CycleCrash: A Dataset of Bicycle Collision Videos for Collision Prediction and Analysis

Nishq Poorav Desai   Ali Etemad   Michael Greenspan

Queen's University, Canada

{n.desai, ali.etemad, michael.greenspan}@queensu.ca

Abstract

Self-driving research often underrepresents cyclist collisions and safety. To address this, we present CycleCrash, a novel dataset consisting of 3,000 dashcam videos with 436,347 frames that capture cyclists in a range of critical situations, from collisions to safe interactions. This dataset enables 9 different cyclist collision prediction and classification tasks focusing on potentially hazardous conditions for cyclists and is annotated with collision-related, cyclist-related, and scene-related labels. Next, we propose Vid-NeXt, a novel method that leverages a ConvNeXt spatial encoder and a non-stationary transformer to capture the temporal dynamics of videos for the tasks defined in our dataset. To demonstrate the effectiveness of our method and create additional baselines on CycleCrash, we apply and compare 7 models along with a detailed ablation. We release the dataset and code at https://github.com/DeSinister/CycleCrash/.

1. Introduction

The growing popularity of cycling as a sustainable and healthy means of urban commuting brings inherent risks and safety concerns placing cyclists among the most vulnerable road users [34, 49]. Recent studies show that over 130,000 cyclist injuries occur annually due to crashes, while cyclist fatalities have grown by more than 50% over a decade [11, 31]. While machine learning has provided a means to address many problems related to self-driving vehicles, the development of such data-driven solutions for cyclist safety is hindered by the lack of available data that is targetted for this particular problem [18, 32].

Some recent self-driving vehicle datasets [27, 47, 63, 64] have considered bicycles alongside cars for tasks such as object recognition, time-to-collision, and traffic scene understanding. However, the limited representation of cyclist-related instances in these datasets poses a challenge, as they make up only a small proportion of the data. For exam-
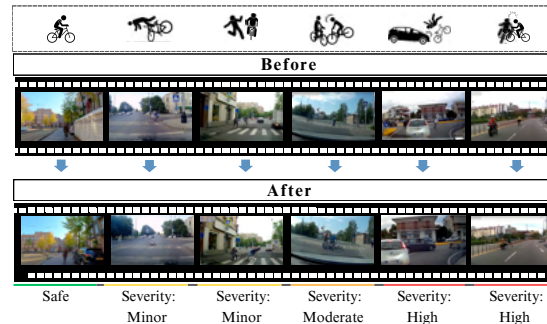

Figure 1. A few samples from CycleCrash showcasing various cyclist-related interactions with different vehicles along with collision severity levels.

ple, among widely used urban driving datasets, DoTA [63] contains only 1,031 frames with at least one cyclist, while KITTI [21] has only 1,627 such frames. Moreover, while datasets like SUTD-TrafficQA [61] explore causal considerations like fault, the existing datasets do not explicitly provide the necessary annotations to facilitate cyclist safety, such as cyclist behaviour risk levels or collision severity.

To address this issue and facilitate deep learning solutions for cyclist safety, we present **CycleCrash**, a novel dataset that explicitly addresses several key challenges in this area. CycleCrash comprises 3,000 videos sourced primarily from the web. In summary, we make the following contributions. (**1**) CycleCrash comprises 3,000 video clips of cyclist-related scenes along with **13 types of annotation**, organized into the 3 categories of *collision-related*, *cyclist-related*, and *scene-related*. Fig. 1 shows different interactions of cyclists in the CycleCrash dataset along with associated collision severity levels. (**2**) We define **9 key tasks** (classification or regression) based on our dataset, with implications on cyclist safety. (**3**) To accurately perform the defined safety-related tasks based on the collected data, we propose VidNeXt, a novel architecture for video representation learning. VidNeXt combines a ConvNeXt feature extractor with a non-stationary transformer for the first time

Dataset	Year	Duration (hr)	Clips	Cyclist Frames	Cyclist BBox	Collision Time	Cyclist Age	Risky Behav.	Severity	Fault	Right-of-way	Cyclist Dir.
KITTI [21]	2012	1.5	22	1,627	✓	✗	✗	✗	✗	✗	✗	✗
DAD [12]	2017	2.4	1,730	✗	✗	✗	✗	✗	✗	✗	✗	✗
CCV [40]	2018	0.1	177	✗	✓	✓	✗	✗	✗	✗	✗	✗
CADP [47]	2018	5.2	1,416	<50k	✓	✓	✗	✗	✗	✗	✗	✗
DACACD [6]	2019	5.0	392	✗	✓	✓	✗	✗	✗	✗	✗	✗
A3D [64]	2019	3.6	1,500	✗	✓	✓	✗	✗	✗	✗	✗	✗
STAG-Nets [27]	2019	8.9	803	✗	✗	✓	✗	✗	✗	✗	✗	✗
BDD100k [66]	2020	1,111.1	100k	13,122†	✓	✗	✗	✗	✗	✗	✗	✗
DoTA [63]	2020	20.0	4,677	1,031	✓	✓	✗	✗	✗	✓	✗	✗
WOD [51]	2020	10	1,000	✗	✓	✗	✗	✗	✗	✗	✗	✗
nuScene [10]	2020	5.5	1,000	<11.8k	✓	✗	✗	✗	✗	✗	✗	✗
CCD [5]	2020	6.3	1,500	✗	✗	✓	✗	✗	✗	✗	✗	✗
RetroTrucks [24]	2020	1.9	474	✗	✗	✓	✗	✗	✗	✗	✗	✗
DADA2000 [19]	2021	6.1	2,000	✗	✗	✗	✗	✗	✗	✗	✗	✗
SUTD-TrafficQA [61]	2021	✗	10k	✗	✓	✓	✗	✗	✗	✓	✗	✗
Lyft Level5 [28]	2021	1,118	170k	20,928†	✓	✗	✗	✗	✗	✗	✗	✗
CycleCrash (ours)	2024	4.1	3,000	436k	✓	✓	✓	✓	✓	✓	✓	✓

Table 1. Comparison of CycleCrash against existing datasets. † indicates partial annotation, ✓ indicates that the label is available, while ✗ indicates that the information/label is unavailable or unknown.

and is, explicitly designed to learn both stationary and non-stationary spatiotemporal information from videos, which allows it to perform accurate prediction and classification. Through extensive experiments, we show that our model outperforms various baselines on most CycleCrash tasks.

