2019

Image-Based Roadway Assessment Using Convolutional Neural Networks

Weilian Song

University of Kentucky, weiliansong19@gmail.com
Author ORCID Identifier:
https://orcid.org/0000-0002-8357-4995
Digital Object Identifier: https://doi.org/10.13023/etd.2019.136

Image-Based Roadway Assessment Using Convolutional Neural Networks

––––––––––––––––––––––––

THESIS

––––––––––––––––––––––––

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science in the
College of Engineering
at the University of Kentucky

By

Weilian Song

Lexington, Kentucky

Director: Dr. Nathan Jacobs, Associate Professor of Computer Science

Lexington, Kentucky

2019

ABSTRACT OF THESIS

Image-Based Roadway Assessment Using Convolutional Neural Networks

Road crashes are one of the main causes of death in the United States. To reduce the number of accidents, roadway assessment programs take a proactive approach, collecting data and identifying high-risk roads before crashes occur. However, the cost of data acquisition and manual annotation has restricted the effect of these programs. In this thesis, we propose methods to automate the task of roadway safety assessment using deep learning. Specifically, we trained convolutional neural networks on publicly available roadway images to predict safety-related metrics: the star rating score and free-flow speed. Inference speeds for our methods are mere milliseconds, enabling large-scale roadway study at a fraction of the cost of manual approaches.

KEYWORDS: urban understanding, transportation engineering, roadway safety, deep learning

Author's signature: Weilian Song

Date: May 2, 2019

Image-Based Roadway Assessment Using Convolutional Neural Networks

By

Weilian Song

Director of Thesis: _____ Nathan Jacobs _____

Director of Graduate Studies: __ Miroslaw Truszczynski __

Date: _____ May 2, 2019 _____

To Rachel, for being the brightest star in my life.

ACKNOWLEDGMENTS

I would like to express my sincere thanks to Dr. Nathan Jacobs for being my advisor. The world of academic research was quite daunting at first, but his enthusiasm and expert guidance have propelled me forward all the way to today. My degree and accomplishments would not have been possible without him.

I would also like to thank Dr. Brent Harrison and Dr. Hasan Poonawala for being on my committee, and for fruitful discussions outside of class. Their expertise in reinforcement learning and robotics control respectively has expanded my knowledge scope and made me think more broadly about applications of deep learning.

Of course, my time at the University of Kentucky is not just about research, and I would like to thank Dr. Judy Goldsmith and Dr. Jerzy Jaromczyk for making my experience here more colorful. Dr. Goldsmith's thought provoking questions have helped shape my decision to pursuit a PhD, and Dr. Jaromczyk has introduced me to the awesome world of competitive programming.

Research is a team effort, and I want to thank my fellow lab members and co-authors for being part of my team. Thank you, Dr. Scott Workman, Armin Hadzic, Tawfiq Salem, Hunter Blanton, Dr. Mei Chen, and many others. You guys are awesome!

Special recognition goes to everyone at the UK Solar Car Team. I have gained much technical and leadership experience from wonderful individuals on the team, including Ethan Toney, Chris Heintz, Monon Rahman, Thomas Barber, and many others.

Lastly, I would like to thank my family and friends for supporting me on my journey. You all have shaped me in more ways than one, and I am forever grateful for that.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

1.1 Importance of Roadway Safety

Development of roads and motorized vehicles has revolutionized our daily lives, reducing travel times and connecting the world in more ways than one. However, as vehicles become cheaper and more performant, the lethality of them has increased as well. Just in the US, over $35\,000$ people died in road crashes in 2015 [3], with millions more injured or disabled. From an economic point of view, road crashes are a significant burden as well, costing over 200 billion dollars [2]. Highway crashes have indeed become one of the primary cause of accidental deaths in the US [41].

With the staggering number of deaths and expense wasted, governments all around the world are investing billions of dollars in road infrastructure improvements. However, with millions of miles of roads, the funding needs to be efficiently distributed to roads that need the most repair. In fact, more than half of the road deaths and serious injuries occur on less than 10% of the road lengths in the world. [3] As a response, many road assessment programs have been founded, using data-driven approaches to discover high-risk road segments and intersections.

Efforts made by these road assessment programs have lead to a reduction in the number

Figure 1.1: Some roads are safer than others.

of fatalities observed [40]. Unfortunately, this reduction has plateaued in recent years. The main challenge that programs face is the inefficiency of conducting safety assessment. It is a costly process to collect raw data (road imagery, degree of incline, LiDAR data), and then perform manual annotation of road attributes such as traffic volume, pavement condition, and roadside hazards. Services like Google Streetview and OpenStreetMap have offered significant time savings when it comes to raw data acquisition, but annotation of data remains a manual process, as it requires a human-level understanding of road data.

1.2 Roadway Assessment using Deep Learning

Thanks to advancements in deep learning technologies, computer agents have achieved near human-level performance in many tasks, such as scene classification and object recognition. Adoption of these technologies in roadway applications have also become popular, due to neural networks' powerful ability at extracting high-level information from imagery data. In this thesis, we propose methods to automate the annotation process of roadway

safety assessment using deep convolutional neural networks. We built two safety-related image datasets–one on Star Ratings and the other on free-flow speeds–and trained neural networks to predict the two metrics. We evaluate our methods both quantitatively and qualitatively, and discuss possible applications of our work in addressing roadway safety issues.

1.3   Thesis Overview

In Chapter 2, we consider the task of predicting the Star Rating of a road from its ground-level imagery. The Star Rating metric is used by the United States Road Assessment Program to rate roads on a five-star scale, which roughly indicates the likelihood and cost of a crash at a particular location. In Chapter 3, we consider the task of predicting the free-flow speed of a road using aerial-level imagery. Free-flow speed is the expected average speed of drivers on a given road, an important metric to consider for understanding driver behavior. Finally in Chapter 4, we discuss future direction of roadway safety assessment, and the in-progress work of combining ground and aerial image modalities for prediction.

Chapter 2

Automated Safety Assessment

2.1 Introduction

An emerging technique for systemically addressing the highway crash problem is represented by the protocols of the United States Road Assessment Program (usRAP). In this process, a trained coder annotates, at regular intervals, various features of the roadway such as roadway width, shoulder characteristics and roadside hazards, and the presence of road protection devices such as guardrail. These annotations are based on either direct observation or from imagery captured in the field and can be used to augment any existing highway inventory of these features. These data are then used to rate the roadway, and in turn, these ratings may be classified into a 5-tier star-rating scale, from highest risk (1 star) to lowest (5 stars). This manual process is laborious and time consuming, and sometimes cost prohibitive. Moreover, the speed and accuracy of the rating process can and does vary across coders and over time.

