Connected Vision for Increased Pedestrian Safety (CVIPS)
Data Management Plan

Data Collection

What data will you collect or create?

(1) **Dataset**: We will be creating (using the high-fidelity simulator CARLA) images and short image sequences of urban intersections containing vehicles, pedestrians and other vulnerable road users (VRUs). All the images and image sequences generated will be annotated, providing important information such as 3D bounding boxes and trajectories for objects of interest in these scenes. This dataset will be used to train and evaluate machine learning algorithms to detect, track and predict the trajectory of VRUs including pedestrians.

(2) **Software**: There will be two categories of software. One is for creating synthetic images and image sequences using CARLA. The other is the code for training and testing selected machine learning algorithms for detecting, tracking and trajectory prediction of pedestrians and other VRUs.

(3) Model: Deep learning models for VRU detection, tracking and trajectory estimation will be developed using the CVIPS dataset.

(4) Publications & Reports: Initially, we will develop reports discussing developed algorithms and numerical results on VRU detection, tracking and trajectory prediction. Later, we plan to publish our research findings in conference and journal venues focused on connected vision for transportation and safety. We will freely distribute these publications in PDF format, via the research group's web page.

How will the data be collected or created?

The CVIPS dataset will be created using the high-fidelity CARLA simulator, simulating cameras installed on multiple vehicles and infrastructure with various VRUs in different environmental conditions in the scene. The simulator allows for annotating of each object of interest in the scene.

Deep learning (DL) models will be developed by training and testing several DL models on the CVIPS dataset and selecting those with acceptable accuracy and computational complexity tradeoffs.

After validation, the code for the selected DL models will be made available in appropriate version-controlled repositories such as *Github*. The numerical experimental results and corresponding reports and conference/journal publications will be shared via the research group's web page.

Documentation and Metadata

What documentation and metadata will accompany the data?

The documentation will include details on the methodology for CVIPS dataset creation, the deep learning (DL) model's implementation, instructions on executing inference with the DL model, and steps to replicate the presented results. For the dataset, we will utilize universally accepted data formats. Videos from all cameras will be collected in a compressed format, accompanied by their camera intrinsic and extrinsic transformation parameters. The data will be organized based on various criteria such as the number of agents involved, the nature of the road intersection, and prevailing weather conditions. Details of the data-creation process will be carefully documented.

Ethics and Legal Compliance

In the initial phase, we anticipate no ethical concerns with our research data since all the data we will be creating and using will be simulated using CARLA. Before we transition to using real-world datasets, we will engage with the university's internal review board (IRB) to obtain appropriate approvals. Any personally identifiable information in such real datasets will be anonymized to ensure privacy.

Storage and backup

How will the data be stored and backed up during the research?

The dataset will be stored on a server in the research group's laboratory. Additionally, the training dataset will be backed up on Google Drive for redundancy and to minimize the risk of data loss.

How will you manage access and security?

Both the research group lab server and Google Drive allow secure access. Only active project researchers will have permission to access the data. Any data gathered in the field will be safely relayed to our data server using recognized secure data transfer methods.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

Data created for training, validation, and testing of algorithms and the developed code and results will be retained for at least 10 years as they provide a baseline for subsequent research to improve the detection, tracking and trajectory estimation of pedestrians and other VRUs in urban intersections. The dataset can be used in future research on collaborative vision approaches for VRU safety. This will benefit researcher areas in traffic safety, autonomous vehicle, and smart cities.

What is the long-term preservation plan for the dataset?
The dataset will be archived and stored in the research group lab server. The code will be stored in standard version control tools –*Github* with a link to the pre-trained model. The technical report and research papers will be stored on *arXiv*. Published papers will also be made available in line with the specific guidelines of the publishers.

Data Sharing

What resources will you require to deliver your plan?
The dataset, dataset generation code, algorithm implementation code, trained model, and technical reports will be posted on code and paper repositories.

Are any restrictions on data sharing required?
We do not expect to require any significant restrictions on data sharing for the data created using CARLA. We might delay posting some technical reports or simulation codes until the corresponding technical papers are published.

Responsibility and Resources

Who will be responsible for data management?
The PI, Vijayakumar Bhagavatula, will have overall responsibility for data management.

What resources will you require to deliver your plan?
We don't expect to need any specialized expertise or extraordinary hardware/software. We plan to use the data storage resources available in the group lab.