2. Related Works

We summarize existing crash-related datasets from the literature in Tab. 1, and observe that none specifically focus on cyclists. Therefore in this section, we present the next most relevant body of work, i.e., datasets and methods focusing on self-driving vehicles, and car collisions.

Car Collision Datasets. Early datasets focusing on car collisions, such as Car Crash Videos (CCV) [40] and its successors [6, 47, 50, 61], include annotations of car collisions viewed from the vantage of fixed perspective cameras such as pole-mounted CCTV cameras. In contrast, newer datasets use dashcam videos widely accessible from internet sources. As observed in Tab. 1, CADP [47] offers the highest number of cyclist frames among the datasets that report this count, with fewer than 50,000 frames. This demonstrates a clear gap, as it encompasses only about one-ninth of the cyclist frames available in the CycleCrash dataset.

Additionally, it is observed that despite containing 3,000 video clips, CycleCrash's total duration is only 4.1 hours, which is shorter than some other datasets containing fewer video clips. This is expected because we temporally crop the video clips in CycleCrash to exclude those parts of the video which do not contain relevant cyclist information.

A wide variety of methods for addressing traffic safety have been previously outlined, such as identifying the occurrence of collisions which was first introduced by the DAD dataset [12], 'time-to-collision' which was coined by the CADP dataset [47], as well as SUTD-TrafficQA dataset [61], which is a video-based question-answering dataset that also includes causal considerations such as 'fault'. However, annotations such as 'right-of-way' have been relatively unexplored. CycleCrash introduces novel annotations such as the 'cyclist behaviour risk index' and 'severity', which have not been utilised before to the best of our knowledge. These annotations can play an important role in responding promptly to avoid collisions, as well as potentially alerting first responders in critical situations. Furthermore, the types of annotations in existing datasets are often fragmented, highlighting the absence of a unified dataset with credible cyclist frames and a comprehensive collection of annotations for cyclist safety. Owing to its specific cyclist-related focus, CycleCrash provides a unique resource to enable increased cyclist safety, e.g., cyclist accident prediction, time-to-collision forecasting, etc., for the autonomous driving and driver assist research community.

Methods. The evolution of urban-traffic-related video classification and detection methods closely aligns with dataset diversity. Early approaches focused on conventional methods such as YOLO [9] for binary collision detection [40]. Convolutional neural nets (CNNs) when in combination with LSTMs [43], or in the form of 3D-CNN (such as ResNet3D, X3D, DensNet3D etc.) [20, 23] show prominent application in urban traffic video-based tasks [3, 42, 62, 67], although they are somewhat computationally expensive. R(2+1)D [55] was introduced to address these computational drawbacks with divided spatial and temporal CNNs.

Transformers [56] have demonstrated tremendous success in the vision domain, and have been adapted for urban-traffic-related video tasks including object detection, multiple object tracking [29], natural driving action recognition [15] and anomaly detection [1, 13, 16, 37, 48] . Examples of such transformer-based methods include ViViT [4], VideoMAE [54, 58] and TimeSformer [8], all of which also apply similar spatiotemporal division potentially inspired by R(2+1)D. However, they still are ineffective in captur-

ing global representation and aren't specifically designed to handle stationary and non-stationary elements of temporal information. In processing the temporal domain, a common pre-processing step is to adopt stationarization to improve predictability, albeit at the cost of losing non-stationary information. We propose VidNeXt to address this issue by using both stationary and non-stationary components for improved video representation learning (see Sec. 4).

3. CycleCrash Dataset

To address the limitations identified in prior works regarding cyclist safety, we introduce CycleCrash. In this section, we describe the collection procedure, data format, annotations, and other details of this dataset. The dataset along with the tool kit (see Appendix A for details) will be made publicly available upon publication of the work.

3.1. Data Collection

We curated the dataset from the web, namely from YouTube [65], Vimeo [57], DailyMotion [14], Facebook [17], Instagram [30], X (formerly known as Twitter) [60], and TikTok [53]. Each video contains one or more cyclists, as viewed from vehicle dashcams. The following are the criteria used to qualify a video for inclusion in our dataset, where the first four constitute a collision or a near-miss:

- There exists a collision or near-miss involving a cyclist and a motor vehicle (i.e. car, motorcycle, bus, etc.);
- There exists a collision between a cyclist and a pedestrian or another cyclist;
- There exists a loss of balance or fall due to potholes, animals, mechanical failure of the bicycle, etc.;
- There exists risky behaviour shown by the cyclist, justifying the need for extra attention to be given by drivers to prevent a collision;
- There exists a cyclist navigating safely in an urban traffic scenario, without depicting risky behaviour or being involved in a collision or near-miss.

The dataset comprises 3,000 video clips, out of which 2,000 are accident-free and 1,000 exhibit potential accidents of various levels of severity and near-misses involving a cyclist of interest. We collected dashcam videos instead of fixed-perspective CCTV footage to facilitate the development and integration of cyclist collision warning and mitigation systems in vehicles.

3.2. Data Format and Structure

CycleCrash is a curated list of video links accompanied by precise start and end timestamps. All videos are sourced explicitly from public posts, ensuring no private videos are used. Following established practices from prior works, such as [5, 63, 64], we do not download or share the videos. Instead, we provide links to the original videos along with the relevant timestamps. We create and include

a library with our dataset, that downloads and pre-processes the videos for uniformity in spatiotemporal properties. Our pre-processing pipeline includes:

Temporal Cropping. Each video is temporally cropped based on the provided start and end times.

File Conversion. All videos are converted to .mp4 format.

Spatial Cropping and Scaling. Videos are rescaled to a uniform resolution of $1,280 \times 720$ pixels using interpolation. When a video is not in a $16/9$ aspect ratio, they are cropped before rescaling to achieve the desired aspect ratio.

