CSE6242 - Data and Visual Analytics

Final Report

Team 64: Safe Street

Section 1: Introduction and Problem Definition

The trending bike movement in the United States is advocated by a large population who is seeking a healthy and green lifestyle, especially in big cities like San Francisco. With its plan for net zero emissions in 2050, San Francisco is the second most bike friendly city in the United States [2, 20]. Meanwhile, as biking becomes a more prevalent mode of transportation, the rate of bicyclist injury has increased drastically by 21% from 2013 to 2017[20]. Cyclists are still extremely vulnerable when comparing to other types of commuters despite existing political policies and initiatives. To boost safety for cyclists and improve the civil infrastructure, urban designers need to know what road features are most relevant to bike accident rate. Such civil projects, however, are expensive to carry out in the Bay Area. To make matters worse, there is a constant shortage of data tools that can support new ideas of transportation professionals. Our solution, Safe Street for Cyclist, aims to fill this gap by identifying key contributors to bike accidents. Our goal is to help stakeholders prioritize their resources and guide them with budget decisions when it comes to city planning. Additionally, our deliverable would also serve as an educational pathway and provides rich insight into San Francisco street safety for rest of the world.

Section 2: Literature Survey

Today, many risk factors are assessed through high-level social studies such as vehicle counts, demographics, and social-economic status [14], but those studies rarely expose the root causes of accidents. One common problem of these approaches is the lack of associating transferable visual street features with accident data.

Other researches collect historical accident data and visualize the number of accidents on a map view. As a result, streets with higher accident frequency would be considered more dangerous [5, 15, 19]. The bias is that areas with higher traffic density are sampled and reported more often, and are thus more likely to be categorized as dangerous. Additionally, there is little visibility into the root cause of an accident.

Inspired by the work in [1, 5, 7], we aim to develop a scalable approach that leverages object detection algorithm [9, 10] to identify visual features of a street and use these features to predict risk score. Our model will be trained with Microsoft's common objects dataset [11], which current practices [3, 7, 16, 18] do not have.

Moreover, we will perform clustering analysis on the detected objects and features extracted from street views. Similar to [9, 13], the significance level and effects of these clusters will be examined, and the significant ones will be the inputs for our final risk model. Our new approach has no sampling bias but covers all bike accident-related data, aiming to provide a unified method of collecting, storing, and transferring the features related to bike accidents. It is a one-stop shop for researchers and users.

Our approach is effective because Google Street Views API is a reliable and cost-effective data source for extracting infrastructure features and it has been used successfully in similar researches [4, 6]. San Francisco's route infrastructure differs from those in cities with high trip shares (like Amsterdam), so analyzing street features is a promising approach for investigating injury risks [18]. Our work will be presented in interactive visualization tools like mapbox.gl, rendering it user-friendly for users with non-technical background.

Section 3: Proposed methodology

The innovations are summarized here, and details can be found in a later section.

a. Computation:
    i. Use object detection model to pull location characteristics from the images
    ii. Combine object detection findings with public road feature data in developing a Random Forest model for each location
    iii. Combine object detection findings with public road feature data in developing a Convolutional neural network model
    iv. Compare the results of both models to select the winning prediction model

b. Data presentation:
    i. Visualize risk score predictions and historical bike collision data on one single interactive map
    ii. Craft user-friendly web page with parallax effect in story-telling; readers are able to interact with page-level filters to display desired data segments

