Learning to Predict Collision Risk from Simulated Video Data

Learning to Predict Collision Risk from Simulated Video Data

Tim J. Schoonbeek[1], Fabrizio J. Piva[1], Hamid R. Abdolhay[2] and Gijs Dubbelman[1]

***Abstract*— We propose an image-based collision risk prediction model and a training strategy that allows training on simulated video data and successfully generalizes to real data. By doing so, we solve the data scarcity problem of collecting and labeling real (near) collisions, which are exceptionally rare events. Domain generalization from simulated to real data is taken into account by design by decoupling the learning strategy, and using task-specific, domain-resilient intermediate representations. Specifically, we use optical flow and vehicle bounding boxes, since they are instinctively related to the task of collision risk prediction and because their simulated-to-real domain gap is significantly lower than that of camera video data, i.e., they are more domain resilient. To demonstrate our approach, we present RiskNet, a novel neural network for image-based collision risk prediction, which classifies individual frames of a video sequence of a front-facing camera as *safe* or *unsafe*. Additionally, we present two novel datasets: the simulated Prescan dataset (which we intend to make publicly available) for training and the YouTube Driving Incidents Database (YDID) for real-world testing. The performance of RiskNet, trained solely on simulated data and tested on the real-world YDID, is comparable to that of a human driver, both in accuracy (91.8% vs. 93.6%) and F1-score (0.92 vs 0.94).**

I. INTRODUCTION

Road accidents are still a major concern of both automotive industry and society as a whole, as globally 1.35 million people lose their lives in traffic every year [1]. Since the number of vehicles is only increasing [2], the need for more competent advanced driver assistance systems and specifically collision avoidance systems grows. The current industry standards in collision risk prediction are limited by their lack in predicting possible collisions under dynamic and complex movement patterns that are frequent on rural and urban roads [3], [4]. In this work, we present a deep data-driven, image-based collision risk prediction model that is designed to handle complex and versatile scenarios.

Significant research is being done on making model-based approaches capable of dealing with highly dynamic scenarios [5]–[9]. In these approaches, a physical or probabilistic representation of the surroundings of the ego-vehicle is constructed, after which the model's parameters are tuned to a dataset. Tuning to a specific dataset suffers from reduced generalization to new scenarios, often requiring updating the parameters for specific scenarios manually or via online self-supervised learning [10]. Additionally, model-based approaches often rely on sensors that are currently infeasible to install in mass-production vehicles, such as highly accurate localization systems [5], [6] or expensive high-end LiDAR sensors [9]. Instead, we rely on a low-cost and widely available sensor, i.e., a camera.

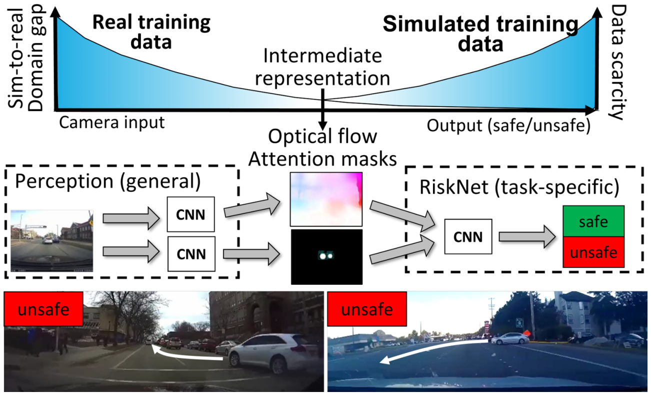

Fig. 1: We propose RiskNet, a collision risk prediction method that can be trained solely on simulated data and generalizes successfully to real data. It classifies images of video streams (bottom part of illustration) in *unsafe*, i.e., risk of collision in the upcoming 1 second, or *safe*. Our approach is based on the concept of decoupled learning (top part of illustration). The Perception module is trained separately on real-world data for tasks where data is not scarce, i.e., optical flow estimation and vehicle detection. RiskNet, for which real training data is scarce, is trained on simulated data using the intermediate representations provided by the Perception module, for which the domain gap between real and simulated data is minimal by design.

The alternative to model-based approaches are deep data-driven approaches, which have shown significant advances in the past decade [11], [17]. Instead of explicitly designing a model, a model is learned from training data. This has shown to be very robust to noisy data from low-cost sensors such as cameras [18]. Furthermore, such deep data-driven models can handle complex and versatile scenarios as long as they are sufficiently represented in the training data. If sufficient representative training data is available, deep data-driven approaches are known to significantly outperform model-based approaches. However, representative real-world vehicle collision data is scarce, and creating a new dataset is highly time-consuming, expensive, and infeasible without a large fleet of vehicles. To overcome this data scarcity, one can choose to utilize simulated training data, where numerous collision samples can be readily generated. This however introduces a sim-to-real domain gap, as it is challenging to capture the visual complexity of the real world in a simulator [19], [20].

[1] Department of Electrical Engineering at Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands. Correspondence to t.j.schoonbeek@tue.nl.

[2] Siemens Industry Software Netherlands B.V., 2595 BN Den Haag, South Holland, The Netherlands

TABLE I: Placement of our work in the existing literature.