Frame Rate Adjustment. Videos are adjusted to a consistent frame rate of 30 frames per second (fps). For videos that were initially higher than 30 fps, our library performs temporal sub-sampling, whereas, for videos with lower fps rates, the library performs up-sampling using interpolation.

Color Normalization. Videos are z-score normalized.

3.3. Dataset Annotation

We conducted a comprehensive survey on possible attributes for safety [33, 48, 52] and compiled a list, which we then narrowed down to 13 annotations based on perceived importance. The 13 different annotations, organized into three categories, are as follows:

Collision-related, which refers to characteristics associated with annotations in the context of a collision or close call. Factors like the right-of-way and time of collision contribute to enabling autonomous vehicles to predict and avoid collisions. They include 'right-of-way', 'time-to-collision', 'type of object involved', 'fault', 'severity';

Cyclist-related, which captures information regarding the cyclist of interest e.g., cyclist behaviour risk index, age or appearance. They include 'cyclist behaviour risk index', 'cyclist age', 'cyclist type' (competitive or recreational), 'cyclist bounding box', and 'direction of the cyclist';

Scene-related, which provides additional information regarding the motion of the vehicle involved in the scene, as well as the cameras used to capture the videos. They include the 'direction of the object involved', 'camera position', and 'ego-vehicle involved'.

We present a detailed description of each annotation in Tab. 2, and examples of annotated videos in Fig. 2 (additional visuals are presented in Appendix F).

3.4. Quality Control

The videos were collected based on search terms such as 'cyclist hitting car', 'bicyclist car accident', 'cyclist accident dashcam', and others (please see the full list in Appendix B). The videos were automatically processed to ensure their frame rates were at least 20 fps, and that their resolutions were greater than 854×480 pixels. Following this, they underwent further human inspection to ensure the conditions in Sec. 3.1 were met. The selected set of videos

No.	Annotation	Description	Options
1	Right-of-way	Binary label capturing the priority between the cyclist of interest and others involved (e.g., vehicle, other cyclists, pedestrian)	Yes, No
2	Time-to-collision	Exact timestamp on which the collision begins if one exists in the video clip	Timestamp (if a collision occurs)
3	Type of object involved	17-class categorization to describe the type of object involved in the interaction	Car, Bus, Train, Cyclist, Pedestrian, Pothole, Animal, etc.
4	Fault	Binary label determining whether the collision was the cyclist of interest's fault	Yes, No
5	Severity	5-class categorization to describe the intensity of the collision	Safe, Minor, Moderate, High, Very high
6	Cyclist behaviour risk index	4-class categorization to indicate the risk levels portrayed by the cyclist	Low, Moderate, High, Very High
7	Cyclist age	3-class categorization of the age of the cyclist.	Young, Adult, Old
8	Cyclist type	Binary label capturing if the cyclist is participating in competitive or recreational cycling	Competitive, Recreational
9	Cyclist bounding box	Bounding box captured around the interested cyclist	Coordinates in x,y,w,h format
10	Direction of the cyclist	5-class categorization of the direction of the cyclist's motion as observed from the ego-vehicle along the medial and lateral directions	Forward (+ve medial motion $x+$), Backward (-ve medial motion $x-$), Left (-ve lateral motion $y-$), Right (+ve lateral motion $x+$), and stationary (no motion)
11	Direction of the object involved	5-class categorization of the direction of the object involved as observed from the ego-vehicle along the medial and lateral directions	Forward, Backward, Left, Right, Stationary
12	Camera position	Categorization based on the camera's position (Front/Back) and the camera carrying vehicle	Front dashcam, Back dashcam, Front cyclist-helmet camera etc.
13	Ego-vehicle involved	Binary label indicating whether the ego-vehicle is involved in a collision or not	Yes, No

Table 2. Descriptions of annotations provided in CycleCrash along with possible options.

were then preprocessed and filtered algorithmically into the common configuration described in Sec. 3.2.

To ensure bias mitigation for subjective annotations such as risk, we provided labellers with detailed instructions to pay attention to traffic rules, legal speed, distance from other objects and vehicles, etc. For age, labellers were instructed to consider the visual appearance of the cyclist. For severity, labellers were asked to pay attention to collision impact and degree of estimated injury caused to the cyclist. When video data lacked sufficient information to confidently determine scores for 'right-of-way', 'fault', 'age', and 'cyclist type', annotators were asked to use '-1'. Each video was annotated by three different labellers with valid driving licenses and sufficient knowledge of traffic conditions. We employed the Latin Square method [7] for labelling, which presented each labeller with an equal distribution of tasks and scenarios, thereby reducing bias and enhancing the quality of the dataset's annotations. For categorical annotations we use the median, while for continuous labels, an average of the values is used. Qualitative labels including 'risk' and 'age' have been previously labelled in other datasets [35, 52], and we followed a similar labelling protocol. Lastly, we follow [2, 46], and use Randolph's Coefficient [44] to measure inter-annotator agreement for 'cyclist age' (0.81), 'severity' (0.76), and 'risk' (0.84), which are considerably higher than previous works and indicate substantial agreement among labellers.

3.5. Tasks and Metrics

We define 9 tasks based on the labels, as follows.

Task 1: Cyclist Behaviour Risk Index Prediction estimates the cyclist's level of risky behaviour based collected for our dataset which is first recorded on a continuous scale of 0 to 1 and subsequently quantized into the 4 classes 'Low', 'Moderate', 'High', and 'Very High', to reduce subjective variability and cast the task as a classification problem rather than regression. We adopt accuracy and F1 as the metrics to quantify the performance on this task.

Task 2: Right-of-way Classification involves classifying the priority in a cyclist's interaction with another vehicle or pedestrian. We treat this task as a binary classification to determine if the cyclist has the right-of-way or not. We use accuracy and F1 to evaluate performance on this task. Although a few prior works [5, 61] have implicitly analyzed right-of-way as a fault-related factor, they have not exclusively outlined right-of-way as a separate task.