3.1 Data Collection & Data Cleaning

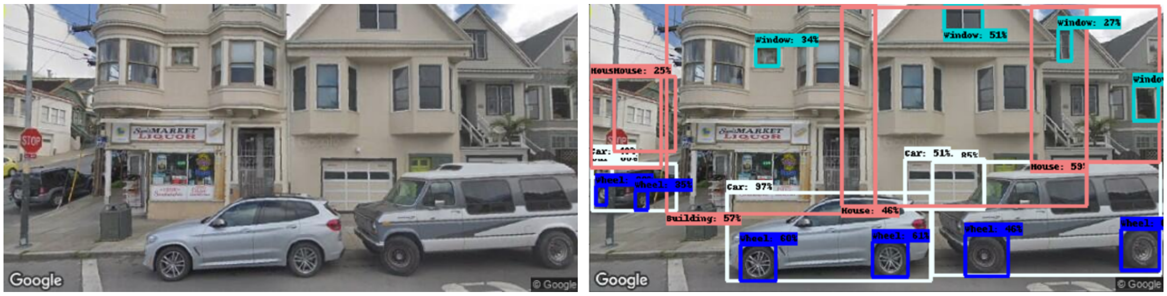
The data collection process begins with literature research and compiling a list of relevant datasets. The original datasets are then transformed to suit our needs. We gathered historical bike collision data in San Francisco between 2008 and 2012 from two pubic data resources: Transportation Injury Mapping System (TIMS) and TransBase. We aggregated these datasets from case-level to street segment level and enhanced it by adding street feature data.

Since both datasets are publicly available, they contain some redundant information and have large amounts of missing data. The cleaning process is done with Python Pandas, NumPy and SQLite3 that includes multiple joins and aggregations from the schema. After a thorough review of the original datasets, we removed columns that are not relevant to road features. Some street features, such as speed limit and parking spot counts, contain missing values (between 1% to 3%). We used KNN at street segment level to impute the missing value so they can be fitted into the model. The idea of using KNN is very intuitive – streets that are closer to each other are more likely to have similar speed limit and parking spot counts. For categorical variables, we replace the value with “missing” to label the null values if the share of missing values is above 30%.

To get the image data for object detection later, we extract a list of latitude and longitude pairs for each intersection in San Francisco and a list of coordinates for each collision. Using the Google Street Views API, we have successfully collected 9,147 images from latitude-longitude coordinates for intersections and 3,852 images for collision locations.

3.2 Computation

To enrich street features, we leveraged a TensorFlow object detection model trained on Google's Open Image V4 dataset to identify possibly relevant objects and street characteristics on all the Google Street Views images collected. For each image, we detected 100 objects, each with its probability score. We ended up with a very sparse frequency matrix, with 169 unique objects detected. It is a sparse matrix because many objects are only detected once. To handle this sparsity, we applied PCA on this matrix to reduce the dimension from 169 to 20. We have selected the top 20 principal components, which explains 99.5% of the variance of the original matrix. From there we were able to cluster streets using K-means into custom groups on those streets. The values of K can be chosen based on the cumulative sum of distance away from centroids of all points, which implies that K between 7 and 10 seem optimal. We will determine the final value of K from model tuning that optimizes the accuracy. The results from K-Mean can now combined into other existing street features in the TransBase dataset and be used in the final prediction model.



Latitude, Longitude: 37.7483528, -122.40871

Once the model dataset is ready, we developed two models to explore which set of street feature(s) are most relevant to a street's risk score.

The first model, using Random Forest Regressor, leveraged both street features and K-Means clusters to predict the number of accidents divided by traffic volume. The processed dataset includes some categorical variables, such as sharrow type, facility type, and lane type. Since Random Forest cannot deal with categorical variables directly, those columns are converted into binary columns using one-hot-encoding. To find the optimal parameters for *K* in K-Means, and *Max Depth, Max Features*, and *Number of Trees* in Random Forest, we used a cross validated grid search in Python to tune the model. We have tried out 300 (4 x 5 x 5 x3) hyperparameters combinations and it turns out that the best set is 7 for *K*, 5 for *Max Depth*, 10 for *Max Features*, and 300 for *Number of Trees*. This best fitted model has a 65% out-of-bag R-squared, which is not very high but also reasonable. Interestingly, we find that the most significant features are estimated number of daily ride volume, K-Means clusters, and speed limit. Finally, since the predicted scores (number of accidents divided by traffic volume) for all segments are in decimal point and very skewed, we applied a log transformation and min max scaler to make the predicted scores to be more interpretable to users. The original score ranges from 0.0078 to 0.0135 with high skewness, but the transformed score ranges from 50 to 100 with much lower skewness.