	Data specifications				*Leveraged modalities*		
	Automotive application	Ego focused	Simulated training data	Temporal	Optical flow	Object information	Depth
DroNet [11]		✓					
FlowDroNet [12]		✓	✓		✓		
D3QN [13]		✓	✓	✓			✓
DSA-RNN [14]	✓			✓		✓	
AdaLEA [15]	✓	✓		✓		✓	
NIDB [16]	✓	✓		✓	✓	✓	
Ours	✓	✓	✓	✓	✓	✓	

In order to circumvent this aforementioned domain gap, we propose the decoupled learning strategy outlined in Fig. 1. Our approach consists of a Perception and a task-specific module, RiskNet. The Perception module is trained with real-world data for two tasks without data scarcity issues, namely object detection [21] and optical flow (OF) estimation [22]. The outputs of this module, i.e., bounding boxes and per-pixel flow maps, are *intermediate representations*. We hypothesize these intermediate representations to be more domain resilient than camera video data as they are not natural images, but representations of motion and type of object in images. This allows us to train the task-specific module, which predicts collision risk, on intermediate representations generated from easily generated simulated collision data. In doing so, we solve the data scarcity problem without having to engage in complex domain adaptation techniques to overcome the simulated-to-real domain gap [20], [23]. Our approach is described further in Section III.

In order to train and validate our work, we present two novel datasets which we intend to make publicly available at [24]. Firstly, we provide the Prescan dataset, which contains simulated (near) collisions and is used exclusively to train RiskNet. Secondly, we provide the YouTube Driving Incidents Database (YDID), which contains real (near) collisions collected on the YouTube video sharing platform. Section IV outlines both datasets in detail. We perform a quantitative and qualitative analysis of the performance of RiskNet on YDID, as further described in Section V.

To summarize, this work contains the following contributions:

- A decoupled learning strategy where the domain gap between simulated and real-world data is reduced significantly by creating multiple task-specific, domain-resilient input representations.
- The successful validation of our decoupled training strategy to collision risk prediction, which is facilitated by our RiskNet deep neural network architecture.
- Two ego-focused, image-based vehicle collision datasets: Prescan and YDID.

II. RELATED WORK

In this section, we discuss existing work on (deep) data-driven collision risk and provide a summary of the most relevant related research.

A. Data-Driven Collision Risk Prediction

Deep learning approaches are popular for many computer vision tasks, such as object detection and semantic segmentation [17], [21]. Recently, several works have also successfully shown its application in collision prediction [4], [11], [14]–[16]. An overview of the most relevant implementations using front-facing cameras is shown in Table I.

Other works have successfully leveraged optical flow [12] or disparity maps [13] to significantly reduce the sim-to-real performance gap. However, the applicability of both works is limited, as they evaluate relatively simple scenarios where collisions with (often static) objects within only a few meters are predicted. However, the noise in optical flow and depth estimation significantly increases with distance, which prevents these approaches from being used in the automotive domain, risk often exists at much greater distances. In order to handle noisy inputs, we propose not only to extract optical flow or depth, but to also extract attention masks based on vehicle detections, to direct our models attention.

The automotive approaches outlined in Table I are based on real-world datasets, which they, together with their trained models, keep private. This prevents a direct comparison of our approach with theirs. AdaLEA [15] and NIDB [16] focus on ego-vehicle collisions and are trained and evaluated on a near-miss incident database, captured with vehicle-mounted driving recorders on more than 100 taxis over a period of 10 years. In total, they collected less than 5 near-miss events per taxi per year, showing the extremity of the data scarcity for the task of prediction collision risk. Both AdaLEA and NIDB fail to consistently classify negative samples (i.e., safe driving), as it is difficult for the neural networks to distinguish safe from unsafe behavior.

B. Intermediate representations for Domain Generalization

A critical problem in supervised deep learning methods is the assumption of the same distribution between the train and test data. In reality, a domain gap exists between two datasets due to many possible differences in the distribution of their data, such as weather, object textures, or traffic density. When the training data is based on simulation but the testing data is real, the so called simulated-to-real, or sim-to-real, domain gap is especially large for camera data as it is difficult to capture the complexity of the natural environment and render them photo-realistically.

FlowDroNet [12] and Xie et al. [13] have successfully shown good generalization from simulated to real-world data by using optical flow or depth as a task-specific input representation. FlowDroNet is closest to our work as it includes non-static objects and supervised learning, therefore we focus on comparing to their work and leave depth as intermediate modality for future work. Flow inherently generalizes better to unseen data compared to camera video data due to its decreased visual complexity, as textures, colors and illuminations do not exist in optical flow [12]. Additionally, optical flow represents motion which is inherently relevant to collision prediction. Despite all the positive traits that optical flow offers with regards to generalization, we consider that object information also plays a key role especially when dealing with complex scenarios, and for this reason we complement optical flow with attention masks derived from vehicle detections.

C. Attention Masks

Vehicle collision risk does not exist without other vehicles, therefore it makes sense to attend our model to areas of the image which contain vehicles. Collision prediction studies have successfully directed their models' attention by taking a crop of each vehicle detection and extracting a set of local features from the crops [14], [15]. However, as our features are extracted from optical flow rather than RGB, significantly fewer features can be extracted from local objects [12]. In other classification problems, research has shown that fusing attention masks in feature space rather than taking crops in the data-space shows increased performance [25]. We adopt this technique and fuse binary attention masks with optical flow early-on in feature space.

III. METHOD

The purpose of our research is to demonstrate that it is possible to train a deep neural network on simulated data for a task where real-world data is particularly scarce, i.e., vehicle collision prediction. We formulate vehicle collision risk prediction as a binary classification problem with as input X_i, a sequence of RGB camera images, and as output the predicted collision risk $\hat{y}_i$. We propose a decoupled learning strategy with a task-specific collision risk prediction model RiskNet and a Perception module, i.e.,

$$\hat{y}_i = \text{RiskNet}(\text{Perception}(X_i)). \tag{1}$$

The Perception module predicts optical flow and object detections and is trained on real data, as it does not rely on task-specific collision data. RiskNet is trained exclusively on the simulated collision dataset $\mathcal{D}_{src}$ and the entire model is tested on the real-world dataset $\mathcal{D}_{target}$. More information on the datasets is provided in Section IV.

