2023 IEEE Symposium Series on Computational Intelligence (SSCI)
Mexico City, Mexico. December 5-8, 2023

Protecting Vulnerable Road Users: Semantic Video Analysis for Accident Prediction

Julian Petzold*, Mostafa Wahby*, Youssef Ziad*, Mostafa ElSheikh*,
Ahmed Dawood*, Mladen Berekovic*, Heiko Hamann†
*Institute of Computer Engineering
University of Lübeck, Lübeck, Germany
Email: petzold@iti.uni-luebeck.de
†Department of Computer and Information Science
University of Konstanz, Konstanz, Germany

***Abstract*—Pedestrians and cyclists are some of the most vulnerable, but also least predictable traffic participants. Due to their ability to move in urban environments with high degrees of freedom and sudden changes of direction, their movement is still challenging to predict. We present a driver assistance system that tackles some of these challenges. Our system consists of a world model made of a variational autoencoder and a long short-term memory network. The world model takes vision and action data from the perspective of the vulnerable traffic participant and generates a visual prediction (image) of their environment up to one second in advance. The second part of our system is a transformer-based description system that takes the predicted perceptions and here, as a showcase, abstracts them down to a textual warning if a collision between car and vulnerable traffic participant seems imminent. Our description system helps contextualize the dangerous situation for the driver and could be extended to other driver assistance systems, such as blind spot detection. We evaluate our system on a dataset generated in simulations using CARLA.**

***Index Terms*—autonomous driving, machine learning, video description, world models**

I. Introduction

Due to the continuously increasing demand for automobiles over the years, road traffic injuries became one of the leading causes of death. According to the World Health Organization's report on road safety in 2018, around 20-50 million people experience non-fatal injuries and 1.35 million people die each year, globally [1]. Introducing roundabouts has been effective in reducing the frequency and severity of accidents in comparison to traditional stop-controlled intersections. According to a study by the Federal Highway Administration, they resulted in a 90% reduction in fatal accidents, a 75% reduction in non-fatal ones, and a 35% reduction overall [2]. However, they led to higher risk than intersections for cyclists due to narrower lanes and stronger curves. A cyclist is 1.4 times more likely to be involved in an accident that leads to an injury in a roundabout [3]. Therefore, developing methods to improve traffic safety became an immediate necessity.

The recent introduction of automated driver assistance system (ADAS) functions, such as adaptive cruise control, lane keeping assist, automatic emergency braking, blind spot detection, etc. have proven to improve road safety [4]. Still, the development of further ADAS functions that consider the safety of vulnerable road users (VRUs), i.e., pedestrians and cyclists, could significantly reduce the casualty numbers mentioned above. In urban areas, VRUs are heavily involved with vehicles in traffic, which can not be properly detected and tracked using the current radar-based systems [5]. However, the recent success of vision-based machine learning solutions indicates their enormous potential for such ADAS functions. Predicting VRU trajectories and behavior is crucial for ensuring their safety as well as vehicle passengers. If an autonomous vehicle can accurately predict the movements of pedestrians, it can take appropriate actions to avoid collisions and ensure safe and smooth driving.

This paper builds upon previous work, where vision-based prediction models were trained to anticipate the behavior of VRUs in a simulated unsignalized pedestrian crossing scenario [6]. In this work, we train prediction models in more dangerous traffic situations. More specifically, we design two scenarios. The first one involves a pedestrian that crosses a signalized 4-way crossing. In the second one, a cyclist enters a roundabout, then exits at the third-next exit. The two situations are considered the most dangerous situations for both pedestrians and cyclists. We collect synthetic data from the perspective of the pedestrian and cyclist. We use the data to train convolutional variational autoencoder (VAE) and long short-term memory (LSTM) networks. The networks can be used to build an ADAS that predicts the traffic and signals a warning in case of predicted dangerous situations.

Finally, we train a video captioning network which provides a base for semantic processing of traffic situations as a showcase. This semantic processing base could be used as input for various ADAS functions and output types, such as an alarm signal for impending collisions, an audio-based lane keeping system, or a textual description of the traffic situation. In our case study, the network textually describes predicted footage and raises warnings in case of anticipated accidents.

II. Related Work

Traffic prediction approaches can be broadly categorized into three perspectives: vehicle, pedestrian, and road infras-

Partially funded by the Federal Ministry of Education and Research (BMBF), project 'NEUPA', grant 01IS19078, and project 'KI-IoT', grant 16ME0091K.

tructure. Many of these approaches rely on datasets that include images, as well as prior information about the actions and environment of the pedestrians being tracked [7]–[9]. These priors provide additional context that can help to account for feature relevance and enhance prediction performance. Here, we present relevant approaches from each category, as well as video captioning.