The second model, based on CNN algorithm, used only street view images. The algorithm was selected due to its capability to capture and analyze visual feature representation of Google Street

View images. We applied several layers of filters to transform the image data from pixel representations to our desired feature dataset, where the objects detected are grouped into clusters with labels.

We enter the map of Google street view images as inputs data and turn them into several matrices to the Convolution Layer. The activation function of the convolutional layer is ReLU. The activation function (ReLU) is ReLU (x) = max (0, x). Behind the convolution layer is a pooling layer. It doesn't have any activation function in the pooling layers.

The combination of convolutional layer + pooling layer can appear many times in the hidden layer. The number of the combination is based on the needs of our model. We have tried several numbers of combination to get more accurate results. We also can flexibly use a combination of convolutional layer + convolutional layer, or convolutional layer + convolutional layer + pooling layer. After several convolutional layers + pooling layers, the next step is a Fully Connected Layer (FC). Fully connected layers' function is as "classifiers" in the entire convolutional neural network. The fully connected layer plays the role of mapping the learned "distributed feature representation" to the sample tag space.

We got the cumulative distribution function (c.d.f) for the distribution of labels (probability of accidents), which describes the distribution of the label in the data set. The distribution is very uneven. Unbalanced samples will make the model underfitting. For example, 90% of the training data are positive and the others are negative, that will cause the model to make all the predictions are positive, so the accuracy can reach 90%. But the model may not work well in the testing data set. We cannot find a good way to solve the problem of uneven samples. In our data set, no matter how the model and hyperparameters were tuned, the model error would always converge to a point, and the predictions were not very convincing.

The possible reason for this result is that our training data set has relatively few samples. In addition, the samples in the training data set are all pictures of the accident-prone areas, and there is a lack of sample images of the general areas (because it is difficult to obtain the sample labels of them). It is very hard to obtain accurate prediction results for a model trained under a small and uneven training data set. But this is also a good attempt, because we can learn how the structure of the Google street view images data set should be to train similar models in the future.

Based on validation results from the two approaches mentioned above, we found the Random Forest model to perform better in making predictions of street rick scores. This suggests the set of specific street features provided in the Transbase dataset are more useful in predicting street safety.

3.3 Visualization

Data is meaningless to the public without visualization. Our challenge is to design an interactive visualization that depicts the hierarchical relationship between findings for each coordinate's features and the street risk index, while preserving spatial relationships between each coordinate and each street.

We decided the platform will be a scrollable, story-telling presentation built on top of JavaScript, html, CSS and D3.js library. The first section showcases interesting insights about historical bike collision cases in San Francisco. When the end user enters the site and starts to scroll down from the first webpage, old content will seamlessly transition into new charts and insights. On each of the pages in this section, we pair a paragraph highlighting our findings with a d3 graph for storytelling. The second section is an interactive geo map of San Francisco showcasing all the historical collision data points. The map is built using MapBox-gl and D3.js. On the map, the streets are color-coded based on their risk level. Each historical collision case is represented by a dot. The end user is also able to use additional toggles to filter the visualization by collision time and facility type.

Ultimately, the goal is to provide value to the end user from the interaction with this platform. For example, an urban planner should be able to learn visual features for civil project analysis and combine his/her professional knowledge to come to a decision that maximizes safety improvement.

Section 4: Evaluation

There will be two ways of evaluating the impact of our project. One way is purely how good our models are, and another is user feedback.

4.1. Accuracy: Baseline & Random Forest

We implemented a random forest and calculated the score on the train set. In order to make sure that the model is not overfitting, a validation set was created.

The best model selected in this case renders a 65% out-of-bag R-squared.

4.2 Accuracy: Baseline & CNN

Scored on Accuracy and Macro-F1: 50% R-sqaured.