To automate this manual process, we propose a deep convolutional neural network (CNN) architecture that directly estimates the star rating from a ground-level panorama. We name our work FARSA, Fully Automated Roadway Safety Assessment. See Figure 2.1 for examples of input panoramas. Note the lack of a physical median, paved shoulder, and

4
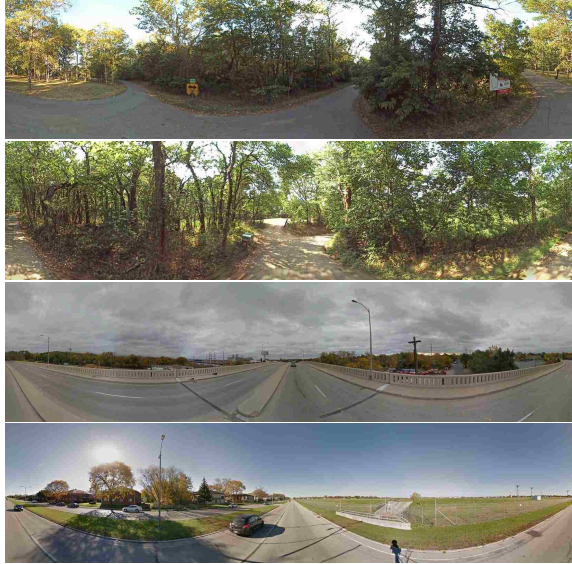
Figure 2.1: Example panoramas that our system uses to estimate the safety star rating. The top two are rated 1 star, the most dangerous. The bottom two are rated 5 star, the safest.

sidewalks in the less safe roads. The key features of our approach are:

- *multi-task learning*: we find that augmenting the network to support estimating lower-level roadway features improves performance on the main task of estimating the star rating;

- *task-dependent spatial attention layer*: this allows the network to focus on particular regions of the panorama that are most useful for a particular task; and

- *unsupervised learning*: we add a loss that encourages the star rating distribution to be similar for nearby panoramas, greatly expanding the number of panoramas seen by the network, without requiring any manual annotation.

We evaluate this approach on a large dataset of real-world imagery and find that it outperforms several natural baselines. Lastly, we present an ablation study that demonstrates the benefits of the various components of our approach.

2.2   Related Work

Our work builds upon work in several areas: general purpose scene understanding, automatic methods for understanding urban areas, and current practice in the assessment of roadway features.

Scene Classification and Image Segmentation   Over the past ten years, the state of the art in scene classification has dramatically improved. Today, most, if not all, of the best performing methods are instances of deep convolutional neural networks (CNNs). For many tasks, these methods can estimate the probability distribution across hundreds of classes in milliseconds at human-level accuracy [13]. A notable example is the *Places* CNN [45], developed by Zhou et al., which adapts a network that was created for image classification [18]. This network, or similar networks, has been adapted for a variety of tasks, including: horizon line estimation [42], focal length estimation [37], geolocalization [21], and a variety of geo-informative attributes [19].

This ability to adapt methods developed for scene classification and image segmentation to other tasks was one of the main motivations for our work. However, we found that naively applying these techniques to the task of panorama understanding did not work well. The main problem is that these methods normally use lower resolution imagery which means they cannot identify small features of the roadways that have a significant impact on the assessment of safety. We propose a CNN architecture that overcomes these problems by incorporating a spatial attention mechanism to support the extraction of small image features.

In some ways, what we propose is similar to the task of semantic segmentation [8], which focuses on estimating properties of individual pixels. The current best methods are all CNN-based and require a densely annotated training set. Constructing such datasets is a labor intensive process. Fortunately, unlike semantic segmentation, we are estimating a high-level attribute of the entire scene so the effort required to construct an annotated

dataset is lower. It also means that our CNNs can have a different structure; we extract features at a coarser resolution and have many fewer loss computations. This means faster training and inference.

Urban Perception and High-Definition Mapping Recently there has been a surge in interest for applying techniques for scene classification [5, 9, 24, 26, 31, 38, 39] and image segmentation to understanding urban areas and transportation infrastructure [10, 23, 36]. The former focuses on higher-level labels, such as perceived safety, population density, or beauty, while the later focuses on finer-grained labels, such as the location on line markings or the presence of sidewalks. Our work, to some extent, combines both of these ideas. However, we focus on estimating the higher-level labels and use finer-grained labels as a form of multi-task learning. Our evaluation demonstrates that by combining these in a single network enables better results.

Current Practice in Roadway Assessment The Highway Safety Manual (HSM) [16] outlines strategies for conducting quantitative safety analysis, including a predictive method for estimating crash frequency and severity from traffic volume and roadway characteristics. The usRAP Star Rating Protocol is an internationally established methodology for assessing road safety, and is used to assign road protection scores resulting in five-star ratings of road segments. A star rating is partly determined based on the presence of approximately 60 road safety features [12]. More implementation of road safety features entails a higher safety rating, and vice versa. There are separate ratings for vehicle occupants, cyclists, and pedestrians, but we focus on vehicle occupant star ratings in this paper.
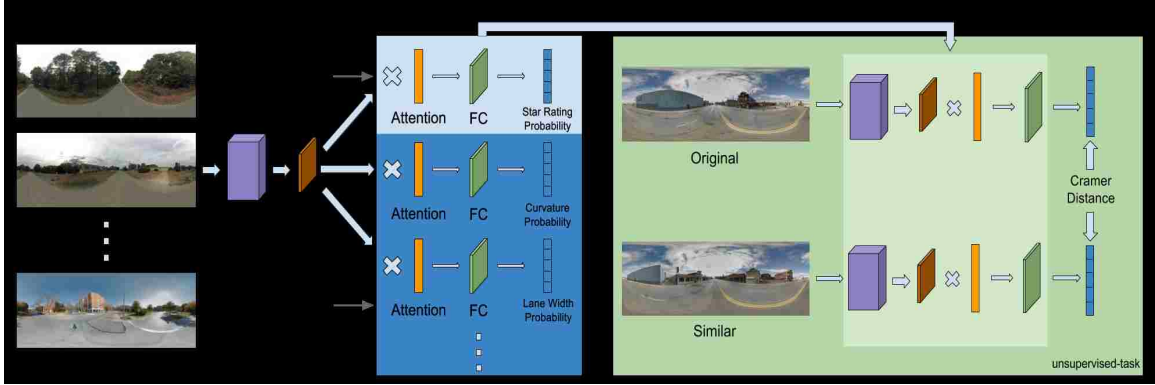
Figure 2.2: Overview of our network architecture.

2.3 Approach

We propose a CNN architecture for automatic road safety assessment. We optimize the parameters of this model by minimizing a loss function that combines supervised, multi-task, and unsupervised component losses. We begin by outlining our base architecture, which we use in computing all component loss functions.

2.3.1 Convolutional Neural Network Architecture

Our base CNN architecture takes as input a street-level, equirectangular panorama (*e.g.*, from Google Street View) and outputs a categorical distribution over a discrete label space. Our focus is on the roadway safety label space, which is defined by usRAP to have five tiers. Other label spaces will be defined in the following section. In all experiments, panoramas are cropped vertically, to reduce the number of distorted sky and ground pixels, and then resized to be 224×960.

The CNN consists of a portion of the VGG architecture [33], followed by a 1×1 convolution, a spatial attention layer, and a final fully connected layer. See Figure 2.2 for a visual overview of the full architecture and the various ways we use it to train our model. We use the VGG-16 architecture [33] for low-level feature extraction. It consists of 5 convolutional blocks followed by 3 fully connected layers, totaling 16 individual layers.

We remove the fully connected layers and use the output of the last convolutional block, after spatial pooling. We denote this output feature map as S_1, which is a 7×30 tensor with 512 channels. S_1 is then passed to a convolutional layer (ReLU activation function), with kernels of size 1×1, a stride of 1, and 512 output channels. The resulting tensor, S_2, has the following shape: $7 \times 30 \times 512$. This gives us a set of mid-level features that we use to predict the safety of a given panorama.