A. Vehicle perspective-based predictions

Vehicle perspective-based prediction approaches typically use monocular RGB images as input, where autoencoders are used to convert such images into a lower-dimensional representation to improve processing efficiency [10]. For instance, Hoy et al. [11] implemented an autoencoder-based approach to track objects in the Daimler Pedestrian Path Prediction Dataset [12] and generated a binary classification of pedestrian crossings/stops. Poibrenski et al. [13], [14] proposed a multimodal approach to trajectory prediction that involves feeding past trajectories and pedestrian scales into a conditional autoencoder with a Recurrent Neural Network (RNN) architecture. Makansi et al. employed mixture density networks to anticipate the behavior of pedestrians in traffic [15]. To achieve this, they utilized semantic segmentation data to establish a reliability prior, which allowed them to identify all potential future locations for a particular object class. They used this information to account for the movements of the ego-vehicle and make predictions about the future positions of pedestrians. Mangalam et al. segmented the pedestrian motion and pose prediction task into two distinct components, namely local and global motion [16]. To tackle these individual subproblems, they utilized an RNN that employs a recurrent encoder-decoder architecture. Yin et al. [17] utilized a transformer network, which is another state-of-the-art encoder-decoder architecture example, to combine various inputs such as ego-vehicle speed, optical flow, and previous pedestrian trajectories to forecast trajectories.

B. Pedestrian perspective-based predictions

Egocentric pedestrian trajectory prediction, where the trajectory is predicted from the first-person point of view, is a challenging task due to the highly dynamic nature of the environment and the limited field of view of the camera. Therefore, researchers proposed incorporating additional sensor data, such as inertial measurement units (IMUs), to improve the accuracy of the predictions. For example, Park et al. [18] introduced an EgoRetinal map representing the surrounding indoors environment, encoding occlusion likelihood, depth information, and semantics. Using a convolutional neural network (CNN), they generate a set of credible trajectories. To confirm occluded spaces, they assess the frequency with which the predicted trajectories are in proximity of the occluded regions. Also, using a person's past locations, body poses, and first-person camera images, Qiu et al. [19] proposed an encoder-decoder framework based on LSTM to predict future trajectories. In traffic scenarios (e.g., road crossing), Petzold et al. [6] conducted a study where they gathered data from the pedestrian viewpoint in simulations to train ANNs. The study utilized a synthetic environment generated through the CARLA traffic simulator [20]. The researchers trained VAE and LSTM networks to predict the positions and trajectories of VRUs in the immediate future, up to one second ahead.

C. Infrastructure-based predictions

Alternative methods for traffic prediction do not rely on either the vehicle's or the pedestrian's perspectives, but rather on infrastructure-based sensors. For example, Zhao et al. [21] introduced a pedestrian tracking system that employs roadside LIDAR data as input to a Deep Autoencoder Neural Network. At road intersections, LIDAR sensors were set up to gather information on pedestrians, including their presence, position, speed, and direction. Also, Sun et al. [22] used an external SLAM system and suggested the T-Pose-LSTM, which enables real-time 2D predictions of pedestrian trajectories. Other approaches used 2D maps for readily available training data [23]. For example, Zhang et al. [24] utilized a standard LSTM to predict pedestrian jaywalking based on video data from a camera placed at a crosswalk. They transformed the perspective of the video data to a 2D map representation and incorporated various factors such as location, traffic light state, and social factors to make predictions. Similar to the previous approach, Vasquez et al. [25] used 2D maps and implemented Inverse Reinforcement Learning (IRL) to safely navigate a mobile robot through pedestrian crowds. Also, Fahad et al. [26] used IRL to generate authentic pedestrian trajectories with social interactions on a 2D navigation grid.

D. Video captioning

Given the increasing importance of video content in our daily lives, video captioning networks have become an active area of research and development, with numerous advanced models being developed and tested to improve their accuracy and performance. For example, Ging et al. [27] presented a cooperative hierarchical transformer architecture (COOT) that incorporates long-range temporal context in a cross-level manner. The authors employed two novel components to model interactions within and between hierarchy levels, specifically an attention-aware feature aggregation module for modeling frame and word interactions, and a contextual transformer for modeling local and global context interactions. Furthermore, they introduced a cross-modal cycle-consistency loss to ensure semantic alignment between clips and sentences. SwinBERT [28] is another example that is composed of two modules, namely the *VidSwin Transformer* and the *Multi-modal Transformer Encoder*. The *VidSwin Transformer* takes in raw video frames as input and produces a spatio-temporal representation of the video as a sequence. This sequence is then utilized as an input for the *Multi-modal Transformer Encoder*, which transforms it into a natural language description. This approach differs from previous research by incorporating a built-in generator for spatial-temporal representations within the transformer architecture. As a result, it can learn using variable numbers of video

tokens and end-to-end training, eliminating the need for offline-extracted video features. The model also uses BERT Transformer [29] for natural language generation, making it a fully transformer-based architecture.

