Intersection Safety for the Vulnerable

Srinivasa Narasimhan (ORCID:0000-0003-0389-1921)
Robert Tamburo (ORCID: 0000-0002-5636-9443)
Khiem Vuong (ORCID: 0009-0006-2474-2270)
Dinesh Reddy (ORCID: 0009-0005-5945-4212)

FINAL REPORT

September 25, 2024

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, under [grant number 69A3552344811] from the U.S. Department of Transportation's University Transportation Centers Program. The U.S. Government assumes no liability for the contents or use thereof.

1. Report No. 466	**2. Government Accession No.**	**3. Recipient's Catalog No.**
4. Title and Subtitle Intersection Safety for the Vulnerable		**5. Report Date** September 19, 2024
		6. Performing Organization Code Enter any/all unique numbers assigned to the performing organization, if applicable.
7. Author(s) Srinivasa Narasimhan, Ph.D.(ORCID: 0000-0003-0389-1921) Robert Tamburo, Ph.D. (ORCID: 0000-0002-5636-9443) Khiem Vuong, M.S. (ORCID: 0009-0006-2474-2270) Dinesh Reddy, Ph.D. (ORCID: 0009-0005-5945-4212)		**8. Performing Organization Report No.** Enter any/all unique alphanumeric report numbers assigned by the performing organization, if applicable.
9. Performing Organization Name and Address Carnegie Mellon University 5000 Forbes Avenue Pittsburgh, PA 15213		**10. Work Unit No.**
		11. Contract or Grant No. Federal Grant No. 69A3552344811
12. Sponsoring Agency Name and Address Safety21 University Transportation Center Carnegie Mellon University 5000 Forbes Avenue Pittsburgh, PA 15213		**13. Type of Report and Period Covered** Final Report (July 1, 2023-June 30, 2024)
		14. Sponsoring Agency Code USDOT

15. Supplementary Notes
Conducted in cooperation with the U.S. Department of Transportation, Federal Highway Administration.
Enter information not included elsewhere, such as translation of (or by), report supersedes, old edition number, alternate title (e.g. project name), hypertext links to documents or related information in the form of URLs, PURLs (preferred over URLs - https://archive.org/services/purl/help), DOIs (https://www.doi.org/), insertion of QR codes, copyright or disclaimer statements, etc. Edit boilerplate FHWA statement above if needed.

16. Abstract
The goal of the proposed work is the development of computational methods that can be used towards enhancing the safety of vulnerable road users (VRUs) at intersections. To accomplish this goal, we envision a cyber-physical system that detects VRUs, calculates a vulnerability score (based on spatial and temporal risk), then takes appropriate action or actions to minimize the opportunity for injury. The core of the system is based on automatically detecting VRUs in visual data captured from cameras. Accomplishing this task for automation requires a couple of core components. First, understanding the scene to asses vulnerability requires real-world measurements. Unfortunately, outdoor security and traffic cameras are challenging to keep calibrated, and therefore, they do not have any calibration data required for traditional 3D computational methods. To address this issue and take advantage of the widespread use of outdoor cameras, we have developed a method to automatically calibrate any outdoor camera from street-level images. Second, detecting people with vulnerabilities requires an annotated dataset of VRUs, e.g., person walking with cane, user of a wheelchair, bicyclist, etc. There are a handful of datasets publicly available, but not nearly enough with annotated VRUs. To fill the dataset gap, we have developed a framework to use time-lapse videos from stationary cameras to synthesize realistic scenarios (with and without occlusion) by extracting unoccluded objects and compositing them back into the background image at their original positions. Using this method, public datasets can be augmented with the generated photorealistic synthetic images.

17. Key Words	**18. Distribution Statement**
Machine Learning, Computer Vision, Intersections, Vulnerable Road Users, Calibration	No restrictions. This document is available through the National Technical Information Service, Springfield, VA 22161. Enter any other agency mandated distribution statements. Remove NTIS statement if it does not apply.

19. Security Classif. (of this report)	**20. Security Classif. (of this page)**	**21. No. of Pages**	**22. Price**
Unclassified	Unclassified	Enter the total number of pages in the report, including both sides of all pages and the front and back covers.	Refers to the price of the report. Leave blank unless applicable.

1. Project Investigators and Contributors

This is a summary report combining multiple sub-projects conducted at the University Transportation Center at Carnegie Mellon University in relation to the overall project titled "Intersection Safety for the Vulnerable". The participants who contributed to the different sub-projects are listed below.