Task 3: Cyclist Collision Anticipation predicts whether the sequence will result in a collision or not within a specified timeframe of t seconds, where t is a predefined horizon. Prior works [5, 6, 24, 40, 61, 63] have explored this task in other contexts using horizons ranging from 0.33 to 0.64 seconds, whereas we standardize on a horizon of 1 second. We use binary accuracy to quantify performance on this task.

Task 4: Time-to-Collision Prediction is a regression task that involves predicting the exact moment when a collision with a cyclist will occur following the completion of the
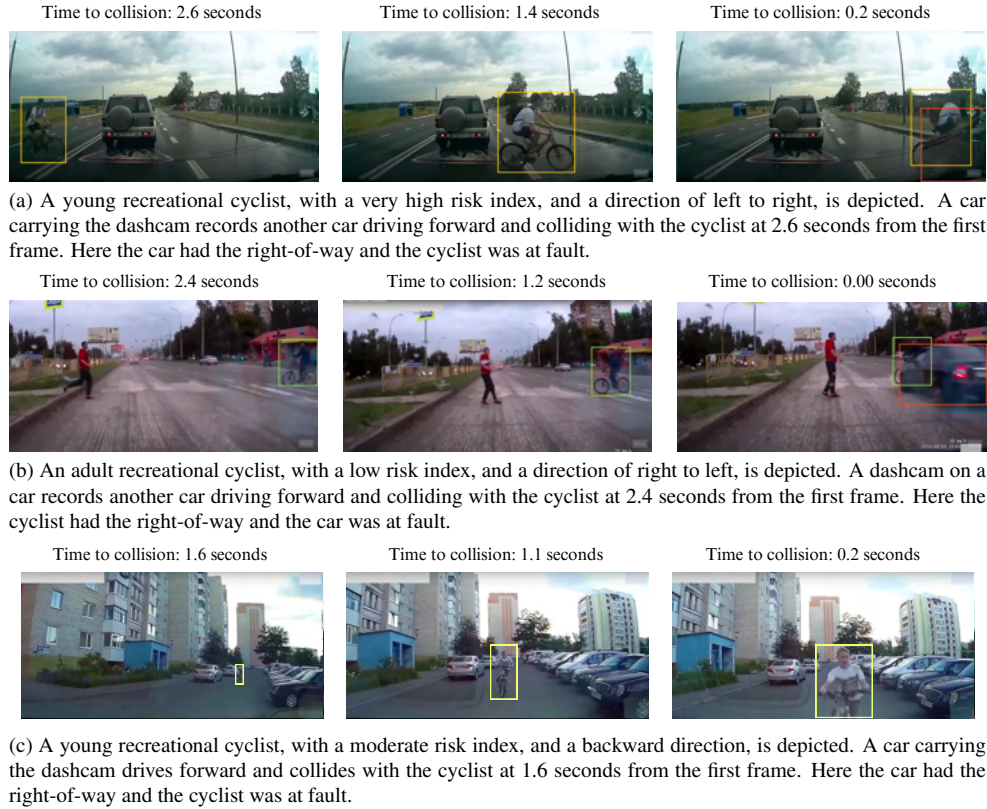
Time to collision: 2.6 seconds   Time to collision: 1.4 seconds   Time to collision: 0.2 seconds

(a) A young recreational cyclist, with a very high risk index, and a direction of left to right, is depicted. A car carrying the dashcam records another car driving forward and colliding with the cyclist at 2.6 seconds from the first frame. Here the car had the right-of-way and the cyclist was at fault.

Time to collision: 2.4 seconds   Time to collision: 1.2 seconds   Time to collision: 0.00 seconds

(b) An adult recreational cyclist, with a low risk index, and a direction of right to left, is depicted. A dashcam on a car records another car driving forward and colliding with the cyclist at 2.4 seconds from the first frame. Here the cyclist had the right-of-way and the car was at fault.

Time to collision: 1.6 seconds   Time to collision: 1.1 seconds   Time to collision: 0.2 seconds

(c) A young recreational cyclist, with a moderate risk index, and a backward direction, is depicted. A car carrying the dashcam drives forward and collides with the cyclist at 1.6 seconds from the first frame. Here the car had the right-of-way and the cyclist was at fault.

Figure 2. Sample frames from 3 video clips along with descriptions and annotations from the CycleCrash dataset.

video sequence. We use MSE for the evaluation of this task.

Task 5: Severity Classification indicates the impact of an accident on the cyclist into 5 classes: 'safe', 'minor', 'moderate', 'high', and 'very high'. We employ both accuracy and F1 as evaluation metrics for this task.

Task 6: Fault Classification involves characterizing the fault in the case of a collision or a near-collision between the cyclist and the other object involved. We treat this as a binary classification task to estimate whether the cyclist was at fault. We use accuracy and F1 to evaluate performance on this task.

Task 7: Cyclist Age Classification involves classifying the age of the cyclist of interest into one of the three categories: 'young', 'adult', or 'old'.

Task 8: Direction of the Cyclist Detection classifies the final direction of the cyclist in the given video into five different possible classes: 'forward', 'backward', 'left', 'right', or 'stationary'. We employ both accuracy and F1 as evaluation metrics for this task.

Task 9: Direction of the Object Involved Detection is a similar task to 'Direction of the Cyclist Detection', which aims to estimate the final direction of the other object involved with the five possible categories.

For the task of 'Cyclist Behaviour Risk Index Prediction', 'Direction of the Cyclist Detection', and 'Cyclist Age

Classification', all 3,000 videos (both collision and safe) were considered. For the rest of the tasks, the 1,000 collision or near-miss videos were used. We provide the results for a multi-task implementation with all 9 tasks in Appendix C and discuss future research directions in Appendix D.

3.6. Data Statistics

The proposed CycleCrash dataset contains 3,000 videos, with durations varying from 1.5 to 21 seconds, totalling over 436,347 frames. We present the key statistics of our dataset in Fig. 3 and Fig. 4. Fig. 3 (i) illustrates the distribution of time-to-collision in seconds, while Fig. 3 (ii) highlights the distribution of the duration of all video clips in the dataset. Fig. 3 (iii) provides a breakdown of the types of objects involved in cyclist collisions. We observe that in the majority of videos, the other object involved is the car as expected. Fig. 3 (iv) presents the distribution of instances across different cyclist behaviour risk indexes, suggesting that in most cases, cyclists show reasonably low-risk behaviour across the dataset. Fig. 3 (v) depicts the perceived age distribution of the cyclists in the videos, indicating that the majority of subjects are adults. Finally, Fig. 3 (vi) presents the fault distribution, demonstrating that fault is reasonably similar between cyclists and others.