III. Methods

Our contributions can be separated into three components: a realistic simulation scenario with fully controllable pedestrians containing roundabouts and traffic lights in CARLA [20], a prediction model [30] for cyclist vision in traffic, and a captioning and data collection method for swinBERT (see Section II-D) that transforms it into a driver assistance system for collision prediction.

In our approach, we collect action and vision data for pedestrians and cyclists in CARLA. Using this data, we train a VAE-LSTM vision prediction model [6] to generate predicted vision data. This in turn is the input to our textual description model, which interprets a sequence of actual and predicted vision data. If our description model predicts an anomalous scene, it warns the driver of a potential accident.

A. CARLA-based traffic scenarios

Our data generation and collection methods are based on [6]. We designed a map that resembles a German town, as, for our case, this ensures realism and facilitates a later transfer of our pedestrian perception prediction model to real-world data. The map consists of a 4-way intersection and a roundabout, providing relevant scenarios for collecting pedestrian and cyclist data. To generate traffic scenario data, the map is populated by a specific traffic participant whose perspective we use to collect data and multiple other pedestrians and vehicles (see III-B). Depending on their mode of locomotion, the data-collecting traffic participant is either called an ego-pedestrian or an ego-cyclist. The ego-pedestrian's movement is controlled by two finite state machines (FSMs). The other traffic participants, including the ego-cyclist, are controlled by CARLA's AI systems. The roundabout has 5 exits, 2 car lanes and a cyclist lane, which we modeled in CARLA as a modified car lane. Cars are only spawned on car lanes and cyclists are only spawned on the cyclist lane. To prevent cars from merging onto the cyclist lane, we have disabled automatic lane switching.

The 4-way intersection contains three traffic lights with realistic light patterns. CARLA does not provide pedestrian behaviors that are compatible with traffic lights out of the box. Therefore, we implemented the waiting behavior of pedestrians at intersections by disabling their AI controllers whenever they approach a red traffic light. This is determined by the orientation and position of the pedestrian. The ego-pedestrian is controlled by two FSMs. One FSM controls the body movements and the other FSM controls the head. We define the ego-pedestrian's starting position. The ego-pedestrian then moves along the sidewalk, until they encounter the traffic light. They turn right by 90° and check the status of the traffic light. When the traffic light is green, the ego-pedestrian looks left to check if any vehicle is approaching. If the intersection is clear, the ego-pedestrian crosses the street. While the light is red, the body FSM stays in a waiting state (*look*) and the ego-pedestrian stands still in front of the traffic light, mimicking a waiting behavior similar to the pedestrians controlled by CARLA. After the traffic light the ego-pedestrian follows the sidewalk. See Fig. 1 for the ego-pedestrian's body FSM and see Fig. 2 for the ego-pedestrian's path.

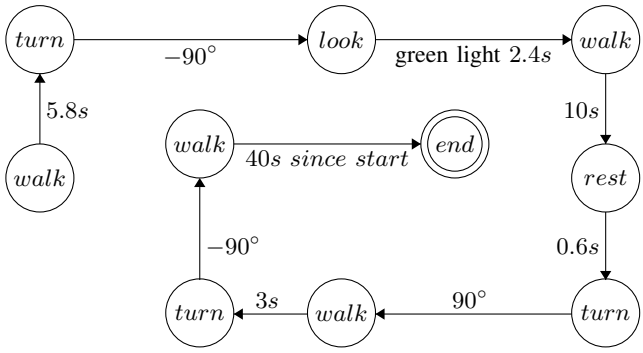


Fig. 1: A finite state machine describing the ego-pedestrian's movement in the recorded traffic light scenarios. Turn angles are given in clockwise direction.

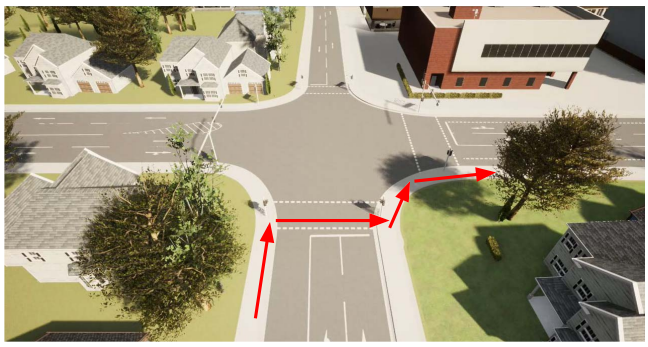


Fig. 2: A top-down view of the ego-pedestrian's path through the traffic light environment in CARLA.