We define collision risk $\hat{y}_i$ as

$$\hat{y}_i = (p_i^{\text{safe}}, \, p_i^{\text{unsafe}}), \tag{2}$$

where $p_i^{\text{safe}} \in [0, 1]$ is the score for the classification *safe* and $p_i^{\text{unsafe}} = 1 - p_i^{\text{safe}}$, ensured via a softmax function, the score for *unsafe*. Every sample with a p_i^{unsafe} score exceeding a threshold t_H is classified as *unsafe*.

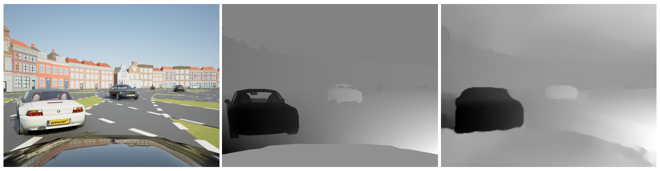

Fig. 2: Simulated camera input (left), ground truth (center) and CNN estimated (right) horizontal optical flow.

A. Decoupled learning

To solve the data scarcity problem without inducing a significant domain gap, we propose a decoupled learning approach. The first module, Perception, is defined as

$$(F_i, A_i) = \text{Perception}(X_i), \tag{3}$$

where F_i is a sequence of optical flow images and A_i a sequence of binary attention masks, which indicate vehicle detections. F_i and A_i are our proposed intermediate representations, which we hypothesize to be more domain resilient, as they are not natural images but lower dimensional representations of those images. Moreover, to train this general Perception module, we can use data that does not necessarily have to include (near) collisions. Such real-world data is readily and abundantly available [26], [27]. The second module, RiskNet, is defined as

$$\hat{y}_i = \text{RiskNet}(F_i, A_i), \tag{4}$$

and predicts the collision risk based on optical flow and attention masks as intermediate representations. We train RiskNet exclusively on simulated data because its inputs (F_i, A_i) and labels y_i can readily be generated with a simulator, whilst real-world vehicle collision data is scarce.

B. Perception module

Our Perception module consists of two neural networks, one for estimating optical flow and one for creating attention masks based on vehicle detections.

a) Optical flow: We define optical flow F_i as

$$F_i = (f_{t-(T-1)}, \cdots, f_{t-1}, f_t)_i \in \mathbb{R}^{T \times 2 \times H \times W}, \tag{5}$$

where f_t is the estimated optical flow at time t, T the total number of frames, H the height and W the width of each flow estimate. As flow consists of a horizontal and vertical component, two $H \times W$ images are provided at every t. Since optical flow is estimated based on two observed frames, we require $T + 1$ RGB images for T optical flow frames.

We experiment with training on ground truth optical flow provided in $\mathcal{D}_{src}$ (with pre-processing similarly to [12]) and training on flow estimated by LiteFlowNet [22] on the simulated camera data. Fig. 2 shows a comparison between ground truth and CNN estimated optical flow. For $\mathcal{D}_{target}$, F_i is always estimated by LiteFlowNet. We avoid training on raw optical flow vectors and instead normalize the vectors for each frame to 8-bit unsigned integers, as the raw, absolute optical flow values depend largely on the input video parameters, such as frames per second (FPS) and resolution.

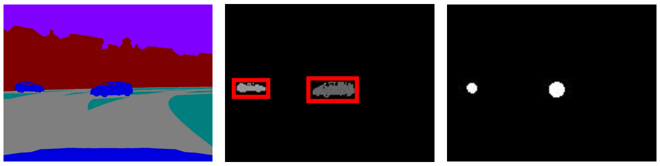

Fig. 3: Attention mask estimation in the simulated training data. Ground truth semantic segmentation (left) is used to create object detections (middle), around which attention masks are created (right).

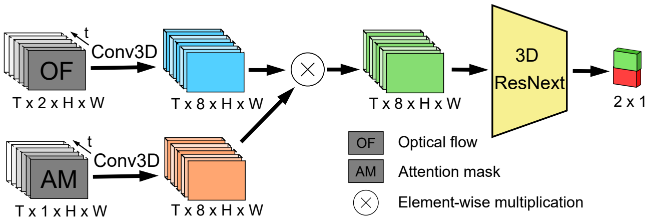

Fig. 4: Architecture of the task-specific module, showing early fusion via element-wise multiplication after a single Conv3D layer, before the 3D CNN feature extractor [30].

b) Attention masks: Let $\mathbb{B} = \{0, 1\}$ be a binary set, we define attention mask A_i as

$$A_i = (a_{t-(T-1)}, \cdots, a_{t-1}, a_t)_i \in \mathbb{B}^{T \times 1 \times H \times W}, \quad (6)$$

where a_t is the estimated attention mask mask at time t. The binary attention masks for $\mathcal{D}_{target}$ are obtained based on vehicle detections from Faster R-CNN [21]. For each vehicle, a circular mask is drawn inside it's estimated bounding box. As demonstrated in Fig. 3, we apply a similar approach for the simulated $\mathcal{D}_{src}$, but use the semantic segmentation already provided in the dataset to extract the vehicle bounding boxes. This results in overlapping cars being grouped together in a single bounding box during training, the effects of which we do not explicitly investigate. But, for future work, we recommend using per-vehicle simulated bounding boxes instead of using simulated semantic segmentation.

Since the simulated $\mathcal{D}_{src}$ provides us with highly accurate semantic segmentations and therefore near-perfect attention masks, whilst real-world detections are likely to contain errors, we add false attention masks (with a random radius) N_f times with a probability p_n to each A_i sequence. Additionally, by opting for circular rather than oval or rectangular attention masks, we keep information about the shape of the detected object to a minimum, preventing the model from overfitting on unintentional biases in the dataset.

