A Video Analytics Infrastructure Platform for Connected Vehicles andTransportation Planning

Srinivasa Narasimhan
https://orcid.org/0000-0003-0389-1921

Robert Tamburo
https://orcid.org/0000-0002-5636-9443

Project ID: #217

FINAL RESEARCH REPORT

Contract # 69A3551747111

1 Problem

Self-driving cars and smart cities are the future of mobility. Once considered science fiction and fantasy, self-driving cars offer independence for seniors and people with disabilities, greater road safety, cost savings through ride sharing, increased productivity, reduced congestion, and reduced fuel use and carbon emissions. The technology enabling self-driving cars is rapidly improving. However, to get the best performance from self-driving cars, especially in busy urban areas, they need to be connected to a smarter roadway infrastructure. Smart cities have the potential to not only improve road safety and mobility, but also enable pro-active maintenance and better planning for the future, and provide better services.

Human error causes more than 90% of crashes, so it is not surprising to see why there has been so much excitement surrounding the development of advanced driver-assist systems and autonomous driving vehicles. Car crashes and fatalities have been on a downward trend for the past few years, but there are still 6 million car crashes per year with 3 million injuries and 90 deaths per day (CDC). Intersections are particularly dangerous sections of the roadway system due to being points of conflict between vehicles, pedestrians, and bicyclists. According to the Federal Highway Administration, more than 25% of all fatal crashes and 50% of crashes occur at intersections resulting in over 3 million accidents at a cost of over $100 billion per year, and according to the Center of Disease Control pedestrians are killed every 88 minutes.

This research is focused on instrumenting the infrastructure with image sensors and edge computing with capabilities for sharing timely information – computed by novel developed algorithms – about the road environment. Visual data contains extremely valuable information, but must be processed and interpreted at the edge because of bandwidth limitations. Through the field of computer vision, large amounts of visual data can be mined to extract useful information for understanding the world. This information is just a small fraction of the imaging data, which is not bandwidth limited. Cameras located high above the road in the infrastructure, for example, on utility or traffic signal poles provide a broad view of the road free of occlusions captured by sensors on vehicles at ground level. The infrastructure platform automatically captures, processes and interprets visual data, which can be wirelessly transmitted to connected vehicles for use by approaching (semi-)autonomous vehicles or to work stations for real-time observation. For example, detected stopped vehicles, fallen pedestrians, etc can potentially be relayed to connected vehicles for appropriate evasive maneuvers and vehicle counts, near misses, etc can be used by planners to improve mobility.

Advanced algorithms have been developed in the areas of pedestrian/vehicle detection and 3D/4D reconstruction, which yields valuable information such as object type (vehicle type, pedestrian, bicyclist, etc), speed, density, trajectory, and anomalies (accidents, large debris, etc). Algorithms were developed to address multiple challenges. First, multi-view camera systems require that the cameras are temporally synchronized and calibration. Synchronization requires additional cumbersome cables during installation or precise software control of the cameras. Camera calibration is a tedious and disruptive process, which cannot be regularly repeated in an outdoor deployment. Second, the road environment has many objects which occlude the sight line of the camera to vehicles and people making detection, tracking, and reconstruction of objects of interest very challenging. The algorithms that were developed overcome all of these challenges to provide the opportunity for high level analytics for a rich understanding of the road environment. Several camera/computer systems were deployed in the Pittsburgh area for research and development.

2 Approach and Methodology

2.1 System Deployments for Development and Data Collection

We aimed to build a scalable platform for video analytics. It is intended to serve as a test bed for computer vision research and cloud-to-edge computing systems. An example compute box contains four NVidia Tegra TX-2 GPUs, and four Intel NUCs. Both GPUs and CPUs are used as a heterogenous system. Data can be stored locally on hard drives. The compute box can manage up to eight 4 megapixel Gigabit ethernet cameras with high quality lens optics. Video ingestion and analytics are performed in real-time. A pipeline-based programming model defines complex operations on multiple camera streams. Support is provided for multiple camera operation modes. There are currently four deployments within the City of Pittsburgh, which includes a total of 18 cameras. Three deployments are on Carnegie Mellon University's campus and one deployment is at the intersection of 5th Ave and Craig Street. Cameras capture data necessary for algorithm development for object detection and air quality assessment.

Figure 1: Deployment of cameras and edge computers to the Craig Street and Fifth Avenue intersection in Pittsburgh, PA. Eight cameras were mounted on traffic signal poles. The cabinet containing the computers and other equipment were mounted on a nearby utility pole.

These works are part of "Platform Pittsburgh," which is a collaborative effort between researchers at Carnegie Mellon University, the City of Pittsburgh, and other partners with the goal of creating a "living laboratory" for conducting smart city research and analytics. As Pittsburgh continues to adopt state of the art technology, this project harnesses the power of edge computing and visual data. The goal is to lay the groundwork for a testbed in order to develop applications and produce statistics towards improving quality of life. We hope to foster interdiscplinary collaborations along with feedback from the community to solve real-world challenges in transportation mobility and safety. An integrated approach will be taken that uses computer vision, machine learning, and simulations to produce data to city planners, traffic engineers, and decision makers. Research will be directed towards applications that improve efficiency, health, safety, and overall quality-of-life. The website[1] describes the project, provides an FAQ, and has live streams of video feeds and computed analytics.

[1] `http://platformpgh.cs.cmu.edu/live_stream/`

With assistance from the the City of Pittsburgh Department of Mobility and Infrastructure, one intersection has been instrumented with 8 cameras in order to capture a full view of the intersection. Cameras are mounted on traffic signal poles in order to provide the best view of the intersection. A cabinet containing electronic equipment is mounted on nearby poles (Figure 1). Business class wired internet is provided through a partnership with Comcast. Wired internet permits transferring large amounts of visual data for training and evaluation purposes. Intersections are of primary interest to this project because approximately 2.5 million accidents occur at intersections, which accounts for nearly 40% of all nationwide accidents. Initial deployments will be at intersections around the Carnegie Mellon University campus in Oakland due to the large number of vehicular, pedestrian, and bicyclist traffic. Another set of cameras are on the rooftop of a CMU building (III). Two cameras are directed towards and intersection while the third camera observes a construction site across the street (Figure 1 Right). Cameras were also deployed inside of the CIC building to observe Forbes Avenue. Handheld videos were also captured from the ground with GoPros and Smartphones. Another source of data were online, live video feeds. For example, a live YouTube feed in Jackson, Wyoming[2] allowed us to demonstrate generalizability of our algorithms that were trained on Pittsburgh streets. Access to all of this data allowed us to train models and develop algorithms. Once algorithms were developed, they were deployed to the compute nodes for testing. Publicly available datasets are described in Section 3.