B. Data collection

To collect data for our prediction models, we generate 1000 episodes with 1000 time steps each in both scenarios. The first scenario is a cyclist traversing a roundabout and the second scenario is a pedestrian crossing a traffic light. We sample the number of traffic participants populating each episode from a uniform random distribution. This guarantees varied but realistic scenarios. For the cyclist episodes, we sample a uniform distribution unif$\{70, 100\}$ for the number of cars, unif$\{150, 220\}$ for pedestrians and unif$\{25, 35\}$ for other cyclists. The same is done for cars (unif$\{20, 60\}$) and pedestrians (unif$\{150, 300\}$) in the traffic light variant. Starting locations and destinations for the AI traffic participants are randomly generated within within the bounds of expected traffic behavior. Vehicles are placed on the street and pedestrians are placed on the sidewalk.

To collect data in each time step t, we attach a camera to the head joint of the ego-pedestrian and -cyclist, capturing

their view of the surroundings. We use this camera to collect semantic segmentation data using CARLA's own semantic segmentation camera [20] and RGB data. Both RGB and semantic segmentation images have a resolution of 450×850. The semantic segmentation can identify 26 types of objects. These include the 23 base classes from CARLA, a class for crosswalks (present at the entrances and exits of the roundabout) and two classes for traffic lights that encode the state of the traffic light for the approaching pedestrian (red or green). To obtain the traffic light state in semantic segmentation, we match the RGB and segmentation images pixelwise. To determine the traffic light state, we use the average color of all RGB pixels corresponding to pixels of the traffic light class in the segmented image. Additionally, we save the ego-agent's action vector a_t in each time step. For the ego-cyclist, a_t consists of the cyclist's speed, steering angle (direction of movement) and yaw angle (yaw angle of the camera). For the ego-pedestrian, we save movement speed, body rotation (direction of movement) and head rotation (yaw angle of the camera).

In the roundabout episodes we simulate a cyclist entering the roundabout from the south-most entrance and exiting it at the north-most exit, as shown in Fig. 3. As the ego-cyclist navigates the roundabout, the other AI-controlled vehicles and cyclists also navigate the roundabout, entering and exiting at randomized times and streets. In multi-lane roundabouts in Germany, the outermost lanes have the right of way. Therefore, a car exiting the roundabout has to wait for passing cyclists on the outermost lane. Since CARLA does not support this behavior out of the box, we implemented it post-hoc and discarded any recorded episodes in which cars cut off the cyclist. We do not count the discarded scenarios against the size of our dataset. An example frame of the ego-cyclist's perceptions is shown in Fig. 4.

In the traffic light episodes, the ego-pedestrian approaches the intersection from the south on the left side of the street (see Fig. 2). They cross a traffic light to their right and then continue eastward on the sidewalk. During the episode, other vehicles and pedestrians also traverse the intersection with random start and target points. Before we spawn all traffic participants and start to collect data, we run the scenario for a random number of ticks (unif$\{0, 1600\}$). This ensures that the traffic lights are in a random state at the beginning of each episode.

Due to constraints in compute resources both during training and execution of our models, we resize all images down from 450×850 pixels to 45×85 pixels before training. To avoid losing details like street markings or distant traffic lights during the resizing, we transform the semantically segmented images from their original RGB encoding to a 26-channel encoding beforehand, representing all classes present in the original image's pixels proportionally. Our dataset Ψ_p contains the traffic light scenarios with the ego-pedestrian p. Each sample $\psi_p \in \Psi_p$ contains the action vector a, an RGB image x_{RGB} and a semantically segmented image x_{sem}. Our dataset Ψ_c contains the roundabout scenarios with the ego-cyclist c. Here, each sample $\psi_c \in \Psi_c$ contains the action vector a and a semantically segmented image x_{sem}. Both datasets contain 1m samples each.