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(\hat{y}_i = y_i)$$

$$Macro - F1 = \frac{1}{C}\sum_{c=1}^{C} F1_c = \frac{1}{C}\sum_{c=1}^{C} \frac{2 \times Precision_c \times Recall_c}{Precision_c + Recall_c}$$

where:

N is the number of streets segments in the training set,
ŷi and yi are the predicted and true label of the i-th street, respectively,
𝟙 denotes the indicator function,
C is the number of possible labels,
and F1c are the precision, recall and F1 score of class c, respectively.

4. 3 User Study:

Our webpage is meant to provide an enriched set of safety features for cyclists to gauge the safety level of a street segment. We also want to enable urban designers to obtain more detailed information about San Francisco's road infrastructure and bike collision cases. WTherefore, ied out to 2 groups of prospective end users for feedback, namely the cyclists and the urban designers. Due to the limitations imposed by running the project locally, we shared a video walkthrough of the project and highlighted the key street features. Then we followed up with a survey.

To see the survey: https://www.surveymonkey.com/r/K3S6TNF

We separated the results into 2 groups (cyclists and urban designers), where each group contains 10 validate responses.

The response is positive with an average score of 8. This suggests our audience would recommend to their friends/colleagues.

4.3.1 Cyclist feedback & GUI User Friendly

Feedback highlights:

"The information provided is right to the point, the page had my attention the whole time. The color choice is great, and the map is interesting to play around!"

"Easy to use, the graphics are easy to understand."

"Details are cool, the findings are very interesting."

4.3.2 Urban Designer Feedback

Several professional urban designers were shown our tool. They tested the functionalities of the tool to determine if the project is indeed useful. They also provided valuable comments that could better our project in the future.

Feedback highlights:

"The map alone has so much information to offer, I can use the toggle to get the details that I am looking for so easily to locate."

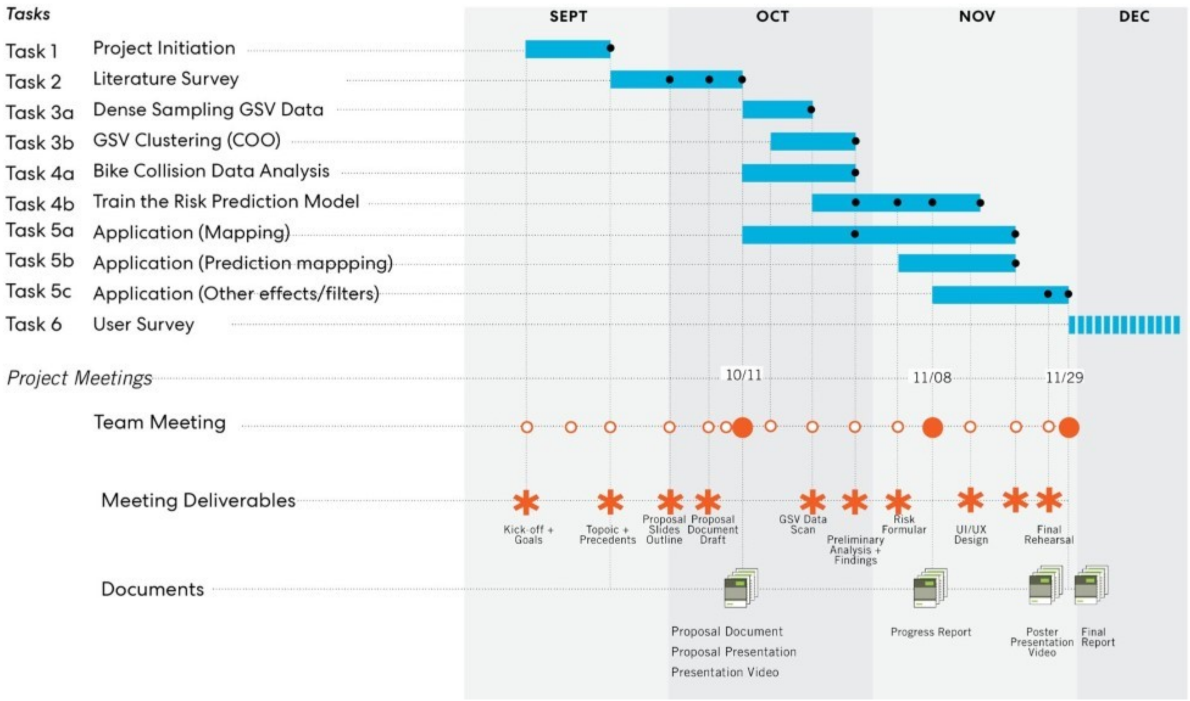
"Lots of information are contained and but not too flattery."