As the target safety rating may depend on where in the image a particular roadway feature is present, we introduce an attention layer to fuse the mid-level features. Specifically, we use a learnable vector a that takes the weighted average of S_2 across the first two dimensions. Our process is as follows: we flatten the spatial dimensions of S_2 (210×512) and multiply it by a (1×210). This process is akin to global average pooling [20], but with location-specific weights. The output (1×512) is then passed to a task-dependent fully connected layer with K outputs.

2.3.2 Loss Function

A key challenge in training large neural network architectures, especially with small datasets, is avoiding overfitting. For this work, we propose to use a combination of supervised, multi-task, and unsupervised learning to train our model. The end result is a total loss function:

$$L = \lambda_s L_s + \lambda_m L_m + \lambda_u L_u.$$

Each component loss function processes panoramas using the base architecture defined in the previous section. In all cases, the parameters of the VGG sub-network and the subsequent 1×1 convolution are tied. The attention and final fully connected layers are independent and, therefore, task specific. The remainder of this section describes the various components of this loss in detail.

Supervised Loss

The first component, L_s, of our total loss, L, corresponds to the main goal of our project: estimating the safety of a given roadway from an input panorama. Each panorama is associated with a star rating label, $l \in \{1, \ldots, 5\}$. We apply the network defined in Section 2.3.1 with $K = 5$ outputs, representing a categorical distribution over star ratings. Our first loss, L_s, incorporates both a classification and regression component. For classification, we use the standard cross-entropy loss between the predicted distribution, $\hat{y}$, and target distribution, y:

$$L_{s^1} = -\frac{1}{N} \sum_{i=1}^{N} y_i(l_i) \log \hat{y}_i(l_i) \tag{2.1}$$

where N is the number of training examples. For regression, we use the Cramer distance between $\hat{y}$ and y:

$$L_{s^2} = \frac{1}{N} \sum_{i=1}^{N} \|F(\hat{y}_i) - F(y_i)\|_2^2 \tag{2.2}$$

where $F(x)$ is the cumulative distribution function of x. Each component is weighted by λ_{s^1} and λ_{s^2}, respectively:

$$L_s = \lambda_{s^1} L_{s^1} + \lambda_{s^2} L_{s^2}. \tag{2.3}$$

Multi-task Loss

The second component, L_m, of our total loss, L, represents a set of auxiliary tasks. We selected M auxiliary tasks with discrete label space for learning, specifically: area type, intersection channelization, curvature, upgrade cost, land use (driver/passenger-side), median type, roadside severity (driver/passenger-side distance/object), paved shoulder (driver/passenger-side), intersecting road volume, intersection quality, number of lanes, lane width, quality of curve, road condition, vehicle parking, sidewalk (passenger/driver-side), and facilities for bicycles. All images were annotated by a trained coder as part of the usRAP protocol.

For each new task, the prediction process is very similar to the safety rating task, with its own attention mechanism and final prediction layer. The only difference is the output size of the prediction layer which varies to match the label space of the specific task.To compute L_m, we sum the cross-entropy loss across all tasks: $L_m = \sum_{t=1}^{M} L_t$ where L_t is the loss for task t.

Unsupervised Loss

The third component, L_u, of our total loss, L, represents Tobler's First Law of Geography: "Everything is related to everything else, but near things are more related than distant things." Given two geographically close street-level panoramas, their star ratings should be relatively close to each other, due to similar road safety features present in both imagery. To satisfy this constraint, we encourage the network to produce identical output distributions for adjacent panoramas. Empirically, we find that optimizing for the unsupervised loss improves the accuracy of our final network.

The key feature of this loss is that it does not require the panoramas to be manually annotated. Therefore, we can greatly expand our training set size by including unsupervised examples. This is important, because due to the small size of the safety rating dataset and a large number of parameters in the network, we found it impossible to update VGG layer weights without overfitting when using only supervised losses.

To define this loss, we first build a set, $U = \{(a_i, b_i)\}$ of panorama pairs. The pairs are selected so that they are spatially close (within 50 feet) and along the same road. For each pair, we compute the Cramer distance between their rating predictions, $(\hat{y}_a, \hat{y}_b)$, as in:

$$L_u = \frac{1}{|U|} \sum_{(a,b) \in U} \|F(\hat{y}_a) - F(\hat{y}_b)\|_2^2. \tag{2.4}$$

2.3.3 Implementation Details

Our network is optimized using ADAM [17]. We initialize VGG using weights pre-trained for object classification [28]. We experimented with weights for scene categorization [44]

and found that object classification was superior, and both were significantly better than random initialization. We allow the final convolutional block of VGG to optimize with learning rate 0.0001, while all task-specific layers use a learning rate of 0.001. We decay the learning rates three times during training, by a factor of 10 each time.

Through experimentation, we find that optimizing the total loss with $\lambda_{s^1} = 1$, $\lambda_{s^2} = 100$, $\lambda_s = 0.1$, $\lambda_m = 1$, and $\lambda_u = 0.001$ offers the best results. We use ReLU activations throughout, except for attention mechanisms and final output layers, which apply a softmax function. For every trainable layer other than the attention mechanisms, we apply L_2 regularization to the weights with a scale of 0.0005. We train with a mini-batch size of 16, with each batch consisting of 16 supervised panoramas and labels along with 16 pairs of unsupervised panoramas.

2.4   Evaluation

We evaluate our methods both quantitatively and qualitatively. Below we describe the datasets used for these experiments, explore the performance of our network through an ablation study, and present an application to safety-aware vehicle routing.

2.4.1   Datasets

For supervised and multi-task training, we utilize a dataset annotated through the U.S. Road Assessment Program (usRAP), which contains star ratings and auxiliary task labels for 100-meter segments of roads in both *Urban* and *Rural* regions. To obtain the labels for each location, a trained coder visually inspects the imagery for a road segment and assigns multi-task labels using a custom user interface. During this process the coder is free to adjust the viewing direction. The star rating for each location is then calculated from the auxiliary labels using the Star Rating Score equations. For more information on dataset annotation methods, please refer to [1] [12] [25].
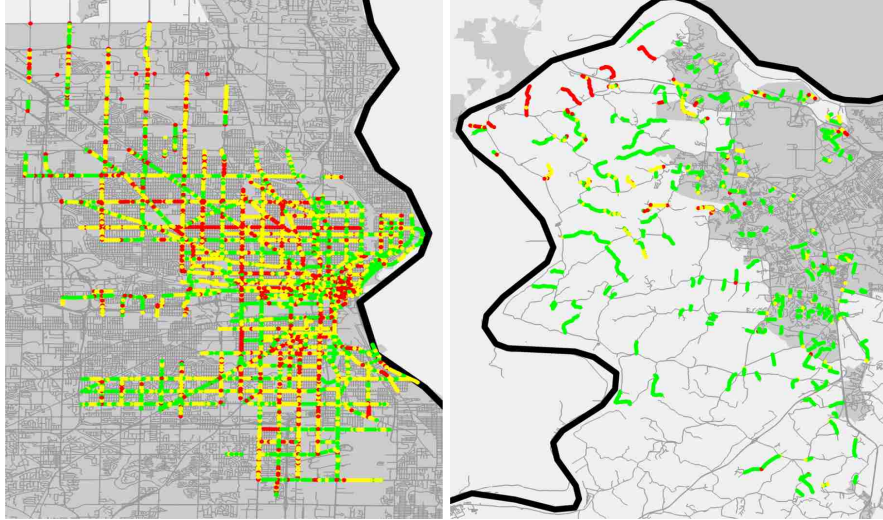
Figure 2.3: Maps showing the locations and star ratings (red is highest risk, yellow is average risk, and green is least risk) of the panoramas in our *Urban* (left) and *Rural* (right) study regions.