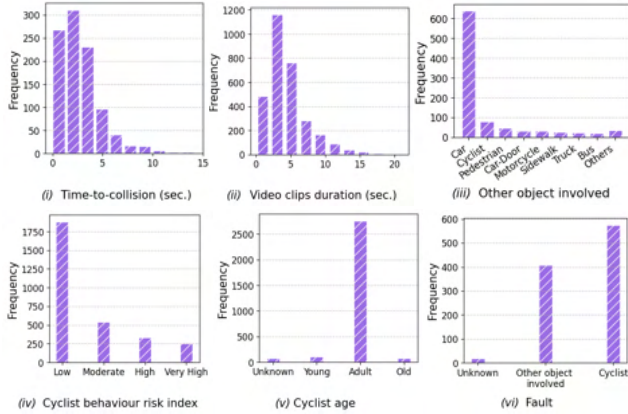Fig. 4 (i) displays the frequency of collisions based on

Figure 3. Distribution of CycleCrash data for (*i*) time-to-collision, (*ii*) duration of video clips, (*iii*) other objects involved, (*iv*) behaviour risk index, (*v*) age, and fault.
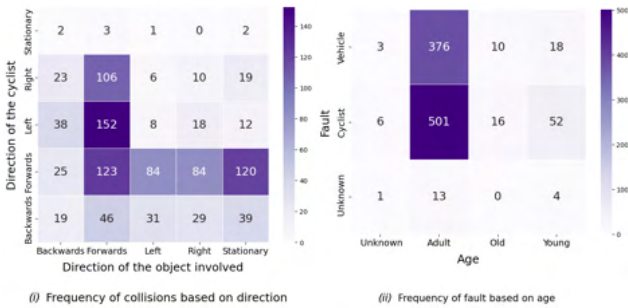

Figure 4. Relationship between (*i*) direction of cyclists and objects involved in collisions, (*ii*) fault and age.

the directions of the cyclist and the other objects involved relative to the ego-motion of the dashcam. It is observed that the highest number of collisions occur when the direction of the cyclist is toward the left and the direction of the other object involved is forward. Additionally, we observe that the lowest collisions occur when the cyclist is stationary. To clarify, the occurrence of rare accidents between stationary cyclists and stationary objects ($n = 2$) refers to instances where the object, e.g., a traffic sign, falls on a stationary cyclist. Moreover, Fig. 4 (*ii*) presents fault against cyclist age, where we observe that young cyclists are 3 times more likely to be at fault compared to adults, while this factor is 1.6 times for older cyclists. Additional statistics are provided in Appendix E.

4. Proposed Method: VidNeXt

Videos, like most time-series data, contain both stationary (information with constant statistical moments in time) and non-stationary components. State-of-the-art video learning networks, however, have not been designed with explicit learning of stationary and non-stationary informa-

tion in mind. This, we believe, can inhibit robust forecasting [36,59,68,69], which is a key factor for the tasks defined based on our dataset. To address this issue and effectively perform the tasks facilitated with CycleCrash, we design VidNeXt (see Fig. 5a), a model that employs a dedicated spatial learner followed by temporal encoding of both stationary and non-stationary information.

The spatial component of VidNeXt is a ConvNeXt [38] feature extractor. This encoder is a CNN which borrows design elements from transformers and is, therefore, more compatible with them [38] through leveraging layer normalization and GELU activations [26]. We empirically find that the base version of ConvNeXt [38] works best for our purpose. This is followed by an encoder-decoder transformer [36] comprising transformer blocks with de-stationary attention (see Fig. 5b) which was originally designed for time-series forecasting.

The temporal component in VidNeXt, as demonstrated in [36], involves the normalization of a series of embeddings (see Fig. 5a), along with the computation of statistical moments such as mean and variance. This normalization process results in stationary series x' containing stationary information. Subsequently, the attention mechanism calculates attention scores based on this stationary series. To incorporate the non-stationary component into the attention scores, they are rescaled using re-scaling factors (Δ) and (τ) in the self-attention block, referred to as 'de-stationary attention'. These factors are determined via a projector based on the raw data x and the statistical moments calculated during normalization. Following this, the output embeddings from the encoder-decoder architecture (y') are de-normalized using the previously calculated means and variances. Finally, an MLP head is applied to the video embeddings according to the specific requirements of one of a number of tasks.