As explained before, the real-world validation set $\mathcal{D}_{target}$ likely contains false detections, resulting in noisy attention masks. Therefore, we apply prior filtering to the attention masks with prior knowledge, similarly to [28]. We construct our prior by calculating the sum of all road and vehicle pixels from the GTA V dataset [29], normalizing the values between 0 and 1 and finally rounding all values to the closest integer. We then element-wise multiply the binary prior and attention masks during interference on $\mathcal{D}_{target}$, to e.g. suppress vehicles that are detected in the sky.

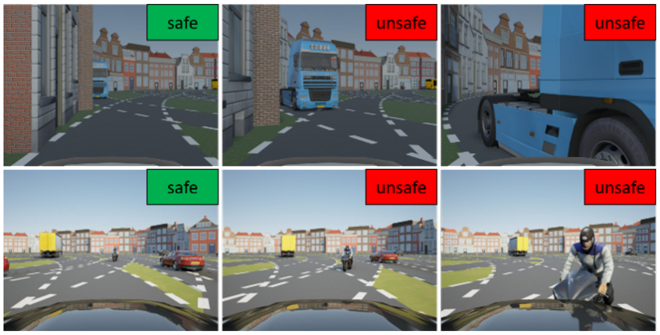

Fig. 5: Samples from the Prescan (source) dataset $\mathcal{D}_{src}$, displayed at the used sampling rate of 10 Hz.

C. RiskNet collision risk model

RiskNet, the proposed task-specific module as described by (4) and in Fig. 4, is a two-stream spatio-temporal 3D CNN. RiskNet takes estimated optical flow F_i and attention masks A_i as inputs.

The two streams of optical flow and attention masks each pass through a single 3D convolution (Conv3D) layer with 8 output channels, kernel size 3, zero padding to maintain the original input dimensions and a stride of 1. The outputs of these two Conv3D layers are fused via element-wise multiplication, which acts as a soft attention mechanism proposed in [25]. The fused features are fed into a 3D ResNext feature extractor with 18 layers [30]. Finally, a fully connected layer predicts a probability for the *safe* and *unsafe* classes respectively, which are automatically normalized.

For training we use the binary cross entropy loss, which for sample i is defined as

$$\mathcal{L}_i = -[y_i \log(p_i^{\text{safe}}) + (1 - y_i) \log(p_i^{\text{unsafe}})], \quad (7)$$

where y_i is the ground truth label at time t. More details of the hyperparameters used in training are provided in Section V.

IV. DATASETS

As there are no publicly available collision datasets that focus on ego-vehicle collisions, we compose two new datasets. The Prescan source dataset $\mathcal{D}_{src}$ contains simulated video data obtained from the Prescan simulator [31] and is used for training RiskNet. The generalization of our models is tested on $\mathcal{D}_{target}$, the YouTube Driving Incidents Database (YDID). We intend to publish $\mathcal{D}_{src}$ to the public at [24]. Each video frame in both datasets is provided with a *safe* or *unsafe* label. This section outlines the datasets and their labeling techniques.

A. Prescan

The Prescan dataset consists of nearly one hour of driving footage containing cars, motorcycles, buses and trucks, simulated with the Prescan simulator [31]. Samples from the Prescan dataset are shown in Fig. 5. The dataset contains a wide range of simple movement patterns, with all vehicles having a constant steering angle and velocity. The aim of the design of the simulated data is to have a high level

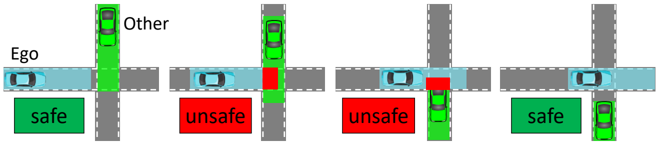

Fig. 6: *Prescan labeling criteria.* The spaces which each vehicle will occupy within the next 1.0 second are drawn for each frame. If there is an overlap in space, the situation is labeled unsafe. In a typical crossing where two vehicles pass near each other, this results in a transition from safe to unsafe to safe.

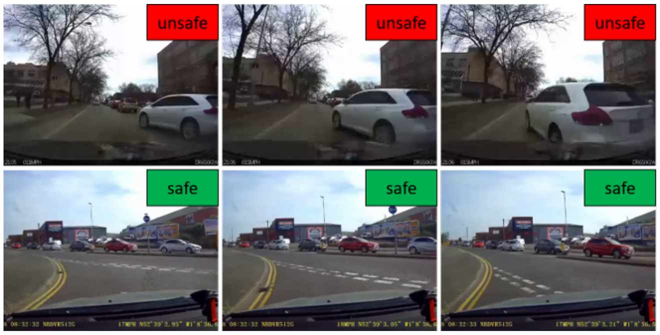

Fig. 7: Samples from $\mathcal{D}_{target}$, the YouTube Driving Incidents Database (YDID), displayed at the used sample rate of 10 Hz.

of randomness in the simulated motion patterns, such that there is a high probability that real motion patterns are a subset of the simulated motion patterns. This concept is often referred to as domain randomization [32]. In total, 189 different runs were simulated at 20 FPS with a duration up to 30 s, depending on whether a crash occurs, upon which the simulation is stopped. In total, $\mathcal{D}_{src}$ contains 58 904 *safe* and 7788 *unsafe* frames.