- Srinivasa Narasimhan (CMU RI Professor, ORCID: 0000-0003-0389-1921)
- Robert Tamburo (CMU RI Senior Project Scientist, ORCID: 0000-0002-5636-9443)
- Khiem Vuong (CMU RI Ph.D. Student, ORCID: 0009-0006-2474-2270)
- Dinesh Reddy (CMU RI Ph.D. Student, ORCID: 0009-0005-5945-4212)

2. Introduction

Vulnerable road users (VRUs) are considered people that are not in a vehicle and are, consequently, at a higher risk for serious injury because they have less crash protection than a vehicle occupant. Pedestrians, bicyclists, motorcyclists, and road workers are common VRUs. Vulnerable road users can be further categorized by their degree of mobility, perception, and cognition and vulnerabilities also have spatial and temporal dependencies, which can be categorized from very low risk to very high risk. For example, running on a sidewalk in the middle of the day has an associated very low risk. However, jaywalking across the road during rush hour might have a high or very high risk.

The goal of the proposed work is the development of computational methods that can be used towards enhancing the safety of VRUs at intersections. To accomplish this goal, we envision a cyber-physical system that detects VRUs, calculates a vulnerability score, then takes appropriate action or actions to minimize the opportunity for injury. For example, if a person falls out of their wheelchair in the middle of a signalized intersection, all of the traffic signals would stay red, emergency medical vehicles would be dispatched, and audiovisual warnings would be broadcast.

The core of the system is based on automatically detecting VRUs in visual data captured from cameras. Accomplishing this task for automation requires a couple of core components. First, understanding the scene to asses vulnerability requires real-world measurements. Unfortunately, outdoor security and traffic cameras are challenging to keep calibrated, and therefore, they do not have any calibration data required for traditional 3D computational methods. To address this issue and take advantage of the widespread use of outdoor cameras, we have developed a method to automatically calibrate any outdoor camera from street-level images. Second, detecting people with vulnerabilities requires an annotated dataset of VRUs, e.g., person walking with cane, user of a wheelchair, bicyclist, etc. There are a handful of datasets publicly available, but not nearly enough with annotated VRUs. To fill the dataset gap, we have developed a framework to use time-lapse videos from stationary cameras to synthesize realistic scenarios (with and without occlusion) by extracting unoccluded objects and compositing them back into the background image at their original positions. Using this method, public datasets can be augmented with the generated photorealistic synthetic images. We are currently leveraging some effort from this project for Phase 1B of the USDOT ITS Intersection Safety Challenge[1].

3. Traffic Camera Calibration

Content for this section has been summarized from our work published at the 2024 Winter Conference on Applications of Computer Vision. For more detail on this work, see [1].

1.1 Method

The first objective is to construct a metric 3D reconstruction of the scene surrounding a chosen traffic camera's location, typically an intersection. To perform the reconstruction, we leverage Google Street View (GSV) to build the scene's geometry around a specific GPS location. GSV is a street-level database consisting of millions of panorama images. Every panorama image has a 360° horizontal and 180° vertical field-of-view and is geo-tagged with GPS coordinates. The top left of **Figure *1*** for samples of panorama images and top right of **Figure *1*** for overview of framework.