The motivation behind disentangling complementary elements, namely stationary and non-stationary information, stems from two main factors. First, stationarization aims to attenuate non-stationarity, leading to improved predictability. Second, while stationarized series may offer better predictability, they may offer limited insights for forecasting real-world spurious events. This limitation can result in over-stationarization, where transformers generate indistinguishable temporal attention for different series. Thus, re-incorporating intrinsic non-stationary information such as statistical moments of spatial embeddings is crucial for enhancing overall predictability. The non-stationary transformer in [36] was originally used for time series forecasting with an input time series sequence of length 96 and a prediction sequence of length 48. We modify the time series sequence length to match the number of frames in the input, which in our case is 30. We also set the prediction length to a value of 1, to indicate a 1D vector representing
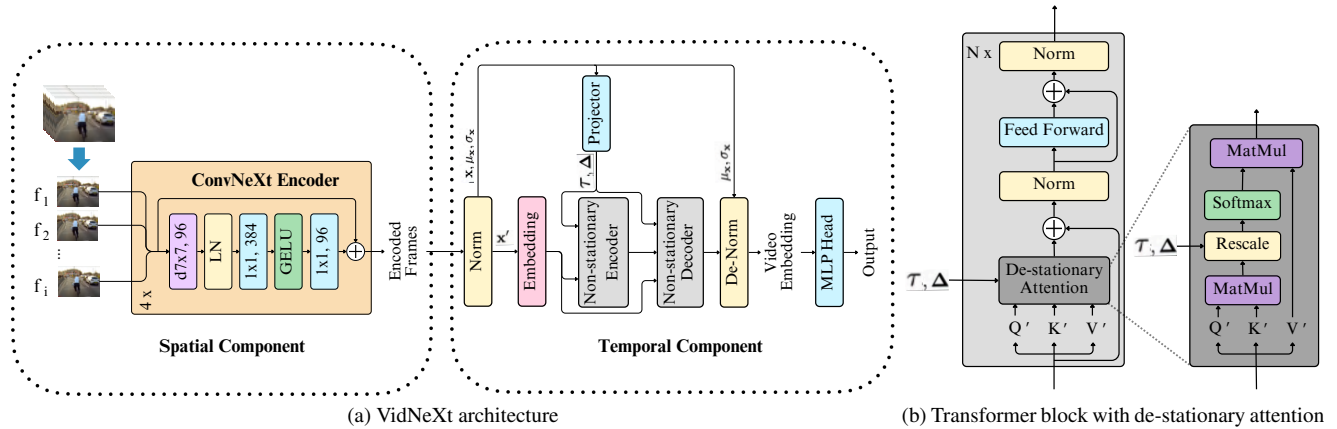
(a) VidNeXt architecture	(b) Transformer block with de-stationary attention

Figure 5. **(a)** The architecture of the proposed method, VidNeXt is presented. First, ConvNeXt is used to encode the video frames. Next, the frame embeddings are normalized for stationarity. Non-stationary information is reintegrated into the transformer blocks via rescaling factors τ and Δ, determined by the projector based on the previous normalization layer. **(b)** Architectural details of transformer with de-stationary attention.

the whole video. We use the same output embedding dimension of ConvNeXt as the input embedding dimension length of the encoder and decoder.

5. Experiments

Dataset Splits. We randomly partition the videos into training and testing sets with $7:3$ ratio. Each video is then segmented into 1-second sequences with 0.5-second overlaps. We provide the code to obtain the fixed splits to allow fair comparisons.

Baseline Models. For baseline comparisons, we have selected models widely adopted in the community for video processing tasks: ResNet50 3D [23], R(2+1)D [55], X3D [20], TimeSformer [8], and ViViT [4]. ResNet50 3D, R(2+1)D, and X3D are used as common convolutional video learners. Notably, we use the X3D-XS, X3D-S, and X3D-M versions, which use 4, 13, and 16 input frames respectively. Additionally, we include TimeSformer and ViViT as competitive transformer-based methods. All the baseline models input a common frame size of 224×224. In addition to the baselines, we present two **ablation variants** to isolate the impact of the spatial and temporal components of VidNeXt. In the first variant (ConvNeXt+VT), we use a vanilla transformer (VT) instead of a non-stationary transformer (NST) for the temporal component, while the second variant (ResNet+NST) involves ResNet-18 [25] instead of ConvNeXt for the spatial component.

Training Details. We train all models on 30-frame (1 sec.) video clips. The networks are trained using batch size 32, AdamW [39] optimization, and ReduceLROnPlateau scheduling with learning rate $2e^{-6}$ for 50 epochs. For all the encoder backbones used in this study, including baseline models and components of our proposed method, we

use the pretrained weights provided with their public distributions [4,8,20,23,25,38,55]. As indicated earlier, all video frames are initially 1280×720 pixels. Following [22,45], we first perform augmentation on all video sequences by randomly cropping at 700×700 pixels followed by random flipping with a probability of 0.25. The 'Time-to-collision Prediction' task is inherently imbalanced because segmentation tends to produce most labels in videos with shorter time-to-collision values and fewer with longer ones. To address this, we upsampled videos by duplicating and augmenting brightness, contrast, saturation, and hue to the training data. We use the spatial dimension of 224×224 pixels. We utilize cross-entropy loss for classification tasks and mean squared error for the regression tasks. Additionally, we perform mixed-precision training [41] on two 40 GB NVIDIA A100 GPUs to save the computation overhead. We use the available implementations of the baseline models in PyTorch.

Results. We perform the 9 tasks described earlier in Sec. 3.5 and present the results in Tab. 3. We observe that for the majority of tasks defined in CycleCrash, VidNeXt outperforms prior methods. This is evident as the proposed method outperforms the others on most tasks, for instance 'Cyclist Behavior Risk Index Prediction', 'Right-of-way Classification', 'Cyclist Collision Anticipation', 'Severity Classification', 'Fault Classification', 'Cyclist Age Classification', and 'Cyclist Direction Detection'. Among two tasks where VidNeXt does not perform the best, it nevertheless yields the second-best performance in metrics such as MSE in 'Time-to-collision Prediction' and F1 in 'Direction of the Object Involved Detection'. We notice a clear difference between the performance of the ablation variants and the proposed method in almost every task, with the highest margin of improvement being up to 8.4% in ac-