Labeling protocol. As illustrated in Fig. 6, a frame is labeled *unsafe* if there is any overlap in occupied space between the ego and any other vehicle within a time horizon of 1.0 s. Since all vehicles are driving with constant velocity and steering angle, this approach to labeling is similar to the time-to-event (TTE) metric widely used in the automotive risk domain [4], [33]. The labels are limited to vehicles within the ego-vehicles field of view, as we do not demand the system to estimate collision risk when an unobserved vehicle is going to collide within 1 second. Additionally, note that no distinction is made between near-miss and collision events, both events are labeled *unsafe*.

B. *YouTube Driving Incidents Database (YDID)*

The YDID contains 16 videos with a total duration 154 s, collected on the online video sharing platform YouTube. Some samples are shown in Fig. 7. The variety between the videos is significant, both with regards to the scenarios and camera positions and quality. YDID contains scenarios recorded on rural, urban and highway roads, in both left and right hand drive countries. Additionally, the lighting, camera positions and image quality differ greatly per sample. All videos are 30 FPS and cropped to the same aspect ratio as $\mathcal{D}_{src}$, resulting in an image size of 360×480 pixels (H×W).

Labeling protocol. In the real-world dataset, vehicles actively try to avoid collisions by braking and steering abruptly. Therefore, the YDID can not be labeled according to the same approach as the simulated dataset. To solve this problem, we asked 41 experienced drivers from different ages, nationalities and genders to annotate our clips. The annotators first see a clip completely, after which they are encouraged to rewind the video and indicate in which frames they believe another vehicle poses an urgent risk to the ego-vehicle. The definitive annotation for each frame in the dataset is determined by the majority voting over all 41 labelers. This approach to labeling is similar to pedestrian intention estimation datasets, such as proposed in [34], the main difference being that our drivers had future knowledge advantage. Considering this advantage, we take the human annotator consensus as an upper bound. In total, the YDID contains 3604 *safe* and 1008 *unsafe* frames. Out of the 16 videos, 13 videos contained any *unsafe* labels.

V. EXPERIMENTS

Three sets of experiments are performed on the newly presented YDID dataset, which has not been used to train the collision risk model RiskNet:

- *Comparison to state of the art.* We study the performance of our approach on a frame level and compare this to both FlowDroNet [12] and human performance.
- *Component analysis.* Here, we study the contribution of each individual proposed component via several ablation experiments and test our hypothesis that optical flow generalizes better to an unseen domain than raw RGB images.
- *Qualitative analysis.* We perform a qualitative analysis, to thoroughly understand RiskNet's performance in different scenarios.

A. *Performance Metrics and Evaluation Protocol*

Since we approach collision risk prediction as a per-frame binary classification problem, we evaluate our experiments with classification metrics. Specifically, the average precision (AP) [35] and the area under the receiver operating characteristic curve (ROC-AUC) [36] metrics are used to evaluate the performance of each model configuration. These metrics are more robust to unbalanced test sets than F1-score and accuracy [37]. Both ROC-AUC and AP evaluation metrics are inadequate to quantify human performance, as our drivers did not indicate confidence levels for their predictions. Therefore, we use F1-score and accuracy to compare our approach against human performance. For these threshold-based metrics, we found $t_H = 0.8$ most suitable, based on the ROC and AP curves.

The videos in the simulated $\mathcal{D}_{src}$ are shuffled and divided into five equal partitions. RiskNet is trained three times for each experiment, every time on a different set of four partitions (80% of $\mathcal{D}_{src}$). For each metric, the mean ± standard deviation of the best checkpoint from each of the

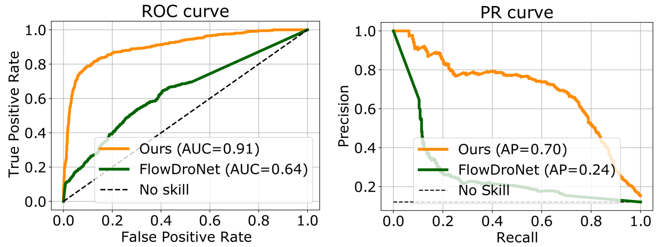

Fig. 8: Comparison of ROC curve and PR curve for our best performing model and FlowDroNet [12] on YDID.

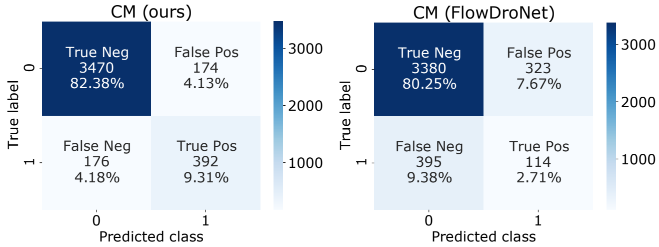

Fig. 9: Comparison of confusion matrices (CM) for our best performing model and FlowDroNet [12] on YDID.

three models is reported. We define the best checkpoint as the one with the highest AP on $\mathcal{D}_{target}$.

B. Implementation Details

In all experiments with a 3D CNN, our models take sequences of $T = 8$ frames sampled at 10 FPS as input for both training and inference. Since even collision datasets with numerous crashes tend to suffer class imbalance between the positive and negative samples, due to the period of safe driving leading up to incidents, we apply horizontal flipping and re-sampling to reduce class imbalance. Specifically, we over-sample the minority class, i.e., we take t={0, ..., 7}, t={2, ..., 10}, etc. for *unsafe* samples and take no overlap for *safe* samples, i.e., t={0, ..., 7}, t={8, ..., 15}, etc. During inference, we evaluate sample every t.

For the 2D CNN experiments, only a single frame and a 2D ResNet-18 [38] are used. We use a KITTI pre-trained LiteFlowNet [22] to estimate optical flow and Faster R-CNN ResNet-50 FPN [21] pre-trained on COCO train2017 [26] for the vehicle detections. All detections with less than 80% classification score, as well as all bounding boxes with an area lower than 1% of the total image area, are discarded.