2.2 Dynamic 3D Reconstruction of Vehicles

Video cameras are becoming increasingly common at urban traffic intersections. This provides us a strong opportunity to reconstruct moving vehicles crossing those intersections. There has been a rich history of detection, tracking, and reconstruction of vehicles. Their performances are progressively improving thanks to recent advances in deep learning. In particular, detection of parts of vehicles (wheels, headlights, doors, etc.) across multiple views is becoming increasingly reliable. However, the detected part locations are still not precise enough to directly apply triangulation based 3D reconstruction methods, and are incomplete in the presence of occlusions. For the same reason, tracking via per-frame detection is not stable enough to be useful for structure-from-motion approaches. Detected part locations are referred to as structured points. There has also been significant work on tracking feature points in structure-from-motion approaches applied to a video from a single moving camera. But corresponding these features across wide-baseline views is near impossible given that each camera sees only parts of a vehicle (front, one side, or back) at any given time instant. These feature points do not often have a semantic meaning (like the structured parts); these are unstructured points.

For this work, a comprehensive framework was developed that fuses (a) incomplete and imprecise structured points (part detections) across multiple views with (b) precise but sparse single-view tracks of unstructured points, to reconstruct moving vehicles even in severe occluded scenarios. This framework is called "CarFusion" and it consists of three main stages: (1) a novel object-centric (as opposed to feature-centric) RANSAC approach to provide a good initialization of the 3D geometry of the structured points of the vehicle, (2) a novel algorithm that fully exploits the complementary strength of the structured and unstructured points via rigidity constraints, and (3) closing-the-loop by reprojecting the reconstructed structured points to all views to retrain the part

[2] `https://www.youtube.com/watch?v=1EiC9bvVGnk`

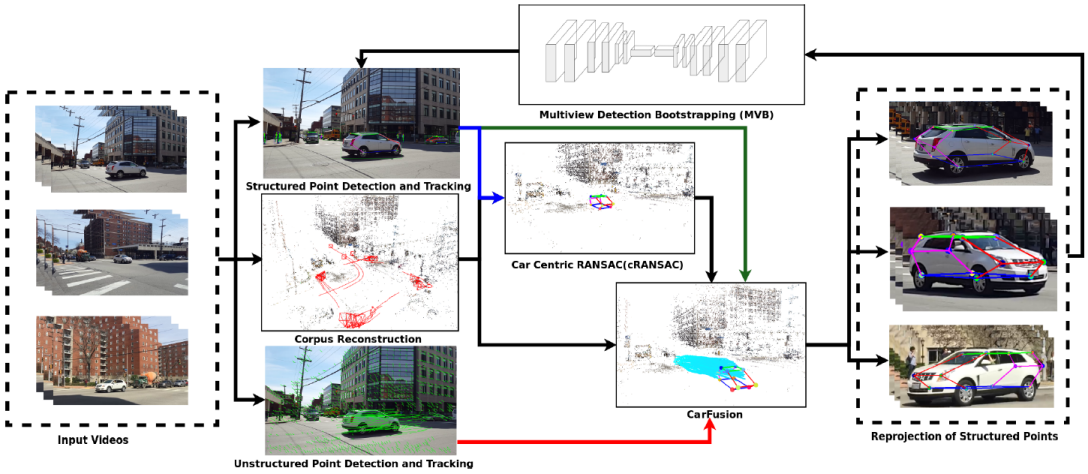

Figure 2: Overall pipeline for dynamic 3D reconstruction of multiple cars from uncalibrated and unsynchronized video cameras. We fuse the structured points (detected vehicle parts) and the tracks of the unstructured feature points to obtain precise reconstruction of the moving vehicle. The reconstructions are reprojected into all the views and are used to bootstrap and improve the detectors.

detectors. A full end-to-end system was implemented that also includes a pre-processing stage to self-calibrate and synchronize the cameras by adapting recent prior works. A detailed overview of our system is illustrated in Figure 2. Results of this work are described in Section 3.1.

2.3 Occluded Vehicle Keypoint Detection

Virtually any scene has occlusions. Even a scene with a single object exhibits self-occlusions - a camera can only view one side of an object (left or right, front or back), or part of the object is outside the field of view. More complex occlusions occur when one or more objects block part(s) of another object. Understanding and dealing with occlusions is hard due to the large variation in the type, number and extent of occlusions possible in scenes. As such, occlusions are an important reason for failure of many computer vision approaches for object detection, tracking, reconstruction and recognition, even today's advanced deep learning based ones. The computer vision community has collectively attempted numerous approaches to deal with occlusions for decades. Bad predictions due to occlusions are dealt with as noise/outliers in robust estimators. Many methods provide confidence or uncertainty estimates to downstream approaches that need to sort out whether the uncertainty corresponds to occlusion. But it is hard to predict performance as they usually do not take occlusions explicitly into account.

We developed a framework called "Occlusion-Net" to explicitly predict 2D and 3D keypoint locations of the occluded parts of an object using graph networks, in a largely self-supervised manner (Figure 3). Our method receives as input, the output of any detector (e.g., using the MaskRCNN architecture) that has been trained on a particular category of object with human supervision of only visible keypoints and their types (e.g., front, back, left, right). Implicitly, then, the key points that are not labeled are assumed to be invisible. This is the only human supervision used in this work. The detector usually provides an uncertainty of all key point locations. We first show that