Method	Risk		Right-of-way		Collision		Time-to-coll.	Severity		Fault		Age		Direction		Object	
	Acc.↑	F1↑	Acc.↑	F1↑	Acc.↑	F1↑	MSE↓	Acc.↑	F1↑	Acc.↑	F1↑	Acc.↑	F1↑	Acc.↑	F1↑	Acc.↑	F1↑
TimeSformer [8]	65.74	41.79	60.20	55.71	66.45	69.69	1.41	36.49	23.01	59.65	51.03	93.77	66.68	47.19	31.38	45.02	29.00
ViViT [4]	65.12	39.06	52.84	53.74	57.01	69.92	1.33	47.51	24.47	53.37	50.42	93.56	66.34	36.29	27.99	46.30	26.34
ResNet50 3D [23]	65.76	39.53	59.41	53.97	63.10	60.24	1.38	56.60	26.12	59.37	54.91	94.21	54.86	46.30	30.12	43.27	27.77
R(2+1)D [55]	66.54	39.56	60.31	53.42	67.71	63.33	1.43	56.63	25.46	50.53	52.62	94.41	53.24	47.49	30.36	40.75	25.48
X3D-M [20]	64.76	38.75	59.83	57.57	63.72	61.08	1.44	54.45	24.70	52.16	52.19	94.34	53.78	47.82	31.85	42.72	23.79
X3D-S [20]	63.37	36.28	60.10	56.90	61.49	61.13	1.47	51.80	24.09	60.47	51.88	94.38	54.30	45.62	30.05	42.03	22.38
X3D-XS [20]	64.77	37.23	59.37	53.43	60.59	60.73	1.47	51.39	23.77	56.10	52.59	93.87	52.37	46.77	30.22	41.73	26.57
ConvNeXt+VT	64.89	40.05	61.13	54.00	63.50	65.06	1.56	53.80	26.54	56.74	55.72	94.55	66.78	46.46	32.62	42.85	25.16
ResNet+NST	67.18	40.74	61.77	58.62	60.79	62.28	1.39	53.88	24.67	57.17	54.08	94.52	53.49	45.12	28.48	44.17	26.91
VidNeXt (Ours)	66.20	41.96	64.28	57.51	64.84	70.84	1.38	59.66	31.78	65.16	52.51	94.57	67.88	47.94	31.20	42.31	28.37

Table 3. Combined experimental results for tasks 1 through 9. The methods above the line are baselines based on prior works, while those below the line are VidNeXt and its **ablation** variants.

Trained on	Vehicles	Splits	Tested on	Eval.	Accuracy	F1
CCD [5]	Cars	1	CCD	-	76.14	66.23
			CycleCrash	Finetuning	55.76	25.28
			CycleCrash	Linear	59.78 (+4.0)	44.28 (+21.5)
DoTA [63]	Cars, Cyclist	3	DoTA	-	77.61	84.03
			CycleCrash	Finetuning	56.57	60.75
			CycleCrash	Linear	59.51 (+2.9)	62.12 (+1.4)
CycleCrash (Ours)	Cyclist	1	CycleCrash	-	67.71	63.33
			CCD	Finetuning	54.81	54.17
			CCD	Linear	62.44 (+7.6)	53.28 (−0.9)
			DoTA	Finetuning	63.62	73.14
			DoTA	Linear	68.78 (+5.2)	81.22 (+8.1)

Table 4. Results of cross-dataset evaluation for collision anticipation using R(2+1)D.

Figure 6. t-SNE plot of representations from CycleCrash and two other datasets.

curacy and 14.4% of F1. We also observe that among the other baselines and ablation variants, ConvNeXt+VT and ResNet+NST generally emerge as the second-best or third-best models following VidNeXt.

We perform cross-dataset evaluations on the 'Collision Anticipation' task using CCD [5], DoTA [63], and Cycle-Crash datasets to investigate the transferability of learned representations. Both CCD and predominantly involve only car collision events, and so the extent to which collisions with cars and collisions with cyclists share visual cues remains to be explored. We opt for the R(2+1)D model due to its strong performance in this specific task and perform both finetuning(updates all model parameters) and linear evaluations (only updates the final layer, keeping the rest of the model frozen). The results of these evaluations are presented in Tab. 4. We observe from this table that training on CycleCrash and testing on CCD and DoTA yields considerably better results than the other way around, indicating that CycleCrash contains more information and situations regarding collisions, filling a unique gap in the landscape of datasets in this area. This experiment highlights Cycle-Crash's potential as a rich resource for pretraining models focused on collisions and vehicle safety

We evaluate the distribution of the embeddings of our dataset with respect to CCD and DoTA, in Fig. 6 we visualize 1,000 random samples from each of the datasets using t-SNE. We use a standard Kinetics400 [33] pre-trained
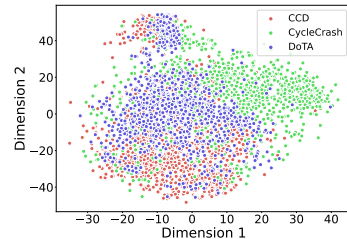
model for this purpose. We observe in this figure that CCD and DoTA datasets considerably overlap with one another. However, while both CCD and DoTA overlap with Cycle-Crash, our dataset occupies a wider area of the latent space, indicating that it contains information not present in either CCD or DoTA. This finding highlights the diverse and distinct scenarios provided in our dataset.

6. Conclusion

In this paper, we present the CycleCrash dataset, a unique collection of 3,000 video clips capturing cyclists in various real-world scenarios, with a particular emphasis on potential cyclist safety in urban settings. To the best of our knowledge, CycleCrash is the first dataset to focus on cyclist-based collision prevention and understanding. The dataset includes 13 annotations for collision, cyclist, and scene-related labels, and defines 9 collision prevention and mitigation tasks. Additionally, we propose Vid-NeXt, a novel method for video classification and regression tailored to learn both stationary and non-stationary information. Furthermore, we evaluate the proposed VidNeXt method against seven baselines and ablation variants on Cy-cleCrash, to demonstrate its stronger performance.

References

[1] Erkut Akdag, Zeqi Zhu, Egor Bondarev, et al. Transformer-based fusion of 2d-pose and spatio-temporal embeddings for distracted driver action recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5452–5461, 2023. 2

[2] Jacopo Amidei, Paul Piwek, and Alistair Willis. Agreement is overrated: A plea for correlation to assess human evaluation reliability. In *International Conference on Natural Language Generation*, 2019. 4

[3] Taif Anjum, Daya Kumar, and Apurva Narayan. Spatio-temporal analysis of dashboard camera videos for time-to-accident forecasting. In *International Joint Conference on Neural Networks*, pages 1–8, 2023. 2

[4] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *IEEE/CVF International Conference on Computer Vision*, 2021. 2, 7, 8

[5] Wentao Bao, Qi Yu, and Yu Kong. Uncertainty-based traffic accident anticipation with spatio-temporal relational learning. In *ACM International Conference on Multimedia*, page 2682–2690, 2020. 2, 3, 4, 8

[6] Elizaveta Batanina, Imad Eddine Ibrahim Bekkouch, Youssef Youssry, Adil Khan, Asad Masood Khattak, and Mikhail Bortnikov. Domain adaptation for car accident detection in videos. In *International Conference on Image Processing Theory, Tools and Applications*, 2019. 2, 4