The intermediate representations are reduced to 120×160 (H×W) pixels using bilinear interpolation to speed up the training process. RiskNet is trained with a batch size of 32 on a single NVIDIA Titan Xp GPU for 20 epochs and a checkpoint is saved after every epoch. The model's weights are optimized with the Adam optimizer [39] and a constant learning rate of 1e-5. Finally, noise masks were added with $N_f = 3$ and $p_n = 20\%$.

C. Comparison to State of the Art

In this set of experiments, we compare our work to the closest related work, FlowDroNet [12], as well as to human performance. We train FlowDroNet from scratch on the simulated $\mathcal{D}_{src}$ in order to make a fair comparison. Fig. 8 shows the ROC-curves and PR-curves of both approaches. Compared to FlowDroNet, our approach obtains nearly a 3 times higher average precision and a 42% higher ROC-AUC. Furthermore, the confusion matrices, shown in Fig. 9, demonstrate that we outperform FlowDroNet in all four quadrants. Note that since we evaluate on a strict per-frame level, predicting *unsafe* a single frame too late already results in a false negative.

TABLE II: Comparison of our approach with attention masks (AM) and optical flow (OF) against FlowDroNet [12] and experienced human drivers. *Metric does not adequately outline human performance, as humans did not indicate a confidence level in their predictions.

	YDID performance			
Approach	AP ↑	ROC AUC ↑	F1-score ↑	Accuracy ↑
FlowDroNet [12]	0.24 ± 0.03	0.57 ± 0.06	0.77 ± 0.08	0.75 ± 0.12
Ours (AM)	0.59 ± 0.01	0.88 ± 0.00	0.82 ± 0.00	0.83 ± 0.00
Ours (RGB+AM)	0.53 ± 0.01	0.82 ± 0.01	0.84 ± 0.00	0.83 ± 0.00
Ours (OF+AM)	**0.69** ± **0.00**	**0.90** ± **0.00**	0.92 ± 0.00	0.92 ± 0.02
Human	0.54* ± 0.16	0.86* ± 0.09	**0.94** ± **0.03**	**0.94** ± **0.03**

As outlined in Table II, RiskNet performs best with optical flow and attention masks as intermediate representations, achieving an average precision of 0.69. Interestingly, RiskNet with exclusively attention masks as input obtains an AP of 0.59, whilst RiskNet with RGB and attention masks together only achieves an AP of 0.53. The performance with attention masks alone demonstrates that RiskNet is able to learn valuable information from the sizes and positions of vehicle detections over time, encoded with binary masks, which is highly domain resilient. Secondly, the RGB domain gap is substantial, to the extent that disregarding RGB data altogether and training on attention masks exclusively, results in better performance. This, combined with the performance increase of using optical flow, clearly outlines the benefits of training on task-specific, domain resilient representations.

Finally, a quantitative comparison is made between our approach and human performance on YDID, by testing each drivers' individual performance against the common sense consensus, excluding that individual drivers' performance from the consensus. These findings are also outlined in Table II, and show that on real-world dataset YDID, we approach human performance with a difference of merely 1.8% in accuracy and 0.02 in F1-score. Please note that humans did not indicate a confidence level in there performance, which is why we compare accuracy and F1-score rather than AP. Although our experiments are clearly not extensive enough to truly conclude our system can replace human drivers, they do show that our approach is a very capable and promising direction for future research and development.

D. Component Analysis

We study the contribution of each individual proposed component, the results of which are outlined in Table III. The best configuration of RiskNet is a spatio-temporal (3D) CNN which takes prior filtered attention masks and Lite-

TABLE III: Performance of different configurations of RiskNet on the real-world YDID. We compare RGB to the task-specific intermediate representation optical flow.

Config-		Temporal	Attention	CNN estimated	Prior	YDID performance (**RGB** input)		YDID performance (**OF** input)	
uration	2D CNN	3D CNN	masks	training OF	filter	AP ↑	ROC-AUC ↑	AP ↑	ROC-AUC ↑
a	✓					0.21 ± 0.03	0.61 ± 0.06	0.25 ± 0.01	0.70 ± 0.01
b		✓				0.24 ± 0.03	0.71 ± 0.05	0.16 ± 0.01	0.57 ± 0.02
c		✓	✓			0.52 ± 0.02	**0.84** ± **0.02**	0.61 ± 0.00	0.89 ± 0.00
d		✓	✓	✓		n/a	n/a	0.63 ± 0.00	0.89 ± 0.00
e		✓	✓	✓	✓	**0.53** ± **0.01**	0.82 ± 0.01	**0.69** ± **0.00**	**0.90** ± **0.00**

FlowNet [22] estimated optical flow as input representations (configuration *e*). We observe that our proposed attention masks (configuration *c*) increase RiskNet's average precision by a significant 280% compared to not using attention masks (configuration *b*). Using CNN estimated flow (configuration *d*) yields a 3% improvement in AP compared to the pre-processed ground truth optical flow (configuration *c*), and using priors to filter real-world detection (configuration *e*) results in our highest AP of 0.69. The results of this component analysis confirm our hypothesis that in the automotive domain, where risk originates at distances where optical flow is particularly noisy, vehicle detections are immensely important.

Without explicitly having optimized our system for real-time operation, our approach, including the optical flow and object detection, achieves 8.5 FPS. The object detection model operates at 14 FPS, the optical flow estimation at 26 FPS, and RiskNet itself at 140 FPS. As computation time for the Perception tasks decreases due to continued research efforts on efficient neural networks, so will the computation time of our approach.