"How do you confirm the year street feature data was collected is the same as collision data."

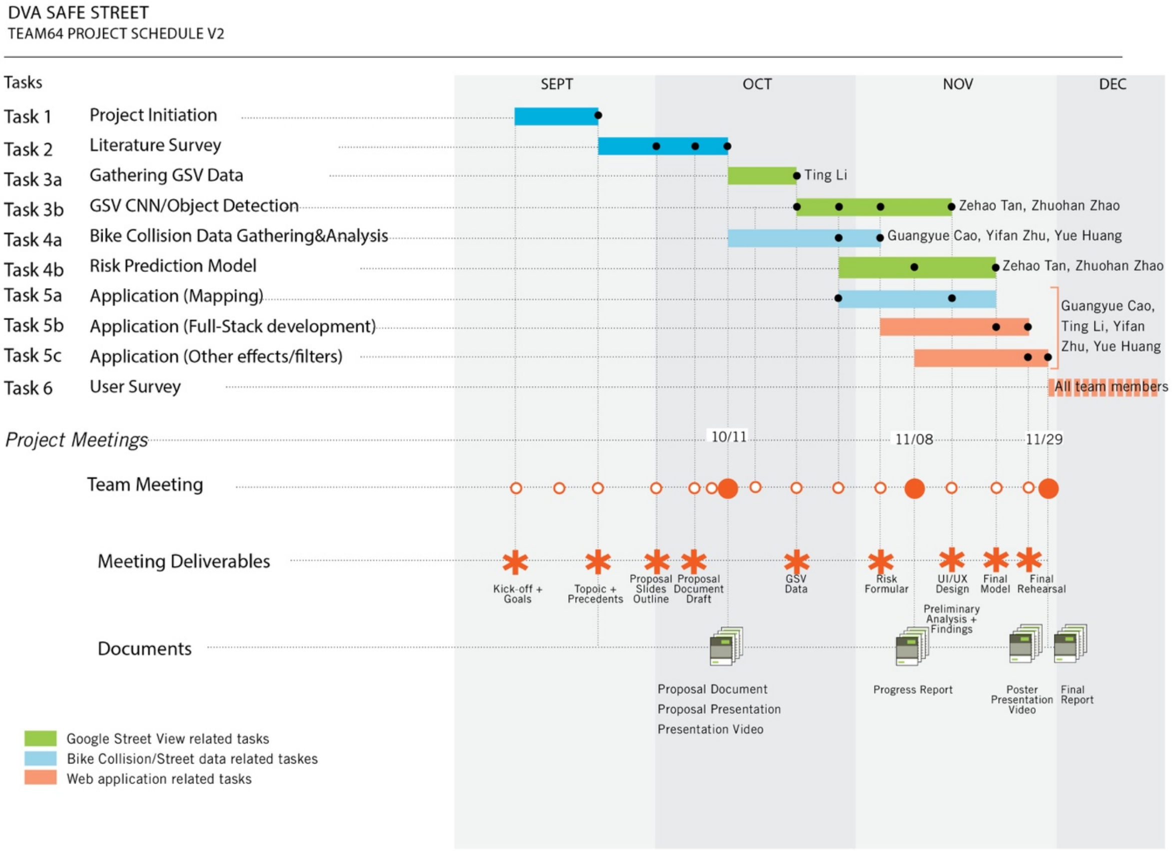
Section 5: Plan of Activities and Distribution of team effort

All team members have contributed similar amount of effort.