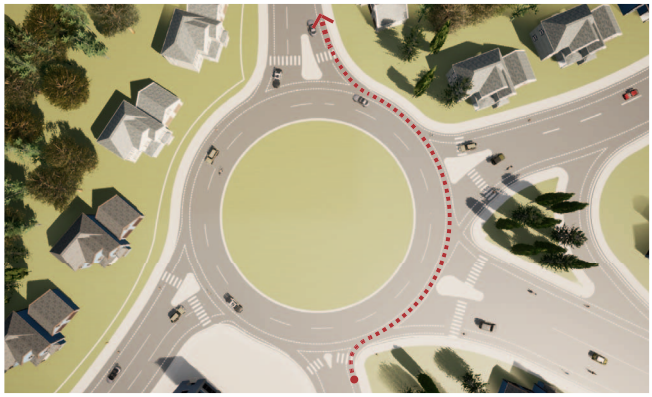


Fig. 3: A top view of the ego-cyclist's path through the roundabout environment in CARLA.

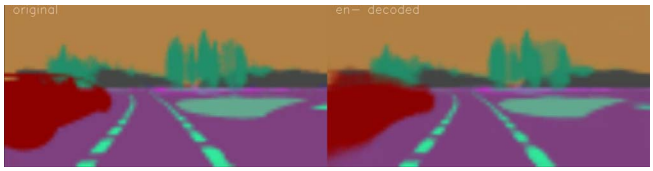


Fig. 4: The ego-cyclist's perception as captured in CARLA (down-sampled to 45×85 pixels, left) and reconstructed by our VAE (right).

C. Training VAE-LSTM prediction models

Our perception prediction models consist of a VAE and an LSTM. They predict a traffic participant's (ego-pedestrian or ego-cyclist) perception at the next time step (60 ms ahead) using images and actions as inputs. Therefore, they constitute a type of world model [30]. For the pedestrian prediction model, the VAE encodes the semantically segmented image x_t captured by a pedestrian at time step t and compresses it into its latent vector representation z_t. The LSTM takes latent vector z_t and action a_t to predict latent vector z_{t+1} at the next time step. Since LSTMs have limited memory capacity [31], the LSTM does not operate directly on high-dimensional image inputs. To reach a desired time horizon for predictions, the LSTM is fed with its own output (and a new action vector $a_{t+1} \in \Psi_p$) repeatedly. The cyclist prediction model operates in the same way.

We create the pedestrian perception model using dataset Ψ_p and the cyclist prediction model using Ψ_c. Both models have the same architecture. Their VAEs consist of 4 convolutional and 4 deconvolutional layers. Their LSTMs consist of a single layer with 512 memory cells. We split both datasets into 86% training data, 10% validation data and 4% test data each. The VAEs are trained on the semantically segmented images of each sample $\psi \in \Psi$ individually, while the LSTMs are trained on sequences of 1000 consecutive samples with each sequence corresponding to one collected episode. The samples for LSTM training consist of the latent vector z

generated by the VAE and the collected action vector a.
During VAE training we use the Kullback–Leibler (KL) divergence [32] not only as regularization loss, but also as reconstruction loss. The pixels in our semantically segmented images represent classes like "sidewalk" or "traffic light (red)". They are categorical variables rather than real-valued variables. For this reason, we did not chose a more typical image reconstruction loss like $L2$ loss. The KL divergence is able to map the different classes to each other without inferring similarities between classes that do not exist. The VAEs were trained for 150 epochs on batches with size 2000. After training we picked the models with the lowest validation loss to prevent overfitting.
To make our model robust against randomness and uncertainty in the environment, we make our LSTM a mixture density network (MDN), similar to Ha and Schmidhuber [30]. An MDN does not produce fixed output values, but a probability density function $p(z)$ containing a mixture of Gaussian distributions. To obtain a prediction for the next time step, we sample from $p(z)$. The level of randomness is controlled by the temperature variable $\tau \in [0, 1]$. If $\tau \approx 0$, $p(z)$ returns the median of the distributions and thus behaves deterministically. A high value of τ makes the prediction task harder, as it introduces uncertainty. During training, this may lead to a more robust prediction network. We train the LSTMs with $\tau = 1 \times 10^{-8}$ for 1.8×10^6 steps on batches with size 2000. We chose the networks with minimum validation loss to prevent overfitting.

D. Training a video description model

We use the video description model swinBERT [28] in our ADAS, as it confers multiple advantages over using a simple binary classifier that only detects an impending collision or not. Using a transformer-based approach may introduce overhead and in the presented use case (see Sec. IV) we only need a binary signal in the end, but swinBERT supports the generalization to other features of traffic. For example, it may easily be extended to cover traffic in residential areas. Furthermore, a video description model may be more robust regarding noise and it is more expressive than a binary classifier. It could cover different types of accidents and express uncertainty. While we use a textual description and post-process it into a binary collision/no-collision signal, the output could also be an alarm signal, a numeric danger estimator or spoken descriptions. Furthermore, the semantic information of the model could also be used for other ADAS use cases, such as a parking assitant or blind spot detection.

