generation of multiple road attributes for all road segments within a given area. This not only improves efficiency but also captures the interdependencies among different road segments and their attributes, leading to better performance than single prediction. Besides, we show contextual information needs to take enough consideration for this task, that we observe building information can help to improve some attributes, like speed limit, indicating the correlation between road attributes and surrounding environments. At the end, the proposed model effectively handles both classification and regression tasks, allowing us to predict 6 road attributes, including road type, road median, number of lanes, road directionality, speed limit, and road width. Through thorough comparison and extensive ablation studies, we demonstrate substantial advancements over previous state-of-the-art methods and numerous baselines. We believe our proposed method is highly deployable and capable of enhancing digital map quality at a large scale, owing to its exceptional efficiency and effectiveness.

REFERENCES

[1] Heba Aly and Moustafa Youssef. 2015. Semmatch: road semantics-based accurate map matching for challenging positioning data. In *ACM SIGSPATIAL*.
[2] Gaetan Bahl, Mehdi Bahri, and Florent Lafarge. 2021. Road extraction from overhead images with graph neural networks. *arXiv preprint arXiv:2112.05215*.
[3] Jonathan Bennett. 2010. *OpenStreetMap*. Packt Publishing Ltd.
[4] Khac-Hoai Nam Bui, Jiho Cho, and Hongsuk Yi. 2022. Spatial-temporal graph neural network for traffic forecasting: an overview and open research issues. *Applied Intelligence*.
[5] Yihua Chen and John Krumm. 2010. Probabilistic modeling of traffic lanes from gps traces. In *SIGSPATIAL*.
[6] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2015. Gated feedback recurrent neural networks. In *ICML*.
[7] Lei Ding and Lorenzo Bruzzone. 2020. Diresnet: direction-aware residual network for road extraction in vhr remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*.
[8] Alexey Dosovitskiy et al. 2021. An image is worth 16x16 words: transformers for image recognition at scale. In *ICLR*.
[9] Emre Eftelioglu, Ravi Garg, Vaibhav Kango, Chintan Gohil, and Amber Roy Chowdhury. 2022. Ring-net: road inference from gps trajectories using a deep segmentation network. In *ACM SIGSPATIAL Workshop*.
[10] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *NeurIPS*.
[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.
[12] Songtao He et al. 2020. Roadtagger: robust road attribute inference with graph neural networks. In *AAAI*.
[13] Yang He, Ravi Garg, and Amber Roy Chowdhury. 2022. Td-road: top-down road network extraction with holistic graph construction. In *ECCV*.
[14] Chidubem Iddianozie and Gavin Mcardle. 2021. Transferable graph neural networks for inferring road type attributes in street networks. *IEEE Access*.
[15] Zohaib Jan, Brijesh Verma, Joseph Affum, Sam Atabak, and Lachlan Moir. 2018. A convolutional neural network based deep learning technique for identifying road attributes. In *IVCNZ*.
[16] Tobias Skovgaard Jepsen, Christian S Jensen, and Thomas Dyhre Nielsen. 2019. Graph convolutional networks for road networks. In *ACM SIGSPATIAL*.
[17] Weiwei Jiang and Jiayun Luo. 2022. Graph neural network for traffic forecasting: a survey. *Expert Systems with Applications*.
[18] Marin Kačan, Marin Oršić, Siniša Šegvić, and Marko Ševrović. 2020. Multi-task learning for irap attribute classification and road safety assessment. In *ITSC*.
[19] Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*.
[20] Konstantin Klemmer, Nathan S Safir, and Daniel B Neill. 2023. Positional encoder graph neural networks for geographic data. In *UAI*.
[21] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*.
[22] Vasileios Linardos, Maria Drakaki, Panagiotis Tzionas, and Yannis L Karnavas. 2022. Machine learning in disaster management: recent developments in methods and applications. *Machine Learning and Knowledge Extraction*.
[23] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: hierarchical vision transformer using shifted windows. In *ICCV*.
[24] Saeed Rahmani, Asiye Baghbani, Nizar Bouguila, and Zachary Patterson. 2023. Graph neural networks for intelligent transportation systems: a survey. *IEEE Transactions on Intelligent Transportation Systems*.
[25] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *Nature*.
[26] Pubudu Sanjeewani and Brijesh Verma. 2021. Single class detection-based deep learning approach for identification of road safety attributes. *Neural Computing and Applications*.
[27] Tao Sun, Zonglin Di, and Yin Wang. 2018. Combining satellite imagery and gps data for road extraction. In *ACM SIGSPATIAL Workshop*.
[28] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*.
[29] Jingjing Yan, Shunping Ji, and Yao Wei. 2022. A combination of convolutional and graph neural networks for regularized road surface extraction. *IEEE transactions on geoscience and remote sensing*.
[30] Wei Yang, Tinghua Ai, and Wei Lu. 2018. A method for extracting road boundary information from crowdsourcing vehicle gps trajectories. *Sensors*.
[31] 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 *ACM SIGSPATIAL*.
[32] 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 *WWW*.
[33] 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 *CVPR Workshops*.