Version 1(as of 10/11/2019)



Updated plan of activities:

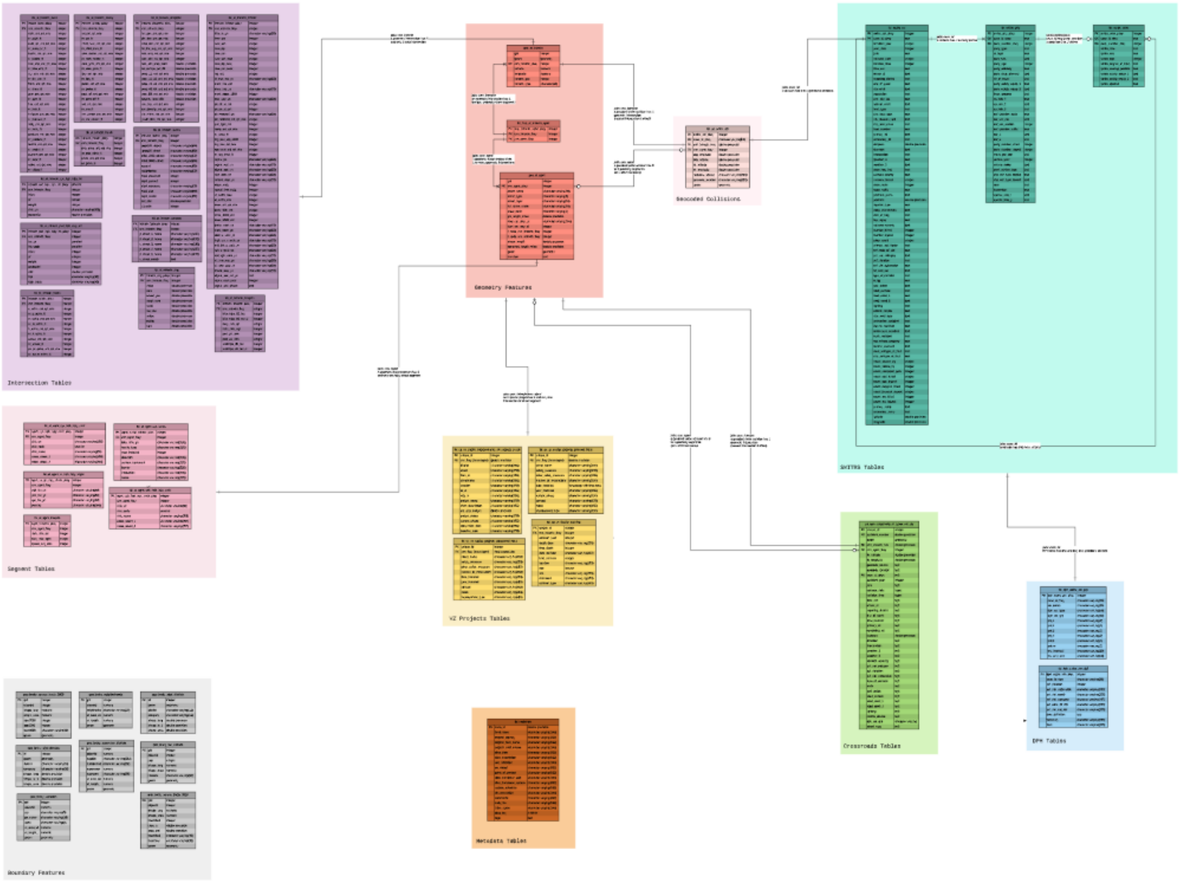


Section 6: Future Work

Due to time limitation for this project, we were not able to bring the final visualization online. With more time provided, we should continue to bring the project's web application online, add more features to allow the end user to filter/search, and enable risk score retrieval given a specific latitude and longitude. It would also allow us to reach out to a larger audience to gather more transparent feedbacks.