The *Rural* area has 1,829 panoramas and the *Urban* area has 5,459 panoramas, for a total of 7,288 samples. Figure 2.3 shows scatter plots of panorama locations for each region, color coded by the safety score manually estimated from the imagery. For unsupervised training data, we uniformly sample road segments from the state surrounding the *Rural* region and then query for panoramas that are less than 50-feet apart along the segment. The result is a dataset of approximately 36,000 pairs of panoramas.

2.4.2 Preprocessing

Panorama Processing For each sample, we download an orbital panorama through Google Street View. We orient panoramas to the heading used during manual coding. This is important because safety ratings are sensitive to the direction of travel. During training only, random augmentation is performed by randomly jittering the direction of travel uniformly between -5 and 5 degrees. Finally, each image is cropped vertically and reshaped to 224×960. The cropping operation removes the unneeded sky and ground portion of the panorama, and we preserve the aspect ratio of the cropped image when reshaping.

13

Table 2.1: Parameter settings for each method.

Method	λ_{s^1}	λ_{s^2}	λ_s	λ_m	λ_u
Baseline	1	10^2	1	0	0
M1	1	10^2	1	0	0
M2	1	10^2	0.1	1	0
M3	1	10^2	1	0	0.001
M4	1	10^2	0.1	1	0
Ours	1	10^2	0.1	1	0.001

Train/Test Split   To create train/test splits for network training and evaluation, we utilize a stratified sampling method for each region's data, 90% for train and 10% for test. A validation set (2% of train) is used for model selection. Corresponding splits from the two regions are combined to form the final sets. We have ensured that locations in the test set are at least 300 meters away from locations in the train set.

Class Weighting   The distribution of labels in our training split is very unbalanced, which led to poor fitting to panoramas with labels in the minority groups (1 star roads specifically). To alleviate this issue, we deploy a class weighting strategy on the star rating loss function to proportionally adjust the loss of each sample based on the frequency of its ground truth label.

We first find the weight vector of each star rating class w through the equation below:

$$w_l = \frac{N}{K * count(l)} \tag{2.5}$$

where w_l is the weight for class l. With the weight vector w, we modify equation 2.1 and 2.2 as follows,

$$L_{s^1} = -\frac{1}{N} \sum_{i=1}^{N} w_{l_i} y_i(l_i) \log \hat{y}_i(l_i) \tag{2.6}$$

$$L_{s^2} = \frac{1}{N} \sum_{i=1}^{N} w_{l_i} \|F(\hat{y}_i) - F(y_i)\|_2^2 \tag{2.7}$$

where w_{l_i} is the weight for class l_i.

Table 2.2: Top-1 accuracy for each method.

Method	Attn.	Multi.	Unsuper.	Acc.
Baseline				43.06
M1	X			43.28
M2		X		43.56
M3	X		X	43.63
M4	X	X		45.68
Ours	X	X	X	**46.91**

2.4.3   Ablation Study

We compare our complete architecture with five variants. The simplest method, *Baseline*, omits the adaptive attention layer and instead equally weights all parts of the input feature map (i.e., global average pooling). The remaining variants include different combinations of adaptive attention, multi-task loss, and unsupervised loss. For each method, we performed a greedy hyperparameters search using our validation set. We initially selected the optimal weighting of λ_{s^2} relative to λ_{s^1} by selecting λ_{s^2} from $(.01, .1, 1, 10, 100)$. We used the same strategy when adding a new feature, while keeping the existing hyperparameters fixed. We train all networks for 10,000 iterations.

Table 2.2 displays the macro-averaged accuracy for all methods on the test set, where for each method we compute the per-class accuracy and average the class accuracies. Our method outperforms all other methods, each of which includes a subset of the complete approach. We observe that the *Baseline*, *M1*, *M2*, and *M3* methods all achieve similar accuracy. The next method, *M4*, which combines the per-task attention and multi-task learning, performs significantly better. It seems that the multi-task loss and attention in isolation are not particularly helpful, but when they are combined they lead to improved performance. We also observe that the unsupervised loss is only significantly helpful when combined with per-task attention and the multi-task loss.

Figure 2.4 shows the test-set confusion matrix for our best method. The results are quite satisfactory except for the 1-star rating category, but that is as expected due to the
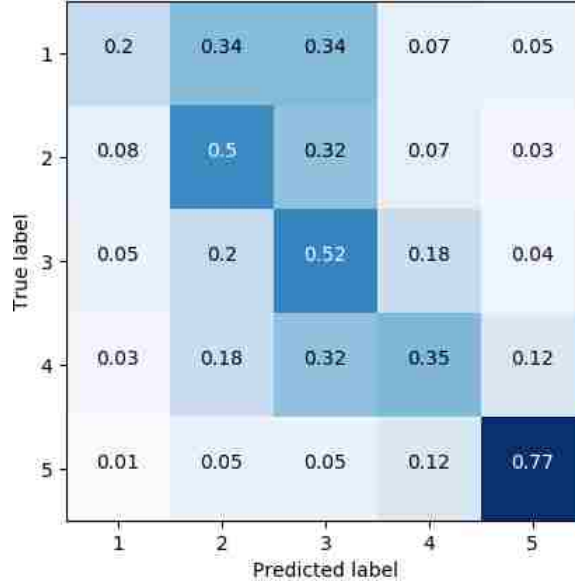
Figure 2.4: The test-set confusion matrix (row normalized) for our method.

imbalanced nature of our training dataset, with only 5.6% of all samples being 1-star roads.

2.4.4 Visualizing Attention

As described in Section 2.3.2, each task has an independent attention mask, a, whose shape (1×210) corresponds to the flattened spatial dimensions of the feature map, S_2, output by the previous convolutional layer. Therefore, a reshaped version of a corresponds to the regions in S_2 that are important when making predictions for a given task. Figure 2.5 visualizes the attention mechanism for our main task and several of our auxiliary tasks, where lighter (yellow) regions have higher weight than darker (red) regions. For example, in Figure 2.5 (e), attention is focused on the left side of the panorama, which makes sense given the task is related to identifying dangerous driver-side objects, such as a drainage ditch or safety barrier.

2.4.5 Multi-task Evaluation

We evaluate the multi-task prediction performance of our architecture. Table 2.3 shows the Top-1 macro-averaged accuracy for each of our auxiliary tasks, along with random

Example Panorama

(a) Star Rating
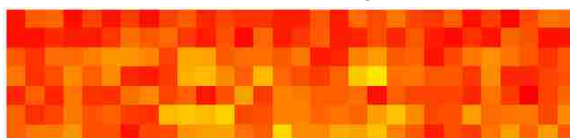
(b) Median Type

(c) Land Use (driver-side)
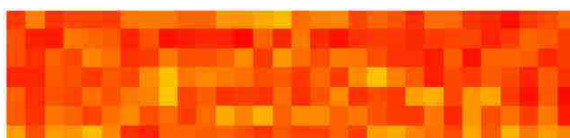

(d) Number of Lanes

(e) Roadside Severity (driver-side object)

Figure 2.5: Heatmaps that visualize the weights of the task-specific attention layers. White (red) corresponds to higher (lower) weight.

Table 2.3: Multi-task evaluation for our architecture. R. = roadside, P. = passenger, and D. = driver.