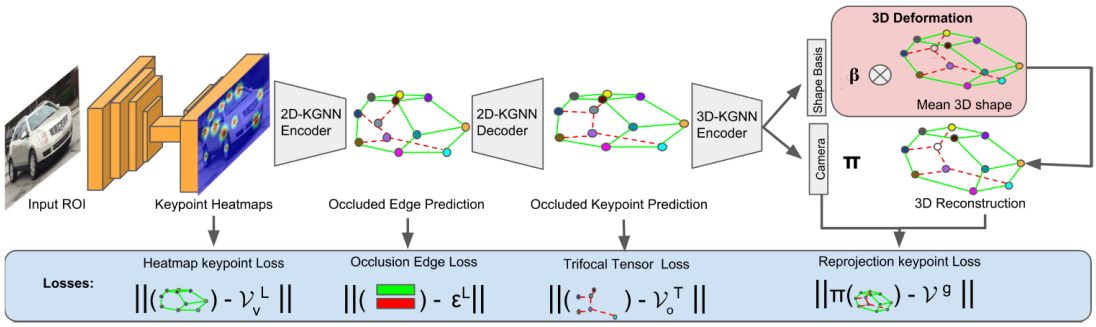

Figure 3: Occlusion-net: We illustrate the overall approach to training a network to improve localization of occluded keypoints. The input is a ROI region from any detector, which is passed through multiple convolutional layers to predict the heatmaps with a confidence score. These confidences are passed through a graph encode-decoder network and trained using multi-view trifocal tensor loss for localization of occluded 2D keypoints. The output from the decoder is passed through a 3D encoder to predict the shape basis and the camera orientation. This network is a self-supervised graph network and trained using reprojection loss with respect to the 2D decoder output.

the distribution of the uncertainties for visible and occluded points overlap significantly, making it hard to predict which key points are occluded at test time. To address this issue, we designed an encoder-decoder graph network that first predicts which edges have an occluded node, and then localizes the occluded node in 2D in the decoder. Visible or invisible edge classification is trained using the implicit non-labeled supervision of occluded points.

We then train the decoder graph network to localize invisible keypoints using multiple wide-baseline views of objects. Our observation is that while some parts may be missing in one view, they are visible and labeled in another view. But how do we provide supervision for a hidden point location in a view? We use two views where a keypoint is seen (and labeled by humans) and compute the trifocal tensor using camera matrices to predict its location in the view where the keypoint is occluded. We call this the Trifocal tensor loss, which is minimized to correct the 2D keypoint positions from the initial detector. Compared to other approaches that use multiple views, our approach explicitly predicts occluded keypoints.

The predicted 2D keypoints (both occluded and visible) are then used in a graph network to estimate the 3D object shape and the camera projection matrix. Similar to previous work, we will estimate the parameters of a shape basis computed a priori of the object of interest. The training is performed in a self-supervised way by minimizing the reprojection loss i.e. error between the reprojection and the predicted 2D keypoint locations. Finally, we train the entire pipeline end-to-end with the aforementioned losses. Results of this work are described in Section 3.2.

2.4 4D Reconstruction and Analysis of Vehicular Traffic

Most traffic analysis systems, vision-based or otherwise, estimate the average number and speed of vehicles passing through a particular location on the road or an intersection. These aggregate analytics report useful traffic patterns to all drivers and can be used to improve the flow of traffic through a city. But individual vehicles, and traffic in general, constitute much more richer behavior than the aggregate statistics suggest. Consider a busy city intersection. How many vehicles turn

left, right, make a U-turn or sudden stop, pass straight or change lanes? How many near-accidents or near-misses occur? Which events are anomalous or rare? How many drivers are conservative or aggressive? Obtaining such a rich description of traffic behavior requires us to go beyond the estimations of number and average speeds of vehicles.

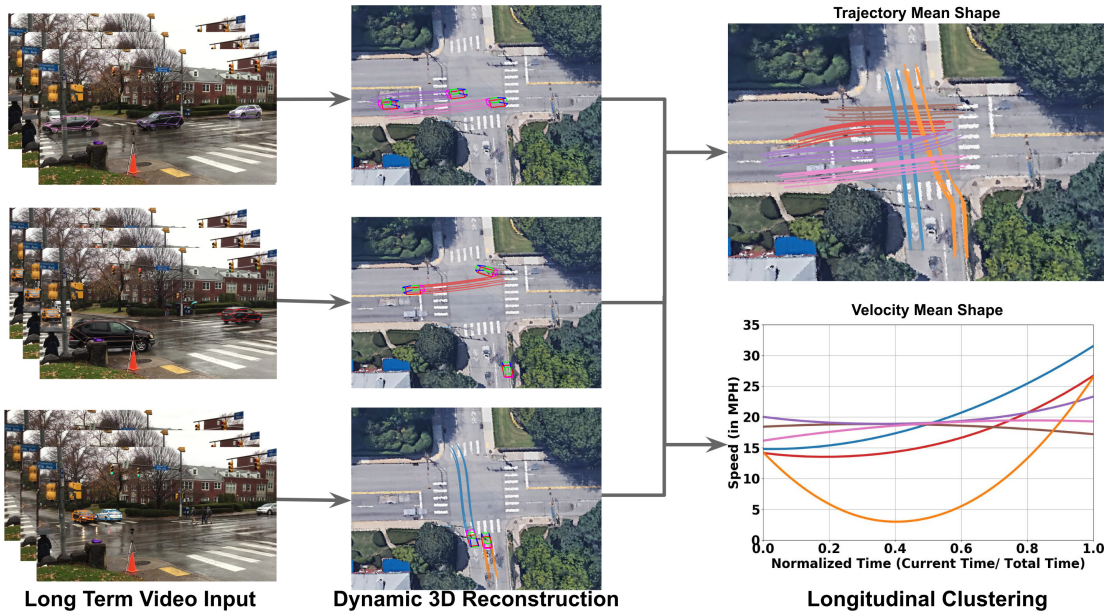

Figure 4: Offline computation of longitudinal vehicle trajectory clusters. The input to the system is a sequence of videos at different time instances from a single stationary camera. We compute dynamic 3D reconstruction(Overlay on the Google map) with the help of 2D keypoint tracks over an extended duration of time.These reconstructions are fit to clustering model and their derivatives are used to compute the velocity profiles. Mean shapes of trajectories and velocities are shown here.

We have developed a novel approach for 4D reconstruction to estimate the 3D trajectories and time-varying velocities of vehicles from a single camera view. A stationary calibrated camera observes the traffic pattern over an extended period of time. The overhead Google Street View images of the location are used to estimate the unknown scale to recover absolute positions and velocities. The vehicle trajectories captured over time are clustered according to their shapes and instantaneous velocity profiles. The clustered trajectories are built progressively and can be used as a model to fit to (or regularize) later vehicle trajectories in real-time. An overview of our approach is illustrated in Figure 4.