[1] https://its.dot.gov/isc

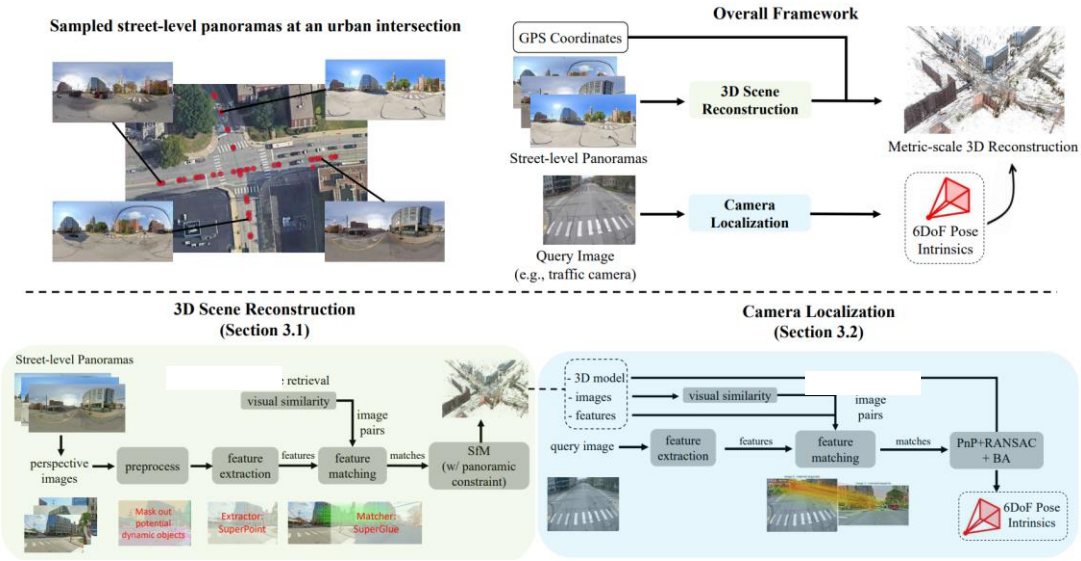

Figure 1: Top: The scene in 3D for a metric-scale representation is reconstructed using street-level panoramas and GPS data from Google Street View. Camera localization to determine intrinsic parameters and camera pose with respect to the 3D scene is performed with a query image from a traffic camera. Bottom: More details on 3D Scene Reconstruction (left) and Camera Localization (right). Illustration credit [1].

To reduce reconstruction errors caused by dynamic objects, semantic segmentation is used to suppress feature extraction around dynamic objects. Images are matched based on the nearest visual neighbor. Panoramic constraints for bundle adjustment are enforced by incorporating known relative poses between frames within the same panorama. This constraint helps reduce camera pose error during 3D reconstruction, which is valuable when working with a limited number of images. A metric scale 3D scene reconstruction is created by geo-registering the reconstruction with an optimized similarity transform. A road plane is estimated by fitting a plane to pixels on lane markings, which are obtained with semantic segmentation. Finally, a traffic camera can be localized by estimating its intrinsic and extrinsic parameters with respect to the scene. To accomplish this, a visual localization pipeline localizes images from the traffic camera within the computed metric scale 3D scene reconstruction.

1.2 Results

We compared our method for estimating intrinsic parameters to state-of-the-art methods (SOTA) by using checkerboard-based calibration as ground-truth (mean error). On average, our method outperforms SOTA methods for calculating focal lengths by at least 122% and principal points by 25%. We also assessed our method to estimate distance by calculating normalized error between pairs of manually selected points on the estimated road plane. Our method outperforms SOTA methods by at least 134% (Max Error %), 239% (Median Error %), 175% (RMSE %). We demonstrated our calibration framework on 100 traffic cameras around the world **Figure *2***. Finally, our method was applied to traffic cameras at intersection in a Pittsburgh suburb to generate activity heatmaps to visualize vehicle activity and to estimate the speed of vehicles as they cross a virtual speed trap (line drawn on ground plane).

Figure 2: Our framework allows 3D scene reconstruction and precise localization of over 100 real-world traffic cameras distributed globally across multiple countries, with the potential to scale to any camera with sufficient street-level imagery. (Left): Highlighting the reconstruction and localization of traffic cameras at specific chosen locations. (Right): Demonstrating 7 cameras positioned within an urban intersection, accurately localized with respect to the reconstructed 3D scene. Illustration credit [1].

1.3 Conclusions

We have developed a scalable framework that enables accurate calibration of real-world traffic cameras. The framework's value in traffic analysis was demonstrated with vehicle speed measures and vehicle activity heat maps, which is valuable information that can be used for improving transportation systems. Since our approach can be applied to any traffic camera with sufficient nearby street-level imagery, e.g., Google Street View, the thousands of cameras across the U.S. have the potential to become a rich source of valuable visual information about vulnerable road users and road activity in general.

4. Generating Realistic Images

Content for this section has been summarized from our work published and presented at the 2024 Computer Vision and Pattern Recognition conference. For more detail on this work, see [2].