Label type	Top-1	Random	% inc.
Area type	97.40	50.39	47.01
Lane width	77.22	33.25	43.97
Curve quality	63.57	32.78	30.79
P. side land use	54.88	28.38	26.50
D. side land use	52.72	28.32	24.40
D. side sidewalk	77.17	57.00	20.17
Vehicle parking	52.39	33.06	19.33
Road condition	51.38	33.35	18.03
P. side sidewalk	60.34	43.54	16.80
Intersection quality	83.06	66.69	16.37
Intersection road volume	28.21	14.18	14.03
D. side paved shoulder	38.61	24.71	13.90
P. side paved shoulder	37.56	24.42	13.14
Number of lanes	45.78	33.13	12.65
R. D. side distance	37.51	25.47	12.04
R. P. side distance	36.45	24.98	11.47
Median type	57.30	46.86	10.44
R. P. side object	38.89	29.59	9.30
Upgrade cost	42.82	33.93	8.89
R. D. side object	49.24	41.15	8.09
Intersect channel	56.25	49.88	6.37
Bicycle facilities	75.19	71.51	3.68
Curvature	25.83	25.52	0.31

performance if we sample predictions based on the prior distribution of labels for that task. The rightmost column shows the relative increase in accuracy for each task over random performance, in some cases with a performance gain of almost 50%.

2.4.6 Safety-Aware Vehicle Routing

While our focus is on road safety assessment for infrastructure planning, our network could be used for other tasks. Here we show how it could be used to suggest less-risky driving routes. We use a GPS routing engine that identifies the optimal path, usually the shortest and simplest, a vehicle should traverse to reach a target destination. Some work has been done to explore semantic routing with respect to scenery [27], carpooling [14], and per-
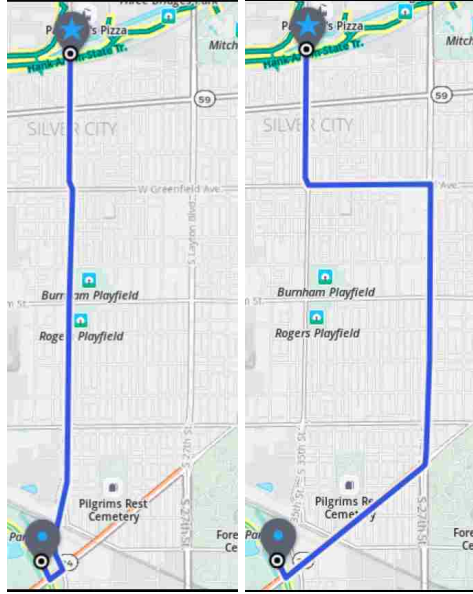
Figure 2.6: Unaltered routing (left) and predicted safety score route (right).

sonalization [32]. We propose routing that employs road safety scores for navigating to a destination, in order to balance speed and safety.

We selected a subset of the panorama test split and used it to influence the Mapzen Valhalla routing engines edge cost calculation. From the subset, each safety score was used as an edge factor corresponding to the safety scores GPS coordinate. When the routing engine searched the GPS graph for a routes traversal cost, it would identify if a cost factor corresponding to a specific edge existed in the subset. Should a cost factor be present, the edge cost of the traversal would be $c_{edge} = c_o * factor$. The routing engine utilizes the augmented edge costs to determine the optimal route, namely, the lowest cumulative cost route. The two routes depicted in Figure 2.6 demonstrate the impact of safety-aware routing in a major US urban area. Figure 2.6-right shows a less risky, but longer, route chosen by the enhanced routing engine, while Figure 2.6-left shows the default route. The enhanced routing engine chooses this route to circumvent a low (1 or 2) safety scoring road and instead travel on a high (4 or 5) scoring road. Figure 2.7-top shows a panorama from the higher risk road. It has numerous issues, including poor pavement condition and small lane widths. Figure 2.7-bottom shows a panorama from the less risky road. This road has
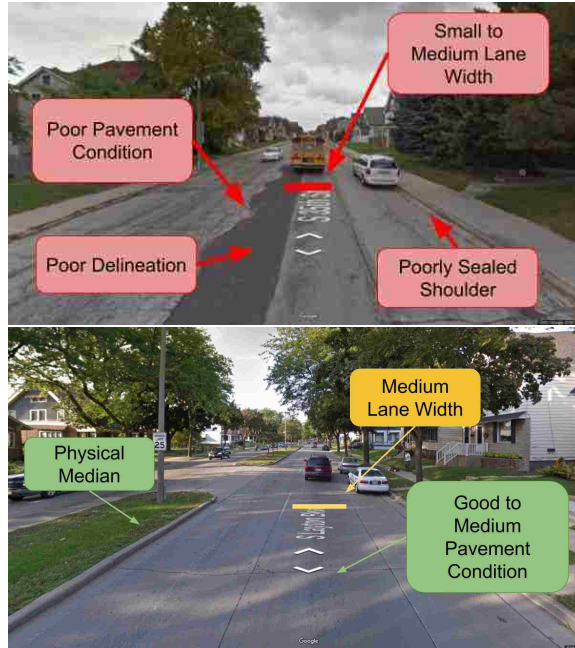
Figure 2.7: Annotation of one star (top) and four star (bottom) roads.

wider lanes, better pavement, and a physical median. While this route may take longer, it clearly traverses a less risky path.

While this is a minor modification to an existing routing engine, we think the ability to optimize for safety over speed could lead to significant reduction in injuries and deaths for vehicle users.

2.5 Conclusion

We introduced an automated approach to estimate roadway safety that is significantly faster than previous methods requiring significant manual effort. We demonstrate how a combination of a spatial attention mechanism, transfer learning, multi-task learning, and unsupervised learning results in the best performance on a large dataset of real-world images. This approach has the potential to dramatically affect the deployment of limited transportation network improvement funds toward locations that will have maximum impact on safety. The outlined approach addresses the main concern of many agencies in deploying sys-

temic analysis (such as usRAP)—cost to collect and process data. As the availability of street-level panoramas is growing rapidly, employing automation techniques could allow many more agencies to take advantage of systemic road safety techniques. For many agencies, there are few options available to prioritize roadway safety investments. Even when and where sufficient data are available, smaller agencies typically lack the expertise to conduct robust safety analysis as described in the HSM [16] for high crash location assessment or in usRAP for systemic study. Our approach could reduce the cost of systemic analysis to these agencies, or help larger agencies assess more of their roads, more frequently. For future contributions, we plan to explore at least three offshoots of this work: 1) apply the proposed method to additional highway safety tasks, 2) integrate overhead imagery, and 3) make assessments using multiple panoramas.

Acknowledgments

Chapter 3

Remote Estimation of Free-flow Speeds

3.1   Introduction

Aside from understanding road conditions, analysis of driver behavior is also crucial for roadway safety assessment. The behavior of an average automobile driver is based on numerous factors. In this chapter, we focus on one aspect of behavior: the speed of travel. We specifically focus on the free-flow speed, which is the average vehicle speed along a roadway when there is no congestion or adverse weather conditions. The free-flow speed is used in a wide variety of planning and regulatory contexts. Standard practice for estimating free-flow speeds relies on the availability of many road features, such as lane widths and curvature. While some of these features are easy to gather, others require expensive surveying equipment and expert annotators, restricting their availability. As a result, in many states, roads are assigned the default speed limit of 35 mph for urban areas and 55 mph for rural areas, which can often be inconsistent with actual operational speeds and lead to undesired driver behavior.