Key to our approach is the optimization that exploits the following observations: (a) vehicles repeatedly follow similar trajectories (lanes) so even though 2D track-lets of individual vehicles may be short or erroneous, groups of vehicles complete the trajectories, (b) vehicles are rigid and approximately lie on a single ground plane making it possible to jointly optimize trajectories of all vehicles and (c) extrinsics of the vehicles (rotation and translation) vary smoothly frame-to-frame regularizing the optimization.

For individual vehicles, the optimization uses detected 2D key points to estimate the coefficients of a PCA shape basis and per-frame 3D rotation and 3D translation. The 3D motions are fit to a low-order polynomial (spline) to obtain complete trajectories and their derivatives yield in-

stantaneous velocities. Recent approaches have either reconstructed vehicles from a single image which often do not produce temporally consistent shapes or poses. Or, they use multiple (four or more) views of an intersection to reconstruct the vehicle trajectories. We are not aware of an approach to obtain longitudinal 3D trajectories of vehicles from a single stationary view.

The estimated longitudinal trajectories of vehicles are then clustered based on their shape and velocity profiles. We propose a novel two stage hierarchical clustering framework to bundle finer details of trajectories. These clusters correspond to vehicles moving in different lanes, turning in different directions, and changing lanes. The velocity profile clusters correspond to vehicles coming to a stop and then accelerating, vehicles changing speed while turning, vehicles rapidly decelerating or accelerating, etc. Using these clusters, we are able to determine anomalous or rare events corresponding to the trajectories and velocities that do not belong to any "usual" cluster. Examples of anomalous events that are automatically detected include: (a) sudden stops, (b) near accidents or misses and (c) erroneous lane changes. We demonstrate the versatility of our approach at four intersections in different cities and suburban areas. The data was obtained from different sources: (a) the AI city challenge, (b) Live YouTube videos, and (c) an intersection captured by us to validate the approach better. We believe this richer understanding of vehicle behavior and traffic pattern can be useful for autonomous vehicles operating in the area, to inform all drivers to expect the unexpected, and the city authorities to plan for smarter and safer cities. Results of this work are described in Section 3.3 along with preliminary work applies our framework to the COVID-19 pandemic in Section 3.4.

3 Findings

3.1 Dynamic 3D Reconstruction of Vehicles

The framework was evaluated on a traffic scene captured with six Samsung Galaxy 6, ten iPhone 6, and six Gopro Hero 3 cameras at 60 fps in a busy intersection for 3 minutes. In total, the algorithm is run on nearly 210000 frames. These videos were captured by 13 people, some of whom carried two cameras. The sequence is challenging as there are no constraints on the camera motion or the vehicle motion in the scene. 2D locations of the structured points were manually annotated for every visible cameras for 2793 frames from different viewing angles in the dataset, which is available to download for research purchases (see below).

The reconstruction of vehicles at a busy intersection was demonstrated with real data captured from an intersection as shown in Figure 5. About 62 vehicles were detected, tracked and reconstructed within a 3-minute duration captured from 21 handheld cameras that are uncalibrated and unsynchronized and were panning to cover wider fields of view. A subset of vehicle structured point trajectories are augmented within the Google Earth image of the intersection. They include cars of different types (sedans, SUVs, hatch-backs, jeeps, etc.) making left and right turns, going straight-ahead as well as changing lanes. Several views of two specific cars in various occluded scenarios are shown with the reprojections of the structured points.

The performance of each stage of our framework was evaluated and compared to alternate methods that rely only on tracking-by-detection or feature based structure-from-motion. Compared were the detection of the structured points before and after multiview bootstrapping with respect to the ground truth labels for two cars observed in three different views. Our multiview bootstrap-

Figure 5: Reconstruction of vehicles crossing a busy intersection, making turns, going straight and changing lanes. A subset of vehicle skeletons (3D detector locations) and their 3D trajectories are augmented within the Google Earth view of the intersection. The reconstructions are reprojected into multiple views of two cars (a sedan and an SUV) demonstrating good performance under partial occlusions.

ping methods showed clear improvements over the baseline method as more confident points are accurately detected. Using our approach, the reprojected points accurately localize the structured points and provide plausible prediction for occluded locations. We attribute this property to the use of symmetry, link length, and rigidity constraints in the reconstruction stage. Although some of the structured points are not visible from any of the views, we are still able to accurately reconstruct the point in 3D due to our left-right symmetry and link length constraints. Without these constraints the reconstruction of the structured points, even fully visible from multiple views, often explodes due to erroneous detection hypothesis. In Figure 6, we illustrate the complete 3D reconstruction of trajectories of structured points on moving cars using our method and the 2D projection to inlier views for several cars. As can be seen from the results we are able to accurately reconstruct the trajectories of the cars over time captured from unsynchronized videos.

Project Website: `https://www.cs.cmu.edu/~ILIM/projects/IM/CarFusion/cvpr2018/index.html`

Publication: N. Dinesh Reddy, Minh Vo, and Srinivasa Narasimhan, "CarFusion: Combining Point Tracking and Part Detection for Dynamic 3D Reconstruction of Vehicle," IEEE Conference