E. Qualitative Analysis

In Fig. 10 we demonstrate the qualitative performance of RiskNet and FlowDroNet on three out of the sixteen videos from the YDID dataset. Video 1 and 2 show scenarios where our model's predictions are as desired and similar to the common consensus ground truth. Video 3 shows false positives between frames 15 and 65, due to a stationary bus in the path of the ego-vehicle. The human annotators did not see this as an unsafe event, likely because the annotators could use information from the clip's end whilst labeling the beginning of the clip, see Section IV for the used annotation policy that allows humans to first watch the entire clip given them an 'unfair' advantage. After passing the bus, the model shows desirable behavior by predicting an increased risk, but exceeding t_H, for a vehicle turning into its lane. Even though the scenarios in the videos differ greatly from the constant velocity, constant radius driving in the simulated training data $\mathcal{D}_{src}$, RiskNet demonstrates good performance. On the other hand, FlowDroNet's predictions are noisy and do not provide insights into the vehicle collision risk. The reason for this poor performance of FlowDroNet is the level of noise in optical flow at greater distances, which clearly shows the need for vehicle detections as additional intermediate representation.

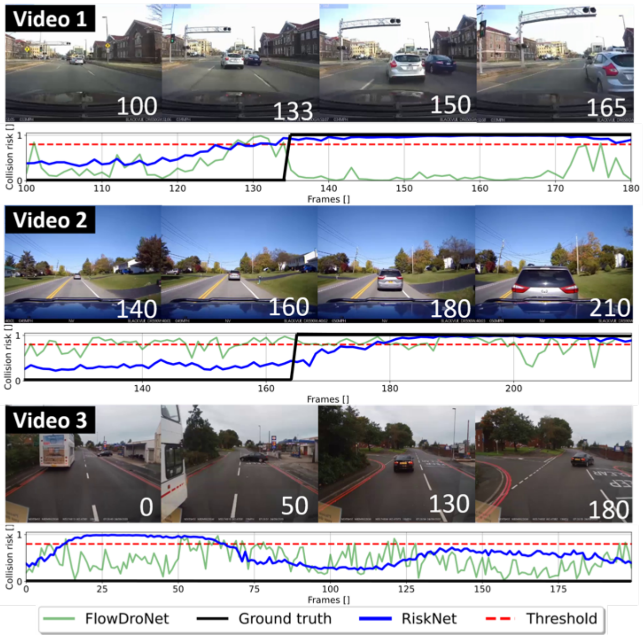

Fig. 10: Results of RiskNet for three YDID videos, containing scenarios which differ greatly from the training data. The black line indicates the ground truth labels, the red dashed line the threshold t_H, the blue line RiskNet's predicted risk and the green line FlowDroNet's predicted risk.

VI. DISCUSSION

In this work, we showed via several experiments that a decoupling learning strategy can be successfully implemented for collision risk prediction, confirming our initial hypothesis that the domain gap between simulated and real data can be bridged by leveraging intermediate, domain resilient representations. Additionally, we showed that our model, trained only on straight-forward but highly random driving scenarios is able to generalize to complex, real-world driving scenarios.

Our approach can be extended by training on more complex training scenarios, specifically removing the constant velocity and radius assumptions. Such additional training data can easily be obtained and is likely to improve RiskNet's performance. In future work, using simulated data makes it possible to generate a multitude of realistic scenarios automatically, and the ability of training on this simulated data without a significant domain gap solves the data scarcity problem. Additionally, although the size and position of the attention masks already give a hint of depth, providing the network with explicit depth estimates could prove to be more robust, especially when dealing with partly occluded vehicles, which inherently have reduced attention masks.

Furthermore, as real-world data is challenging to collect, the YDID contains only 13 positive samples and in total 154 s of driving time, which is not sufficient to make hard claims on the real-world performance of our system compared to human capabilities. Although analysis on more samples is required in order to further demonstrate performance in real-world conditions, our results demonstrate that our approach is capable and a promising direction for future research.

VII. CONCLUSION

Training a collision prediction model on simulated data is attractive, as (near) collisions can be readily generated in simulators, whereas such samples are exceptionally challenging to record in the real world. We proposed to use state-of-the-art approaches in optical flow estimation and object detection, where real-world datasets are abundant, to create task-specific input representations for which the simulation-to-real domain gap is minimal by design. We demonstrated that training a collision risk prediction model on those input representations, generated in a simulator with simple but highly randomized movement patterns, successfully generalized to real-world data. RiskNet predicted risks in complex, real-world scenarios, without requiring any additional domain adaptation techniques, and approaches human-level performance on the real-world YDID dataset.

We believe that with the specific improvements suggested in Section VI, and especially by generating more simulated data training data, which can be done efficiently, the methodology laid down in this work can effectively predict risk in dynamic and complex traffic patterns and thereby contribute to reducing accidents. Additionally, our meta-method can be used for other tasks which require hard to collect, task-specific data.

REFERENCES

[1] *Global status report on road safety 2015*. World Health Organization, 2015.

[2] S. Davis and R. G. Boundy, "Transportation energy data book: Edition 38," Oak Ridge National Lab. (ORNL), Oak Ridge, TN (United States), Tech. Rep., 2020.

[3] M. V. T. Ltd, "Collision warning system," U.S. Patent 9656607B2, May 2017.

[4] W. Wen, H. H. Aghdam, Y. Wang, R. Laganière, and E. Petriu, "A monocular forward leading vehicle distance estimation using mobile devices," in *IV*. IEEE, 2020, pp. 1504–1509.

[5] J. Eggert, F. Damerow, and S. Klingelschmitt, "The foresighted driver model," in *IV*. IEEE, 2015, pp. 322–329.