To automate the costly process of estimating free-flow speeds, we propose a deep convolutional neural network (CNN) that takes as input the overhead imagery of the roadway and coarse-grained road features, including the posted speed limit, the functional classi-

fication, and the type of area. As output, our CNN generates a probability distribution over integer free-flow speeds. For training data, we use aggregated data collected from real-world drivers. At inference time, since the input features are all easily obtainable, it should be possible to rapidly, and inexpensively, estimate free-flow speeds over large spatial regions.

Using a large-scale evaluation dataset, we find that using only image features result in the poorest performance, but that combining imagery and road features result in significantly better performance than road features alone.

3.2 Related Work

Different studies have been proposed to estimate and map properties of the visual world using overhead images. Several authors have proposed different deep learning based approaches for vehicle detection [29, 34] and road extraction [43, 6, 4, 22] from aerial images. Salem et al. [30] introduced an approach for mapping soundscapes of geographic regions using overhead imagery. Greenwell et al. [11] proposed a model that is capable of predicting object histograms from overhead imagery. Several works have addressed the problem of speed estimation. Shuai et al. [15] introduced an approach to estimate the vehicle speed in traffic videos. Most similar to our work, Song et al. [35] proposed a model for road safety estimation based on the usRAP Star Rating Protocol. While this star rating is based on approximately 60 road safety features [12], their network works directly on ground level panorama images. We propose a new method that instead uses overhead imagery with the addition of auxiliary features.

3.3 Approach

We utilize a CNN architecture to estimate the free-flow speed of a given road segment. The neural network uses both aerial imagery and relevant road features as input, and outputs
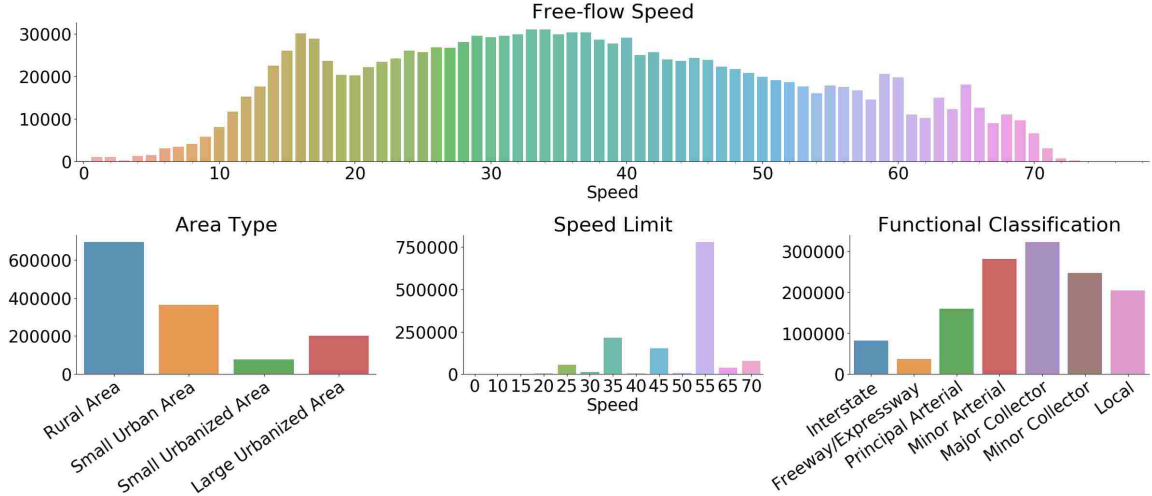
Figure 3.1: Histograms of various road features in our dataset.

a probability mass function over K possible free-flow speeds. We begin by describing the dataset that we use, followed by a more detailed description of the proposed network architecture.

3.3.1 Dataset

Our free-flow speed dataset is obtained through HERE Technologies and further annotated to incorporate coarse-grained road feature data needed for training. To calculate the free-flow speed for a particular road segment, driving speeds of vehicles were monitored during the year 2014. As free-flow speed refers to the speed that a driver can achieve without traffic congestion, we only consider data during non-holiday weekday periods from 9am to 3pm. The speeds are then averaged to obtain the ground-truth for each road segment. We rounded each ground-truth speed to the nearest integer to obtain a discrete label for training. This results in $K = 79$ unique free-flow speeds

For each road segment, we obtain an aerial image through the National Agriculture Imagery Program (NAIP) data. Each image is centered at the beginning of the segment, and covers an aerial view of 400×400 m^2 of land. We orient the images such that the direction of travel is up, and we reshape it to 224×224. In total, our dataset consists of
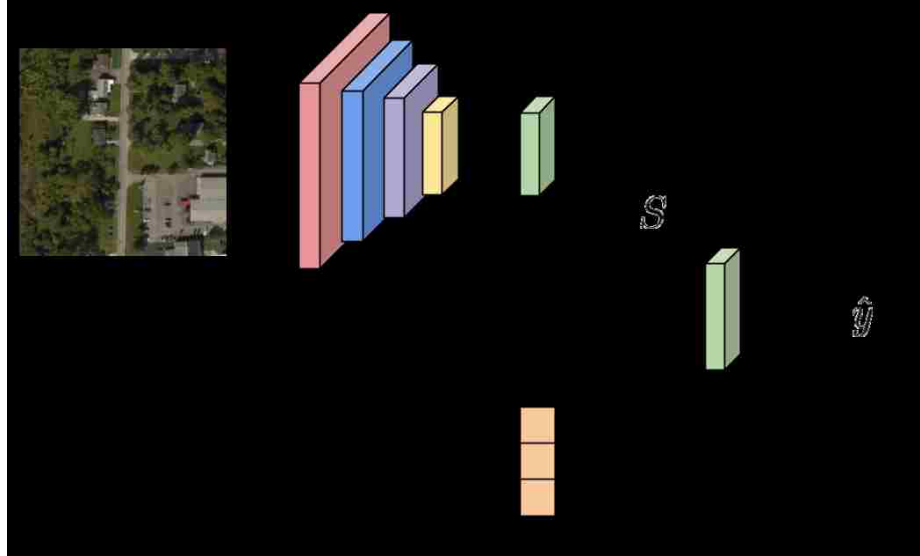
24

Figure 3.2: Network architecture for predicting free-flow speeds.

1 335 135 pairs of road segments and their corresponding aerial images.

Figure 3.1 shows the distribution of the free-flow speeds along with road features used for training. The drastic difference between the distribution of free-flow speeds and speed limits is evident here. In fact, over half of the roads have free-flow speeds significantly slower than the speed limits (more than 10-miles-per-hour). We will examine a few roads with such differences in Section 3.4.2, using predictions obtained from our model.

We split our dataset into training and testing sets. To ensure the sets are disjoint, we partition the dataset by location. The dataset is divided into counties, eight of which are selected for evaluation. The resulting test set is roughly 7% of the total data, and contains a variety of roads with different area types, terrain, and free-flow speeds. A validation set (1% of the training set) is reserved for model selection.

3.3.2 Network Architecture

The backbone of our architecture is the Xception [7] model, which is used to extract high-level features from the aerial image. The resulting feature vector of length 2048 is passed to a dense layer with output size 512, denoted as S, as shown in Figure 3.2. In addition

to aerial imagery, the network is trained using three discrete features related to driving speed: area type, functional classification, and posted speed limit. The three features are concatenated with S before feeding the combined feature vector into the final dense layer with output size K. We denote the output of the final layer as $\hat{y}$.