Appendix

TransBase EER Diagram



References

[1]V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *CoRR*, abs/1511.00561, 2015.

[2]G. Bryden. Causal exploration of bike accidents in the bay area. *Open Journal of Safety Science and Technology*, 02:75–83, 01 2012.

[3]F.-H. Chan, Y.-T. Chen, Y. Xiang, and M. Sun. Anticipating accidents in dashcam videos. volume 10114, pages 136–153, 03 2017.

[4]P. Clarke, J. Ailshire, R. Melendez, M. Bader, and J. Morenoff. Using google earth to conduct a neighborhood audit: Reliability of a virtual audit instrument. *Health Place*, 16(6):1224 – 1229, 2010.

[5]J. DiGioia, K. E. Watkins, Y. Xu, M. Rodgers, and R. Guensler. Safety impacts of bicycle infrastructure: A critical review. *Journal of Safety Research*, 61:105 – 119, 2017.

[6]C. S. Hanson, R. B. Noland, and C. Brown. The severity of pedestrian crashes: an analysis using google street view imagery. *Journal of Transport Geography*, 33:42 – 53, 2013.

[7]K. Hara, V. Le, and J. Froehlich. Combining crowdsourcing and google street view to identify street-level accessibility problems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 631–640, New York, NY, USA, 2013. ACM.

[8]D. Jones and M. Jha. The effect of urban form on traffic accident incidence. pages 212–222, 01 2010.

[9]L. Li, J. Tompkin, P. Michalatos, and H. Pfister. Hierarchical visual feature analysis for city street view datasets. 2017.

[10]J. Liang and R. Urtasun. End-to-end deep structured models for drawing crosswalks.

In *ECCV*, 2018.

[11]T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. Microsoft coco: Common objects in context, 2014.

[12]W. Lu, D. M. Scott, and R. Dalumpines. Understanding bike share cyclist route choice using gps data: Comparing dominant routes and shortest paths. *Journal of Transport Geography*, 71:172 – 181, 2018.

[13]W. E. Marshall and N. N. Ferenchak. Why cities with high bicycling rates are safer for all road users. *Journal of Transport Health*, 13:100539, 2019.

[14]D. Mohan, M. Khayesi, W. H. Organization, G. Tiwari, D. Indian Institute of Technology, and F. Nafukho. *Road Traffic Injury Prevention Training Manual*. Nonserial Publication. World Health Organization, 2006.

[15]W. E. Moritz. Survey of north american bicycle commuters: Design and aggregate results. *Transportation Research Record*, 1578(1):91–101, 1997.

[16]N. Naik, J. Philipoom, R. Raskar, and C. Hidalgo. Streetscore – predicting the perceived safety of one million streetscapes. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 793–799, 2014.

[17]H. Summala, E. Pasanen, M. Räsänen, and J. Sievänen. Bicycle accidents and drivers' visual search at left and right turns. *Accident Analysis Prevention*, 28(2):147 – 153, 1996.

[18]K. Teschke, M. A. Harris, C. C. O. Reynolds, M. Winters, S. Babul, M. Chipman, M. D. Cusimano, J. R. Brubacher, G. Hunte, S. M. Friedman, M. Monro, H. Shen, L. Vernich, and

P. A. Cripton. Route infrastructure and the risk of injuries to bicyclists: a case-crossover study. *American journal of public health*, 102(12):2336—2343, December 2012.

[19]M. Winters and K. Teschke. Route preferences among adults in the near market for bicycling: Findings of the cycling in cities study. *American Journal of Health Promotion*, 25(1):40–47, 2010. PMID: 2080

[20] SF Environment.org. Focus 2030: A Pathway to Net Zero Emissions. Climate Report. July 2019. https://sfenvironment.org/download/focus-2030-a-pathway-to-net-zero-emissions-climate-report-july-2019