1.1 Method

From a time-lapse video, we identify unoccluded objects and extract their 2D attributes such as segmentation and keypoints. We then use off-the-shelf 3D object reconstruction methods to obtain the pose and shape of these objects, constrained by the camera intrinsics and ground plane. After that, we re-insert these non-intersecting 3D objects into the background image at their original positions in a "clip-art" style, arranged based on their distance from the camera to ensure the occlusion configurations are physically accurate and realistic. Finally, we use the clip-art composited image together with its pseudo-ground truth supervision data to learn robust 2D/3D object reconstruction under occlusion. An overview of the method is illustrated in **Figure 3.**

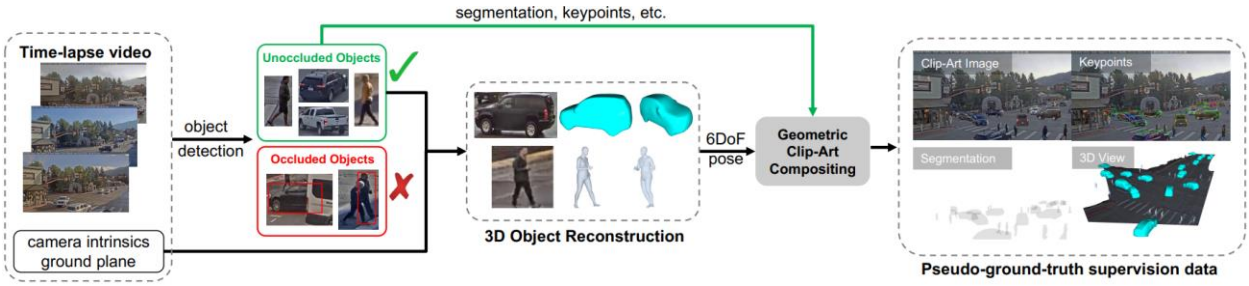

Figure 3: Given a time-lapse video, we automatically generate 2D/3D training data under severe occlusions. We start by detecting each object in the video, and unoccluded (fully visible) objects are identified. Each unoccluded object is then reconstructed using the ground plane and camera parameters. With the 3D pose, unoccluded objects are composited back into the same location (i.e., clip-art style) in a geometrically consistent approach. The composited image and its pseudo-groundtruth from off-the-shelf methods (e.g., segmentation, keypoints, shapes) are utilized to train a model that can produce accurate 2D/3D object reconstruction under severe occlusions.

1.2 Findings

Our method demonstrates significant improvements in both 2D and 3D reconstruction, particularly in scenarios with heavily occluded objects like vehicles and people in urban scenes. Through extensive experiments, we demonstrate the effectiveness of our data in both vehicle and human reconstruction, particularly in scenarios with heavy occlusions. It is important to note that our method does not require any human labeling and hence is easily scalable and serves as an effective method to automatically generate realistic training data for reconstructing dynamic objects under occlusion.

1.3 Conclusions

Our data generation framework is method-agnostic, accommodating various robust object pose estimation alternatives as well as methods for high quality human reconstructions from videos. Moreover, as our pseudo-ground truth data includes metric-scale depth information using ground plane, we can augment existing datasets to enhance the robustness of 3D multi-human reconstruction methods to occlusion. Because our method can automatically generate images with realistic occlusion configurations without human labeling, it can be used to augment datasets. This is especially valuable for training models to detect vulnerable road users because they are not often captured or labeled in publicly available datasets.

5. Publications

- Khiem Vuong, Robert Tamburo, Srinivasa G. Narasimhan. “Toward Planet-Wide Traffic Camera Calibration,” In IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024.
  - URL to Paper: https://openaccess.thecvf.com/content/WACV2024/papers/Vuong_Toward_Planet-Wide_Traffic_Camera_Calibration_WACV_2024_paper.pdf
  - Project Website: https://www.khiemvuong.com/OpenTrafficCam3D/
- Khiem Vuong, N Dinesh Reddy, Robert Tamburo, Srinivasa G. Narasimhan. “WALT3D: Generating Realistic Training Data from Time-Lapse Imagery for Reconstructing Dynamic Objects under Occlusion,” In Computer Vision and Pattern Recognition (CVPR), 2024 (Oral Presentation, Top 0.8%).
  - URL to Paper: https://openaccess.thecvf.com/content/CVPR2024/papers/Vuong_WALT3D_Generating_Realistic_Training_Data_from_Time-Lapse_Imagery_for_Reconstructing_CVPR_2024_paper.pdf
  - Project Website: https://www.cs.cmu.edu/~walt3d/

6. References

[1] Khiem Vuong, Robert Tamburo, Srinivasa G. Narasimhan. “Toward Planet-Wide Traffic Camera Calibration,” In IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024.

[2] Khiem Vuong, N Dinesh Reddy, Robert Tamburo, Srinivasa G. Narasimhan. “WALT3D: Generating Realistic Training Data from Time-Lapse Imagery for Reconstructing Dynamic Objects under Occlusion,” In Computer Vision and Pattern Recognition (CVPR), 2024.