We used swinBERT trained it on our dataset Φ_{desc}. This dataset was collected in a similar way to Φ_p and Φ_c. It consists of 100 semantic segmentation frame sequences of the traffic light episodes and 200 semantic segmentation frame sequences of the roundabout episodes. Most of the sequences show normal and safe traffic situations similar to Φ_p and Φ_c, but 37 pedestrian episodes and 64 cyclist episodes consist of dangerous situations between cars and pedestrians or cyclists. These situations are forced by making the AI-controlled cars ignore VRUs. For the traffic light episodes, the dangerous situations consist of cars making a right turn and thus approaching the traffic light and ego-pedestrian from the left. In the roundabout episodes, cars exiting the roundabout cut off the cyclist or collide with them. Using these episodes, we aim to teach swinBERT how to identify if an accident occurs. We retrain swinBERT by generating videos out of these episodes. To generate ground truth captions we evaluated the action vector a and manually classified the episode as "accident" or not. Furthermore, we manually classified the environment of each episode according to the presence of other traffic participants. We only use one third of an episode per video, equaling 13 s of footage. By using shorter videos, we increase the probability that the captions describe a collision event present in the video, as the event is always relatively short (1 to 3 seconds). After removing anomalous videos, this results in 847 captioned videos.

The captions for the pedestrian episodes are generated in the format shown in Table I. The alternative and optional modes in the sentence structure provide a rich semantic context to the classification and enable an easy extension into other ADAS applications, such as movement prediction for other cars. We used the split for training (86%), validation (10%) and test data (4%) we also used for our prediction model. After training, we chose the network with minimum validation loss to prevent overfitting.

IV. Results and Discussion

We evaluate our ADAS in four experiments. The first two experiments aim at evaluating the text description model's capabilities when using synthetic data taken directly from CARLA. We conduct one experiment based on our traffic light episodes and one experiment based on the roundabout episodes. The second set of experiments uses the same episodes as the first two experiments, but the input data is generated by our VAE-LSTM prediction model instead of CARLA. The experiments can be found in our video.[1]

A. Evaluating our traffic description model on CARLA data

In the first two experiments, we evaluate the performance of our text description model trained on our dataset Φ_{desc} collected in CARLA, while focusing specifically on the accident detection rate. After training, our model has reached 88% accuracy according to the BLEU-4 metric.

For the traffic light episodes with the pedestrian, we collected a total of 37 accident episodes. These episodes were split into training set (20 accidents), evaluation set (8 accidents) and test set (9 accidents). As the first two sets were used in training, we examine our model's performance using only the test set. To evaluate the accident detection performance of our description model, we count an episode as a detected accident if the description contains the word "dangerously". Our model captioned no scenario with the word "dangerously" that did not contain an accident. Therefore, the model's false positive rate is 0. Out of the 9 accidents in the test set, our model detected 3. These accidents are

[1] https://vimeo.com/854150737

separable into three categories: accidents involving the ego-pedestrian, “close calls” that caused the pathing of another pedestrian to react to the presence of a car, and collisions between other pedestrians and a car. The model detects all 3 accidents that involve explicit collisions between other pedestrians and vehicles, but it does not comment on the direction from which the vehicle is approaching. However, the model does not detect 3 close calls, in which a pedestrian stops in their tracks right before the collision and attempts to reroute around the car. Moreover, the description model also does not detect any accidents or close calls involving the ego-pedestrian. We conclude that our description model is sensitive to visible contact between a car and a pedestrian, but is not able to interpret the movement changes of a pedestrian as an accident. Ego-pedestrian collisions are challenging for the description model, because they are often hardly visible on the video input data.

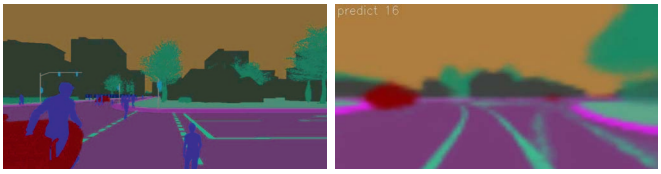
We collected 64 accidents in the roundabout episodes with the cyclist. 55 accident episodes were used in the training set, 4 in the validation set and 5 accidents were evaluated in the test set. Since we used CARLA’s vehicle motion system for the ego-cyclist, all accidents in the roundabout set are close calls. This means that the ego-cyclist braked and no collision occurred. The video description model detected all 5 accidents in the test set and had no false positives. Hence, we assume that our video description model is more sensitive to camera movement in the roundabout episodes than in the traffic light episodes. For 3 out of 5 accidents, the description model was also able to detect if the car approached from the left or from the right.