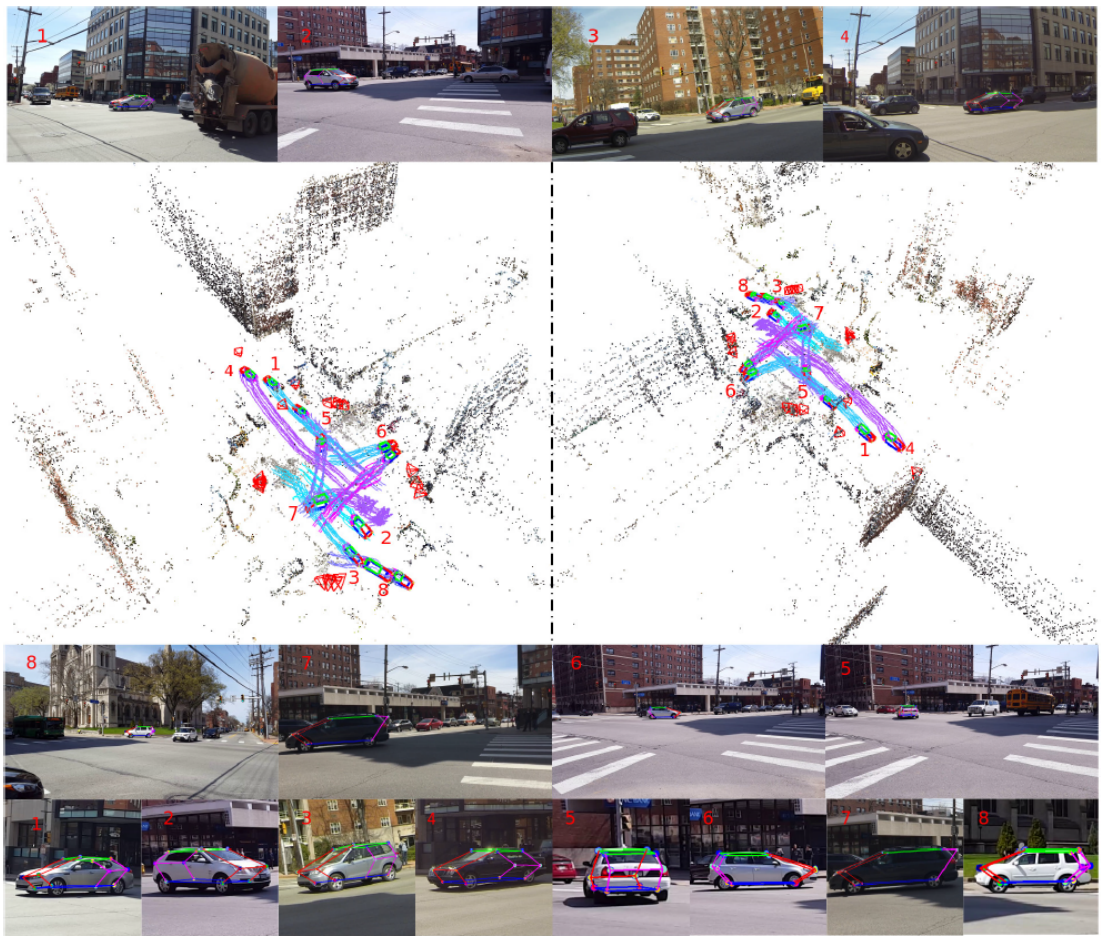

Figure 6: Visualization of 8 reconstructed vehicles using our method. We show the 2D re-projection of the reconstructions onto sample frame containing those cars. All the re-projected points fit the cars well.

on Computer Vision and Pattern Recognition (CVPR) 2018. Publication, poster, and supplementary materials can be downloaded from **Project Website**.

Dataset: A dataset is available that provides 1) 53,000 images captured from 18 moving cameras at multiple intersections in Pittsburgh, PA and 2) manual annotations of 14 semantic keypoints for 100,000 vehicle instances (sedans, SUVs, buses, and trucks). A python script for viewing the labels and python wrapper for reading the data in the common COCO format is also provided. The dataset and code are available on the **Project Website**. Dataset continues to be updated.

3.2 Occluded Vehicle Keypoint Detection

We evaluated our "Occlusion-Net" framework on images of vehicles captured at busy city intersections with numerous types and severity of occlusions. The dataset extends the previous CarFusion dataset (3.1) to include many more city intersections, where 18 views of the intersection are simultaneously recorded. A MaskRCNN car detector is trained using 100000 cars, with human labeled visible keypoints to produce a strong baseline for our method to compare to and build upon.

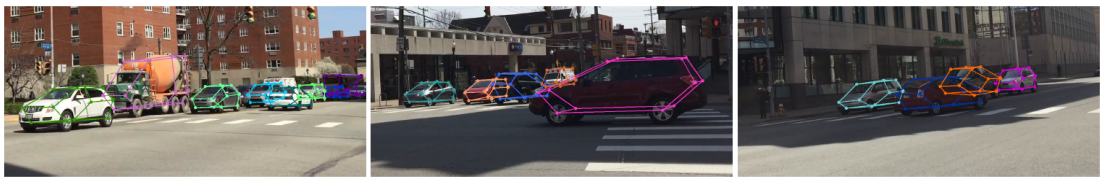

Figure 7: Example results of occlusion-net on sample images of the CarFusion dataset. We accurately localize occluded keypoints under a variety of severe occlusions. See supplementary for additional results. Different colors depict different vehicles in the scene.

Our Occlusion-net significantly outperforms (about 10%) this baseline across many metrics and performs well even in the presence of significant occlusions (see Figure 7). As an interesting exercise, we also show a quantiative comparison of the trifocal loss against human labeling of the 2D occluded point locations and observe that humans label around 90% of the points to lie within the acceptable range of error. We also evaluate our approach on a large synthetic CAD dataset, showing similar performance benefits and improvements of up to 20% for occluded keypoints. Our network is efficient to train and can localize keypoints in 2D and 3D in realtime (more than 30 fps) at test-time. While we have demonstrated our approach on vehicles, the framework is general and applies to any object category.

Project Website: `https://www.cs.cmu.edu/~ILIM/projects/IM/CarFusion/cvpr2019/`

Publication: N. Dinesh Reddy, Minh Vo, and Srinivasa Narasimhan, "Occlusion-Net: 2D/3D Occluded Keypoint Localization Using Graph Networks," IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2019. Publication, poster, and supplementary materials can be downloaded from **Project Website**.

Dataset: Included as part of the "CarFusion Dataset" described above.

3.3 4D Reconstruction and Analysis of Vehicular Traffic

We evaluated our approach on videos captured from four different intersections with wide variation in location and conditions. The videos are obtained from three sources, two from the AI city challenge named as *AICity1* and *AICity2*, a live YouTube feed named as *Jackson* and an intersection captured by us using iPhone 6 named as *Street*. Each video in *AICity1* and *AICity2* is around 5 minutes while the other two are over 10 minutes. These videos are split according to 90-10 ratio over time for offline and online processing respectively. These videos provide various view angles, weather conditions and vehicle motion patterns, showing that our algorithm generalizes well for different settings. Moreover, the Street sequence is captured from two views at the same time, enabling us to perform triangulation to evaluate our single view reconstruction results. We used the live YouTube sequences from Jackson to evaluate anomaly detection.