[6] F. Damerow, T. Puphal, Y. Li, and J. Eggert, "Risk-based driver assistance for approaching intersections of limited visibility," in *ICVES*. IEEE, 2017, pp. 178–184.

[7] A. Paigwar, E. Baranov, A. Renzaglia, C. Laugier, and A. Legay, "Probabilistic collision risk estimation for autonomous driving: Validation via statistical model checking," in *IV*. IEEE, 2020, pp. 737–743.

[8] W. P. Sanberg, G. Dubbelman, and P. H. De With, "Asteroids: A stixel tracking extrapolation-based relevant obstacle impact detection system," *IEEE Transactions on Intelligent Vehicles*, 2020.

[9] P. Wei, L. Cagle, T. Reza, J. Ball, and J. Gafford, "Lidar and camera detection fusion in a real-time industrial multi-sensor collision avoidance system," *Electronics*, vol. 7, no. 6, p. 84, 2018.

[10] W. P. Sanberg, G. Dubbelman, and P. H. de With, "Extending the stixel world with online self-supervised color modeling for road-versus-obstacle segmentation," in *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2014, pp. 1400–1407.

[11] A. Loquercio, A. I. Maqueda, C. R. Del-Blanco, and D. Scaramuzza, "Dronet: Learning to fly by driving," *IEEE RA-L*, vol. 3, no. 2, pp. 1088–1095, 2018.

[12] M. Sperling, Y. Bouteiller, R. de Azambuja, and G. Beltrame, "Domain generalization via optical flow: Training a cnn in a low-quality simulation to detect obstacles in the real world," in *CRV*. IEEE, 2020, pp. 117–124.

[13] L. Xie, S. Wang, A. Markham, and N. Trigoni, "Towards monocular vision based obstacle avoidance through deep reinforcement learning," *arXiv preprint arXiv:1706.09829*, 2017.

[14] F.-H. Chan, Y.-T. Chen, Y. Xiang, and M. Sun, "Anticipating accidents in dashcam videos," in *ACCV*. Springer, 2016, pp. 136–153.

[15] T. Suzuki, H. Kataoka, Y. Aoki, and Y. Satoh, "Anticipating traffic accidents with adaptive loss and large-scale incident db," in *CVPR*, 2018, pp. 3521–3529.

[16] H. Kataoka, T. Suzuki, S. Oikawa, Y. Matsui, and Y. Satoh, "Drive video analysis for the detection of traffic near-miss incidents," in *ICRA*. IEEE, 2018, pp. 3421–3428.

[17] D. de Geus, P. Meletis, and G. Dubbelman, "Single network panoptic segmentation for street scene understanding," in *IV*. IEEE, 2019, pp. 709–715.

[18] S. Zheng, Y. Song, T. Leung, and I. Goodfellow, "Improving the robustness of deep neural networks via stability training," in *Proceedings of CVPR*, 2016, pp. 4480–4488.

[19] Y. Yang and S. Soatto, "Fda: Fourier domain adaptation for semantic segmentation," in *CVPR*, 2020, pp. 4085–4095.

[20] R. Romijnders, P. Meletis, and G. Dubbelman, "A domain agnostic normalization layer for unsupervised adversarial domain adaptation," in *WACV*. IEEE, 2019, pp. 1866–1875.

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *arXiv preprint arXiv:1506.01497*, 2015.

[22] T.-W. Hui, X. Tang, and C. C. Loy, "Liteflownet: A lightweight convolutional neural network for optical flow estimation," in *CVPR*, 2018, pp. 8981–8989.

[23] F. J. Piva and G. Dubbelman, "Exploiting image translations via ensemble self-supervised learning for unsupervised domain adaptation," *arXiv preprint arXiv:2107.06235*, 2021.

[24] RiskNet repo: https://github.com/tue-mps/risknet.

[25] S. Eppel, "Setting an attention region for convolutional neural networks using region selective features, for recognition of materials within glass vessels," *arXiv preprint arXiv:1708.08711*, 2017.

[26] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*. Springer, 2014, pp. 740–755.

[27] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *IJRR*, vol. 32, no. 11, pp. 1231–1237, 2013.

[28] Y. Zou, Z. Yu, B. Kumar, and J. Wang, "Domain adaptation for semantic segmentation via class-balanced self-training," *arXiv preprint arXiv:1810.07911*, 2018.

[29] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *ECCV*. Springer, 2016, pp. 102–118.

[30] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in *CVPR*, 2018, pp. 6546–6555.

[31] "Prescan," Oct 2020. [Online]. Available: https://tass.plm.automation.siemens.com/prescan

[32] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *IROS*. IEEE, 2017, pp. 23–30.

[33] B. Wulfe, S. Chintakindi, S.-C. T. Choi, R. Hartong-Redden, A. Kodali, and M. J. Kochenderfer, "Real-time prediction of intermediate-horizon automotive collision risk," *arXiv preprint arXiv:1802.01532*, 2018.

[34] A. Rasouli, I. Kotseruba, T. Kunic, and J. K. Tsotsos, "Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction," in *ICCV*, 2019, pp. 6262–6271.

[35] K. Kishida, *Property of average precision and its generalization: An examination of evaluation indicator for information retrieval experiments*. National Institute of Informatics Tokyo, Japan, 2005.

[36] J. N. Mandrekar, "Receiver operating characteristic curve in diagnostic test assessment," *JTO*, vol. 5, no. 9, pp. 1315–1316, 2010.

[37] L. A. Jeni, J. F. Cohn, and F. De La Torre, "Facing imbalanced data–recommendations for the use of performance metrics," in *ACII*. IEEE, 2013, pp. 245–251.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.