B. *Evaluating our traffic description model on predicted data*

In the second set of experiments, we applied the methodology described in IV-A to outputs of our prediction models. We took the set of evaluation data already used in IV-A and iteratively applied our prediction models, generating predictions with a lookahead of 1 s. We then generated textual descriptions for the predictions.
The video description model shows comparable performance on recorded and predicted pedestrian data. The model detected 3 out of 9 pedestrian accidents. It recognized explicit collisions, but could not identify ”close calls” or collisions with the ego-pedestrian. The evaluation also yielded one false positive out of 24 non-accident videos. We assume that this false positive occurred due to a prediction artifact that caused a sudden shift in camera perspective. Matching the accuracy on collected cyclist data, our video description model detected all accidents in the predicted cyclist videos. There were no false positives. We believe that the cyclist data is not as susceptible to false positives as the pedestrian data, because the camera movement through the roundabout is smoother than the abrupt movement of the pedestrian at the traffic light.

V. Conclusion and Future Work

We have extended our ADAS toolchain from Petzold et al. [6] to not only incorporate more diverse traffic situations,



(a) ”a pedestrian walks on the sidewalk and then waits at traffic light as a car gets dangerously close.”

(b) ”a cyclist rides around the roundabout with other cars around”

Fig. 5: Two example captions. We generated caption (a) from a recorded traffic light video and caption (b) from a video generated by our roundabout prediction model.

but also extended its functionality to provide an interface between prediction and driver. We have trained a world model that models a pedestrian navigating through a signaled intersection and we have a trained a world model representing a cyclist traversing a roundabout. While we require multiple world models for different traffic situations, selecting the correct model requires only a classification of the observed traffic environment. We have shown that our world models do not only predict regular traffic situations, but are also capable of representing dangerous situations, such as near-misses or collisions. We have introduced a unified transformer-based traffic description model that is capable of differentiating between accidents and safe traffic scenarios in both environments. The semantically rich output of the traffic description model could be used as a base in a wide range of different ADAS systems. In future work, we intend to extend this model to express the degree of uncertainty regarding the prediction and description accuracy. Furthermore, we would like to develop the missing components in our toolchain and close the loop between car and driver. At this time we are able to display warnings about an upcoming traffic situation to a driver, but the input data for our world models is not obtainable yet outside of a simulation. To remedy this, we have to transform the perspective of a car-mounted camera to a VRU’s perspective and we have to generate the VRU’s most probable actions based on information about the environment.

References

[1] World Health Organization et al. Global status report on road safety 2018: Summary (no. who/nmh/nvi/18.20). *World Health Organization*, 2018.

[2] Bruce W Robinson, Lee Rodegerdts, Wade Scarborough, Wayne Kittelson, Rod Troutbeck, Werner Brilon, L Bondizio, Ken Courage, Michael Kyte, and John Mason. Roundabouts: An informational guide. Technical report, United States. Federal Highway Administration, 2000.

[3] C. Schoon, D. de Waard, and F. D. Bijleveld. The safety effects of converting intersections to roundabouts: A review of evidence from the literature. *Accident Analysis & Prevention*, 86:141–146, 2016.

[4] Eunbi Jeong and Cheol Oh. Evaluating the effectiveness of active vehicle safety systems. *Accident Analysis & Prevention*, 100:85–96, 2017.

[5] Shunqiao Sun, Athina P Petropulu, and H Vincent Poor. Mimo radar for advanced driver-assistance systems and autonomous driving: Advantages and challenges. *IEEE Signal Processing Magazine*, 37(4):98–117, 2020.

TABLE I: Construction of captions for normal and dangerous traffic episodes. Optional phrases shown in square brackets.

Classification	Actor and action	Environment	Description of (no) danger
No accident	The ego-pedestrian crosses the street	[with other pedestrians around]	as a car approaches.
			as a car passes from the left.
			as a car passes from the right.
	A cyclist rides around the roundabout	[with other cars around]	.
Accident	The ego-pedestrian crosses the street	[with other pedestrians around]	as a car gets dangerously close to the [ego-]pedestrian.
	A cyclist rides around the roundabout	[with other cars around]	as a car gets dangerously close [from the left/right].

[6] Julian Petzold, Mostafa Wahby, Franek Stark, Ulrich Behrje, and Heiko Hamann. "if you could see me through my eyes": Predicting pedestrian perception. In *2022 8th International Conference on Control, Automation and Robotics (ICCAR)*, pages 184–190. IEEE, 2022.