Figure 8 shows the reconstruction results during the online processing at different time instances. We plot the keypoints re-projected from 3D, vehicle speed profile, along with their 3D location and trajectories. Cars moving straight and turning can all be localized accurately. Also, speed curves record various vehicle dynamic when passing through the intersection. For instance,

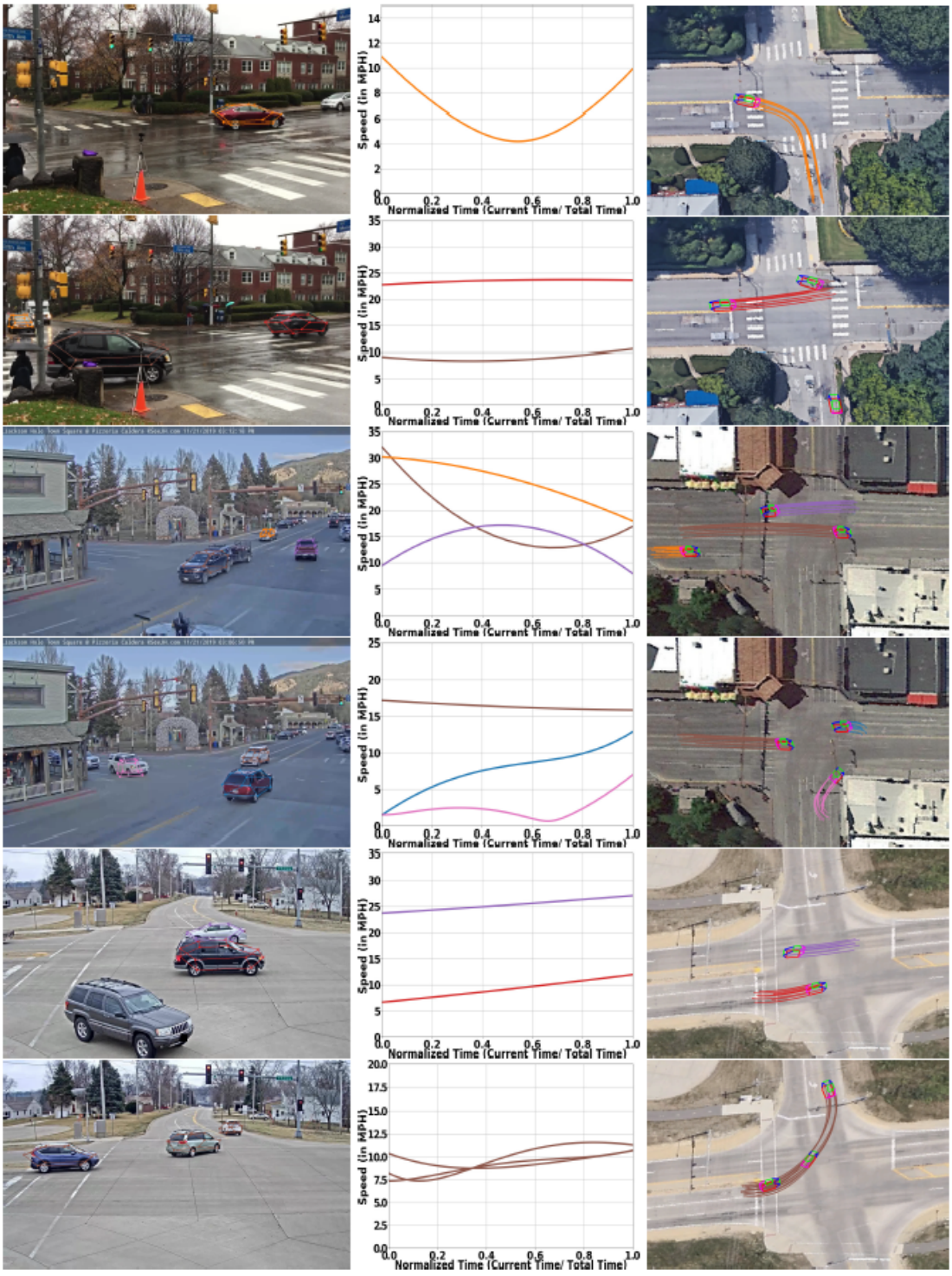

Figure 8: Online results of our reconstruction at different time instances on multiple intersections and their re-projection into the image (Left). We show the velocity profile of the vehicles to better grasp the the behaviour of vehicles (Middle) and their reconstruction being classified accurately with lane specific precision (Right).

we notice turning vehicles usually slow down before turning and then gradually speed up while forward driving cars almost have a constant speed. This work is currently under peer-review. Upon being accepted, a website will be built for this aspect of the project with links to the published paper.

Detecting Anomalous Activity: The The framework can also detect anomalous activity from a single video. During live analysis the likelihood of newly computed 3D trajectories are sampled from each cluster separately and the largest log likelihood value among them show that they are outliers compared to the clustered means. The same can be done with speed profiles. Shown in Figure 9 are automatically detected anomalies from Jackson, Wyoming live stream video sequences. After inspection, it was found that a vehicle made an unwarranted left turn, there was a near collision, and a vehicle and bicyclist collided.