3.3.3 Loss Function

We formulate the task of predicting the free-flow speed of a road segment as a multiclass classification problem with ground-truth free-flow speed label l. With the output of the network defined in Section 3.3.2, we compute the training loss as the standard cross-entropy loss between the predicted distribution $\hat{y}$ and the target distribution y, as shown in the following equation:

$$L = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} y_{i,k} \log \hat{y}_{i,k} \tag{3.1}$$

where N is the number of training examples.

3.3.4 Implementation Details

Our model is implemented using the Tensorflow framework. The network is optimized using the Adam optimizer. We initialize the Xception network with weights pre-trained on ImageNet for image classification. During training, we freeze the weights of the Xception network and optimize only the last two dense layers with learning rate 0.001. We decay the learning rate by a factor of 10 every 5 epochs. *ReLU* activation layers are used throughout the network except for the last layer, which uses the Softmax activation instead. We apply L_2 regularization to the two dense layers with scale 0.00005. We train with batch size 16, for a total of 15 epochs.

Table 3.1: Within-5 accuracy for each method.

Method	Within-5 Acc.
Imagery Only	37.60
Road Features Only	40.07
Combined	**49.86**

3.4 Evaluation

Using the dataset described in Section 3.3.1, we trained our proposed model along with two other variations: one using only imagery as input and the other using only road features. We conducted quantitative and qualitative evaluation on our reserved test set.

3.4.1 Quantitative Analysis

We aim to discover the effect of each input modality in predicting free-flow speeds. We calculate the accuracy of the three trained models (overhead imagery only, road features only, and both) on the test set. Table 3.1 displays the within-5 accuracy for each method, where we consider a prediction to be positive if it is within five-miles-per-hour of the ground-truth speed. As we can see, road feature data is better than aerial imagery for predicting free-flow speeds, but we obtain the best performance when we combine both modalities.

3.4.2 Qualitative Evaluation

Free-flow speed should be similar to the speed limit for a given road, but that is often not the case. We are interested in finding roads where the predicted free-flow speed is drastically different than the speed limit. Figure 3.3 shows such examples. The top two roads have free-flow speeds much lower than the speed limits, most likely due to the curvature and number of intersections present. The bottom two roads have free-flow speeds much higher than the speed limits, which is understandable since the roads are straight, wide, and without congestion. Detection of these discrepancies is very useful, as traffic engi-

(a) Free-flow: 18, Limit: 55   (b) Free-flow: 17, Limit: 55

(c) Free-flow: 27, Limit: 15   (d) Free-flow: 67, Limit: 55

Figure 3.3: Examples of discrepancies in the predicted free-flow speed and the speed limit of roads in miles per hour. Predictions obtained from the combined feature model.

neers can quickly filter through millions of roads to identify ones that need to have their speed limits re-evaluated.

We also compared predictions from our top two models: one trained on both aerial imagery and road features, and the other trained only on road features. Figure 3.4 shows the images that had drastically different predictions from the two models. The images in the right column received higher free-flow speed predictions from the combined model, where the images in the left column received higher free-flow speed predictions from the feature only model. These differences further support the importance of image features for prediction, as imagery can provide hidden, but powerful, information about the road segment. For the left three roads, it is natural for people to drive slower in residential areas or at intersections; for the right three, straight roads and ones parallel to interstates allow drivers to achieve greater speeds. Fine-grained labels such as presence of intersections,

method A: 50, method B: 16   method A: 31, method B: 42

method A: 46, method B: 15   method A: 24, method B: 50
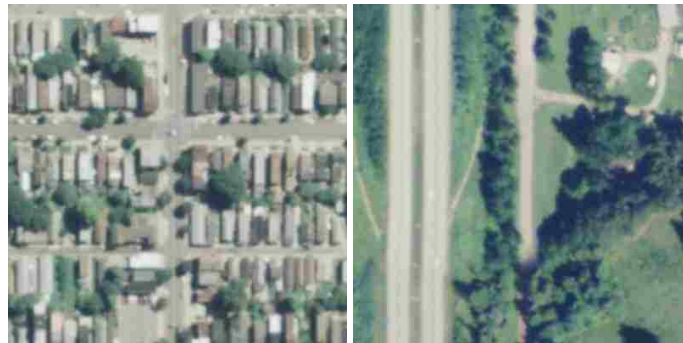
method A: 30, method B: 15   method A: 24, method B: 40

Figure 3.4: Free-flow speed predictions in miles per hour from two methods, where method A is the road feature only model and method B is the combined feature model.

population density, and curvature of the road are hard to obtain, but they can easily be inferred from aerial imagery.

3.5 Conclusion

We introduced a method for estimating the free-flow speed of a road segment, which is important for understanding driver behavior. We demonstrated that a combination of aerial imagery and related road features as input is best for prediction, since it obtained higher accuracy than models trained on either features alone. We also qualitatively analyzed our predictions, and obtained insights on the effect of input modalities along with the relationship between free-flow speeds and speed limits. We hope to extend this work and include ground-level imagery as an additional input to our method, since information such as roadside hazards may not be visible in aerial imagery.

Chapter 4

Conclusion and Future Work

In this thesis, we have looked at two roadway safety metrics: the Star Rating score and the free-flow speed. To predict Star Ratings, we utilized ground-level imagery through Google Streetview, which is available almost everywhere in the US. For the task of estimating free-flow speeds, aerial-level imagery is used instead, which is available at a global scale. Both methods use convolutional neural networks, enabling powerful performance at a fraction of the cost of manual labeling. While these methods are not at the same level as human performance yet, they demonstrated good results on a limited dataset, and should improve with more available data. Each method can greatly assist with assessment of roads on a country-wide scale. Given enough computing resources, one can infer Star Rating scores and free-flow speeds of every road in the US within a single day. Such tools allow rapid evaluation of roads, help keep road information up-to-date, and provide safer roads for drivers.

A natural future direction for automatic roadway safety assessment is to combine different data modalities for prediction. We have seen an example where overhead imagery features are fused with road metadata for prediction of free-flow speeds, and the results are better than either data modality alone. It is of interest to combine different image modalities, as curvature can easily be inferred from an aerial-level image while roadside hazards

are easily distinguished from ground-level imagery. At the same time, the method needs to account for the difference in data availability, as aerial imagery is in greater abundance than ground imagery.

Bibliography

[1] United States Road Assessment Program. `http://www.usrap.org`. Accessed: 2018-1-13. 12

[2] Traffic safety facts. Technical report, National Highway Traffic Safety Administration, 2016. 1

[3] Global status report on road safety 2018. Technical report, Geneva: World Health Organization, 2018. Licence: CC BYNC-SA 3.0 IGO. 1

[4] Rasha Alshehhi and Prashanth Reddy Marpu. Hierarchical graph-based segmentation for extracting road networks from high-resolution satellite images. *ISPRS journal of photogrammetry and remote sensing*, 126:245–260, 2017. 23

[5] Sean M Arietta, Alexei A Efros, Ravi Ramamoorthi, and Maneesh Agrawala. City forensics: Using visual elements to predict non-visual city attributes. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2624–2633, 2014. 7

[6] Favyen Bastani, Songtao He, Sofiane Abbar, Mohammad Alizadeh, Hari Balakrishnan, Sanjay Chawla, Sam Madden, and David DeWitt. Roadtracer: Automatic extraction of road networks from aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4720–4728, 2018. 23

[7] François Chollet. Xception: Deep learning with depthwise separable convolutions. *CoRR*, abs/1610.02357, 2016. 25