[7] Iuliia Kotseruba, Amir Rasouli, and John K Tsotsos. Joint attention in autonomous driving (JAAD). *arXiv preprint arXiv:1609.04741*, 2016.

[8] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *European conference on computer vision*, pages 549–565. Springer, 2016.

[9] Amir Rasouli, Iuliia Kotseruba, Toni Kunic, and John K Tsotsos. Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6262–6271, 2019.

[10] Dor Bank, Noam Koenigstein, and Raja Giryes. Autoencoders. *arXiv preprint arXiv:2003.05991*, 2020.

[11] Michael Hoy, Zhigang Tu, Kang Dang, and Justin Dauwels. Learning to predict pedestrian intention via variational tracking networks. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3132–3137. IEEE, 2018.

[12] Nicolas Schneider and Dariu M Gavrila. Pedestrian path prediction with recursive bayesian filters: A comparative study. In *German Conference on Pattern Recognition*, pages 174–183. Springer, 2013.

[13] Atanas Poibrenski, Matthias Klusch, Igor Vozniak, and Christian Müller. M2p3: multimodal multi-pedestrian path prediction by self-driving cars with egocentric vision. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, pages 190–197, 2020.

[14] Atanas Poibrenski, Matthias Klusch, Igor Vozniak, and Christian Müller. Multimodal multi-pedestrian path prediction for autonomous cars. *ACM SIGAPP Applied Computing Review*, 20(4):5–17, 2021.

[15] Osama Makansi, Ozgun Cicek, Kevin Buchicchio, and Thomas Brox. Multimodal future localization and emergence prediction for objects in egocentric view with a reachability prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4354–4363, 2020.

[16] Karttikeya Mangalam, Ehsan Adeli, Kuan-Hui Lee, Adrien Gaidon, and Juan Carlos Niebles. Disentangling human dynamics for pedestrian locomotion forecasting with noisy supervision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2784–2793, 2020.

[17] Ziyi Yin, Ruijin Liu, Zhiliang Xiong, and Zejian Yuan. Multimodal transformer networks for pedestrian trajectory prediction. In *IJCAI*, pages 1259–1265, 2021.

[18] Hyun Soo Park, Jyh-Jing Hwang, Yedong Niu, and Jianbo Shi. Egocentric future localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4697–4705, 2016.

[19] Jianing Qiu, Frank P-W Lo, Xiao Gu, Yingnan Sun, Shuo Jiang, and Benny Lo. Indoor future person localization from an egocentric wearable camera. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8586–8592. IEEE, 2021.

[20] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017.

[21] Junxuan Zhao, Hao Xu, Jianqing Wu, Yichen Zheng, and Hongchao Liu. Trajectory tracking and prediction of pedestrian's crossing intention using roadside LiDAR. *IET Intelligent Transport Systems*, 13(5):789–795, 2019.

[22] Li Sun, Zhi Yan, Sergi Molina Mellado, Marc Hanheide, and Tom Duckett. 3dof pedestrian trajectory prediction learned from long-term autonomous mobile robot deployment data. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5942–5948. IEEE, 2018.

[23] Dražen Brščić, Takayuki Kanda, Tetsushi Ikeda, and Takahiro Miyashita. Person tracking in large public spaces using 3-d range sensors. *IEEE Transactions on Human-Machine Systems*, 43(6):522–534, 2013.

[24] Shile Zhang, Mohamed Abdel-Aty, Jinghui Yuan, and Pei Li. Prediction of pedestrian crossing intentions at intersections based on long short-term memory recurrent neural network. *Transportation research record*, 2674(4):57–65, 2020.

[25] Dizan Vasquez, Billy Okal, and Kai O Arras. Inverse reinforcement learning algorithms and features for robot navigation in crowds: an experimental comparison. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1341–1346. IEEE, 2014.

[26] Muhammad Fahad, Zhuo Chen, and Yi Guo. Learning how pedestrians navigate: A deep inverse reinforcement learning approach. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 819–826. IEEE, 2018.

[27] Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. COOT: Cooperative hierarchical transformer for video-text representation learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, pages 22605–22618, Red Hook, NY, USA, 2020. Curran Associates Inc.

[28] Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. SwinBERT: End-to-end transformers with sparse attention for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17949–17958, 2022.

[29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[30] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS 2018)*, pages 2455–2467, USA, 2018. Curran Associates Inc.

[31] Jasmine Collins, Jascha Sohl-Dickstein, and David Sussillo. Capacity and trainability in recurrent neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[32] Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.