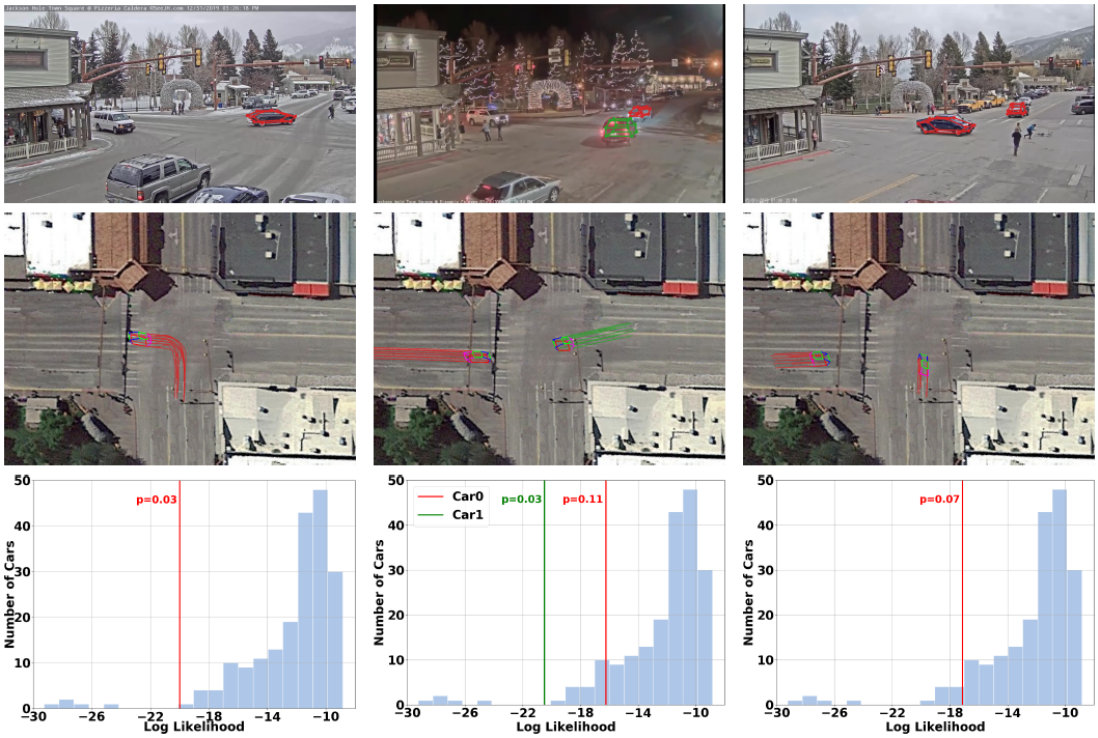

Figure 9: We show automatic anomaly detection on video sequences from Jackson, Wyoming. The plot shows accurate classification of different anomalies like vehicles making a wrong turn (Left), near misses (Middle), and accidents (Right) using our method. The anomalies are detected by a log likelihood(last row) plot of the trajectory with respect to the cluster.

Publication: *Under review*, "Traffic4D: Single View Longitudinal 4DReconstruction and Analysis of Vehicular Traffic," European Conference on Computer Vision (ECCV), 2020.

3.4 Analytics During a Pandemic

The frameworks developed during this research period were meant to provide analytics and information on the road environment. However, we were able to adapt the framework during the COVID-19 pandemic to understand activity related to the health crisis. In the first example, data captured from our cameras mounted on a CMU building were analyzed for vehicle activity in a nearby parking lot. Figure 10 compares parking lot activity one year apart. As one would hope, activity was greatly reduced after non-essential businesses were ordered to close. In the second example, models were trained to detect people, e.g., pedestrians walking on the sidewalk. Using ground plane estimation, camera parameters, and triangulation, the distance between detected people was estimated. Setting a threshold of 6 feet allowed us to determine whether people were practicing social distancing (Figure 11). In the final example, using images captured from our cameras mounted on traffic signal poles at the Fifth Ave and Craig St intersection, we developed an algorithm to determine whether detected people were wearing masks (Figure 12). We believe that the flexibility of our framework can be used during a health crisis such as the COVID-19 pandemic to provide detailed statistics that can inform related actions and corroborate assumptions for pandemic forecasting models.

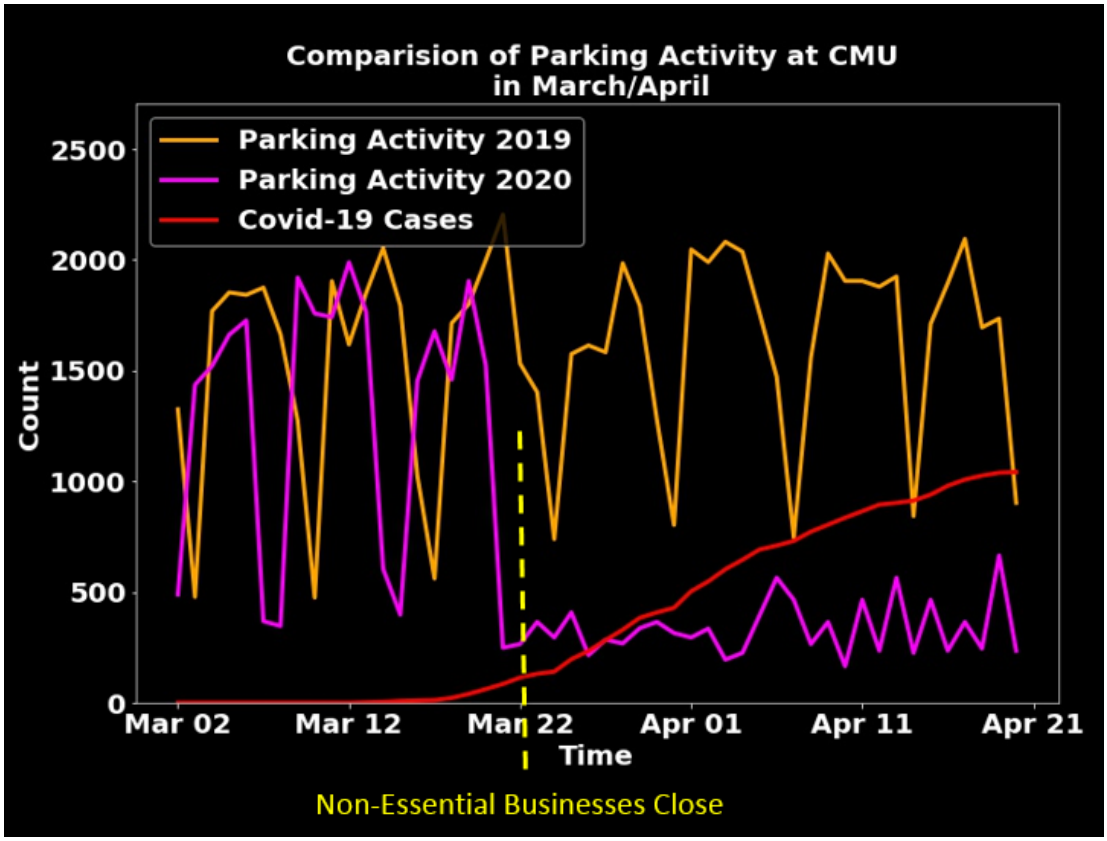

Figure 10: Long-duration video capture from our cameras mounted on the roof of a CMU building. Cameras overlook a nearby parking. Shown here is parking lot activity during COVID-19 pandemic compared to activity a year ago. Spikes for both years occur during the weekends. A noticeable drop in activity occurs when non-essential businesses were ordered to close.