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 6

[9] Abhimanyu Dubey, Nikhil Naik, Devi Parikh, Ramesh Raskar, and César A Hidalgo. Deep learning the city: Quantifying urban perception at a global scale. In *European Conference on Computer Vision*, 2016. 7

[10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 7

[11] Connor Greenwell, Scott Workman, and Nathan Jacobs. What goes where: Predicting object distributions from above. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2018. 23

[12] Douglas Harwood, Karin Bauer, David Gilmore, Reginald Souleyrette, and Zachary Hans. Validation of us road assessment program star rating protocol: Application to safety management of us roads. *Transportation Research Record: Journal of the Transportation Research Board*, 2147:33–41, 2010. 7, 12, 23

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE International Conference on Computer Vision*, 2015. 6

[14] Wen He, Kai Hwang, and Deyi Li. Intelligent carpool routing for urban ridesharing by mining gps trajectories. *IEEE Transactions on Intelligent Transportation Systems*, 15:2286–2296, 2014. 18

[15] Shuai Hua, Manika Kapoor, and David C Anastasiu. Vehicle tracking and speed estimation from traffic videos. In *CVPR Workshop (CVPRW) on the AI City Challenge*, 2018. 23

[16] Joint Task Force on the Highway Safety Manual. *Highway Safety Manual*, volume 1. American Association of State Highway and Transportation Officials, 2010. 7, 21

[17] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2014. 11

[18] Alex Krizhevsky, Ilya Sutskever, and GeoffreyE Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012. 6

[19] Stefan Lee, Haipeng Zhang, and David J Crandall. Predicting geo-informative attributes in large-scale image collections using convolutional neural networks. In *IEEE Winter Conference on Applications of Computer Vision*, 2015. 6

[20] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. In *International Conference on Learning Representations*, 2014. 9

[21] Tsung-Yi Lin, Yin Cui, Serge Belongie, and James Hays. Learning deep representations for ground-to-aerial geolocalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 6

[22] Gellért Máttyus, Wenjie Luo, and Raquel Urtasun. Deeproadmapper: Extracting road topology from aerial images. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3458–3466. IEEE, 2017. 23

[23] Gellért Máttyus, Shenlong Wang, Sanja Fidler, and Raquel Urtasun. Hd maps: Fine-grained road segmentation by parsing ground and aerial images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 7

[24] Nikhil Naik, Jade Philipoom, Ramesh Raskar, and César Hidalgo. Streetscore-predicting the perceived safety of one million streetscapes. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014. 7

[25] B. Nye, E. Fitzsimmons, and S. Dissanayake. Demonstration of the United States Road Assessment (usRAP) as a Systematic Safety Tool for Two Lane Roadways and Highways in Kansas. *Journal of the Transportation Research Forum*, 56(1), 2017. 12

[26] Vicente Ordonez and Tamara L Berg. Learning high-level judgments of urban perception. In *European Conference on Computer Vision*, 2014. 7

[27] Daniele Quercia, Rossano Schifanella, and Luca Maria Aiello. The shortest path to happiness: Recommending beautiful, quiet, and happy routes in the city. *CoRR*, abs/1407.1031, 2014. 18

[28] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 11

[29] Wesam Sakla, Goran Konjevod, and T Nathan Mundhenk. Deep multi-modal vehicle detection in aerial isr imagery. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pages 916–923. IEEE, 2017. 23

[30] Tawfiq Salem, Menghua Zhai, Scott Workman, and Nathan Jacobs. A multimodal approach to mapping soundscapes. In *IGARSS*, 2018. 23

[31] Philip Salesses, Katja Schechtner, and César A Hidalgo. The collaborative image of the city: mapping the inequality of urban perception. *PloS one*, 8(7):e68400, 2013. 7

[32] Wenjie Sha, Daehan Kwak, Badri Nath, and Liviu Iftode. Social vehicle navigation: Integrating shared driving experience into vehicle navigation. In *Proceedings of the 14th Workshop on Mobile Computing Systems and Applications*, 2013. 19

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 8

[34] Lars Wilko Sommer, Tobias Schuchert, and Jürgen Beyerer. Fast deep vehicle detection in aerial images. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pages 311–319. IEEE, 2017. 23

[35] Weilian Song, Scott Workman, Armin Hadzic, Reginald Souleyrette, Eric Green, Mei Chen, Xu Zhang, and Nathan Jacobs. Farsa: Fully automated roadway safety assessment. In *WACV*, 2018. 23

[36] Shenlong Wang, Min Bai, Gellert Mattyus, Hang Chu, Wenjie Luo, Bin Yang, Justin Liang, Joel Cheverie, Sanja Fidler, and Raquel Urtasun. Torontocity: Seeing the world with a million eyes. In *IEEE International Conference on Computer Vision*, 2017. 7

[37] Scott Workman, Connor Greenwell, Menghua Zhai, Ryan Baltenberger, and Nathan Jacobs. DeepFocal: A Method for Direct Focal Length Estimation. In *International Conference on Image Processing*, 2015. 6

[38] Scott Workman, Richard Souvenir, and Nathan Jacobs. Understanding and Mapping Natural Beauty. In *IEEE International Conference on Computer Vision*, 2017. 7

[39] Scott Workman, Menghua Zhai, David J. Crandall, and Nathan Jacobs. A unified model for near and remote sensing. In *IEEE International Conference on Computer Vision*, 2017. 7

[40] Kun-Feng Wu, Scott Himes, and Martin Pietrucha. Evaluation of effectiveness of the federal highway safety improvement program. *Transportation Research Record: Journal of the Transportation Research Board*, (2318):23–34, 2012. 2

[41] Jiaquan Xu, Kenneth D Kochanek, Sherry L Murphy, and Betzaida Tejada-Vera. Deaths: final data for 2014. *National Vital Statistics Reports*, 65(5):1–122, 2016. 1

[42] Menghua Zhai, Scott Workman, and Nathan Jacobs. Detecting Vanishing Points using Global Image Context in a Non-Manhattan World. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 6

[43] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 2018. 23

[44] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99), 2017. 11

[45] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems*, 2014. 6

Vita

Weilian Song

Education

2015–2018   B.S. in Computer Science   University of Kentucky

        Summa Cum Laude, Minor in Mathematics

Publications

Refereed Conference Publications

[1] Weilian Song, Tawfiq Salem, Hunter Blanton, and Nathan Jacobs. Remote Estimation of Free-flow Speeds. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2019.

[2] Weilian Song, Scott Workman, Armin Hadzic, Xu Zhang, Eric Green, Mei Chen, Reginald Souleyrette, and Nathan Jacobs. FARSA: Fully Automated Roadway Safety Assessment. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018.

Abstracts

[1] Weilian Song, Tawfiq Salem, Nathan Jacobs, and Michael Johnson. Detecting the presence of bird vocalizations in audio segments using a convolutional neural network architecture. In *International Symposium on Acoustic Communication by Animals*, 2017.

Appointments

Research Assistant Computer Vision Lab, University of Kentucky

2016–present *Lexington, KY*

Academic Tutor Tau Beta Pi Tutoring Service

2017 *Lexington, KY*

Honors and Awards

- 1st place in 2018 University of Kentucky Student Programming Competition

- Provost Scholarship ($1500/year), based on high school GPA and ACT scores

- Engineering Academic Scholarship ($12,000), based on high ACT scores

- Tau Beta Pi

- Phi Kappa Phi

- Chellgren Student Fellow