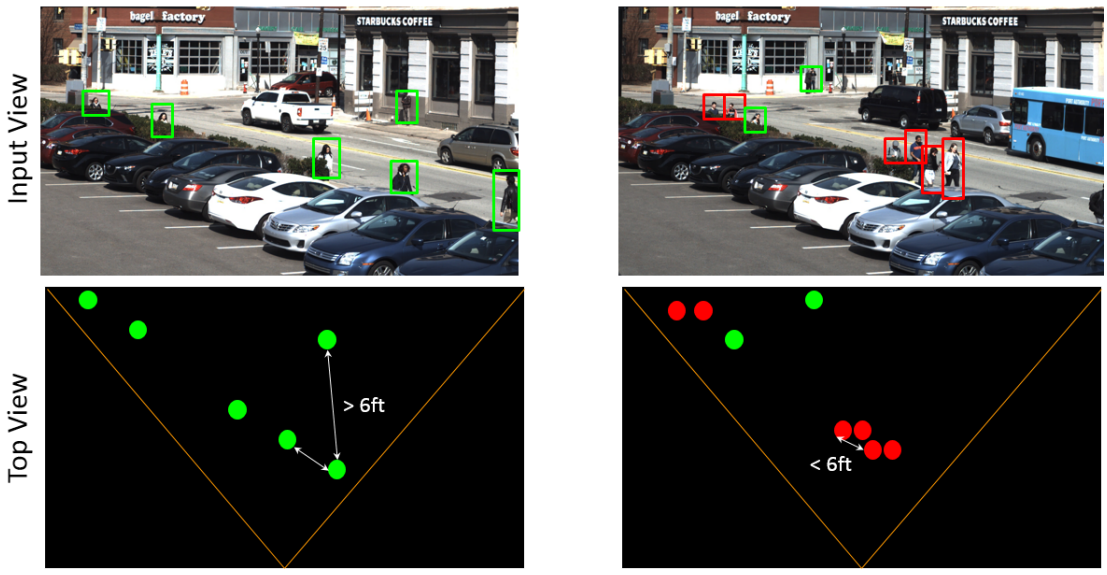

Figure 11: View of a parking lot near CMU's campus captured by our cameras mounted on the roof of a CMU building. Models were trained to detect people and the distance between detected people was estimated using ground plane estimation, camera parameters, and triangulation. A threshold of 6 feet was used to determine whether people were practicing social distancing.

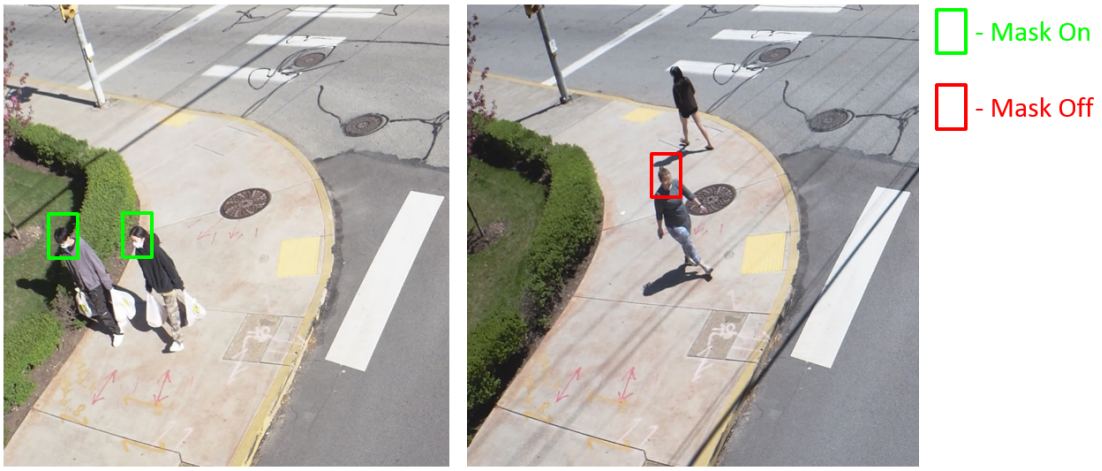

Figure 12: Frames captured with our cameras mount on traffic signal poles at the Fifth Ave and Craig St intersection. An algorithm was developed to determine whether detected people were wearing a mask.

4 Conclusions

We have presented methods to detect, track, and reconstruct vehicles (and people). These methods work with multiple or single cameras and provide accurate detection, localization, and tracking in the presence of strong occlusions. The methods are effective without calibrating or synchronizing the cameras, which is critical for real world systems. To develop, test, and prove the methods, cameras were installed in the infrastructure in Pittsburgh, PA. Cameras were installed at a busy urban intersection and on several CMU buildings. Methods were also applied to a live video stream of an intersection in Jackson, Wyoming. From these computer vision methods, analytics were computed and aggregated over time for analysis of trends. For example, vehicle activity in a nearby parking lot was shown to be much less active during the COVID-19 pandemic this year compared to last year. Performance of the methods were also quantified with ground truth data. We believe this approach to understanding the road environment can be extremely useful in providing rich information to connected vehicles and be a powerful tool for informing and answer questions by city planners, policy makers, and researchers. Examples include: How many vehicles turn left, right, make a U-turn or sudden stop, swerved, pass straight or change lanes? How many near-accidents occur? How many drivers are conservative or aggressive? What is the utilization of a playground? When (and how often) does a crowd cross this intersection? Are there times when children from local daycare cross the street?

5 Recommendations

The future of transportation will include semi- and fully autonomous vehicles that communicate with each other and the infrastructure. Sharing rich information with each other, in real-time, will improve safety and mobility on the road. Rich information can also aid city planners in developing a plan for smart cities. As demonstrated in this project, there are many other purposes that can benefit from this research. For example, we were able to quickly develop algorithms to measure vehicle activity and social distancing during the COVID-19 pandemic. These types of data can be valuable to policy makers during times of crises. This also demonstrates the value of visual data and the great flexibility of the edge computing platform. The integration of wireless communications will prove valuable for semi- or fully autonomous vehicles for their path planning and navigation. High level analytics will be critical in the monitoring of smart city activity. As more and more cities move towards incorporating smart city technology, visual cameras, edge computing, and wireless communication standards will be critical to delivering rich information.