[7] Lorna S Benjamin. A special latin square for the use of each subject "as his own control". *Psychometrika*, 30:499–513, 1965. 4

[8] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *International Conference on Machine Learning*, 2021. 2, 7, 8

[9] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 2

[10] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020. 2

[11] Centers for Disease Control and Prevention, National Center for Injury Prevention and Control. Bicycle safety. https://www.cdc.gov/transportationsafety/bicycle/index.html, 2022. 1

[12] Fu-Hsiang Chan, Yu-Ting Chen, Yu Xiang, and Min Sun. Anticipating accidents in dashcam videos. In *Asian Conference on Computer Vision*, pages 136–153, 2017. 2

[13] Yunpeng Chang, Zhigang Tu, Wei Xie, Bin Luo, Shifu Zhang, Haigang Sui, and Junsong Yuan. Video anomaly detection with spatio-temporal dissociation. *Pattern Recognition*, 122:108213, 2022. 2

[14] DailyMotion. DailyMotion. Retrieved August 12, 2023, from https://www.dailymotion.com/ca. 3

[15] Xiaodong Dong, Ruijie Zhao, Hao Sun, Dong Wu, Jin Wang, Xuyang Zhou, Jiang Liu, Shun Cui, and Zhongjiang He. Multi-attention transformer for naturalistic driving action recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5434–5440, 2023. 2

[16] Keval Doshi and Yasin Yilmaz. Towards interpretable video anomaly detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2655–2664, 2023. 2

[17] Facebook. Facebook. Retrieved August 12, 2023, from https://www.facebook.com. 3

[18] Jianwu Fang, Jiahuan Qiao, Jianru Xue, and Zhengguo Li. Vision-based traffic accident detection and anticipation: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2023. 1

[19] Jianwu Fang, Dingxin Yan, Jiahuan Qiao, Jianru Xue, and Hongkai Yu. Dada: Driver attention prediction in driving accident scenarios. *IEEE Transactions on Intelligent Transportation systems*, 23(6):4959–4971, 2021. 2

[20] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2, 7, 8

[21] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2012. 1, 2

[22] Yong Guo, David Stutz, and Bernt Schiele. Improving robustness by enhancing weak subnets. In *European Conference on Computer Vision*, pages 320–338, 2022. 7

[23] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018. 2, 7, 8

[24] Sanjay Haresh, Sateesh Kumar, M. Zeeshan Zia, and Quoc-Huy Tran. Towards anomaly detection in dashcam videos. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 1407–1414, 2020. 2, 4

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 7

[26] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 6

[27] Roei Herzig, Elad Levi, Huijuan Xu, Hang Gao, Eli Brosh, Xiaolong Wang, Amir Globerson, and Trevor Darrell. Spatio-temporal action graph networks. In *IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1, 2

[28] John Houston, Guido Zuidhof, Luca Bergamini, Yawei Ye, Long Chen, Ashesh Jain, Sammy Omari, Vladimir Iglovikov, and Peter Ondruska. One thousand and one hours: Self-driving motion prediction dataset. In *Conference on Robot Learning*, pages 409–418, 2021. 2

[29] Thomas E Huang, Yifan Liu, Luc Van Gool, and Fisher Yu. Video task decathlon: Unifying image and video tasks in autonomous driving. In *IEEE/CVF International Conference on Computer Vision*, pages 8647–8657, 2023. 2

[30] Instagram. Instagram. Retrieved August 12, 2023, from https://www.instagram.com. 3

[31] Insurance Institute for Highway Safety, Highway Loss Data Institute. Fatality facts 2021 bicyclists. https://www.iihs.org/topics/fatality-statistics/detail/bicyclists, 2023. 1

[32] Irene Isaksson-Hellman. A study of bicycle and passenger car collisions based on insurance claims data. In *Annals of Advances in Automotive Medicine/Annual Scientific Conference*, volume 56, page 3, 2012. 1

[33] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 3, 8

[34] Nazli Kaya, Joelle Girgis, Braden Hansma, and Birsen Donmez. Hey, watch where you're going! an on-road study of driver scanning failures towards pedestrians and cyclists. *Accident Analysis & Prevention*, 162:106380, 2021. 1

[35] Shunsuke Kogure, Kai Watabe, Ryosuke Yamada, Yoshimitsu Aoki, Akio Nakamura, and Hirokatsu Kataoka. Age should not matter: Towards more accurate pedestrian detection via self-training. In *Computer Sciences & Mathematics Forum*, volume 3, page 11. MDPI, 2022. 4

[36] Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. *Advances in Neural Information Processing Systems*, 35:9881–9893, 2022. 6

[37] Yang Liu, Zhaoyang Xia, Mengyang Zhao, Donglai Wei, Yuzheng Wang, Siao Liu, Bobo Ju, Gaoyun Fang, Jing Liu, and Liang Song. Learning causality-inspired representation consistency for video anomaly detection. In *ACM International Conference on Multimedia*, pages 203–212, 2023. 2

[38] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 6, 7

[39] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[40] V. Machaca Arceda and E. Laura Riveros. Fast car crash detection in video. In *Latin American Computer Conference*, pages 632–637, 2018. 2, 4

[41] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. *arXiv preprint arXiv:1710.03740*, 2017. 7

[42] Anirudh Muthuswamy, M Ali Akber Dewan, Mahbub Murshed, and Dharamjit Parmar. Driver distraction classification using deep convolutional autoencoder and ensemble learning. *IEEE Access*, 2023. 2

[43] Veronica Radu, Mihai Nan, Mihai Trăscău, David Traian Iancu, Alexandra Stefania Ghiță, and Adina Magda Florea. Car crash detection in videos. In *International Conference on Control Systems and Computer Science*, pages 127–132, 2021. 2

[44] Justus J Randolph. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Online submission*, 2005. 4

[45] Adrià Recasens, Pauline Luc, Jean-Baptiste Alayrac, Luyu Wang, Florian Strub, Corentin Tallec, Mateusz Malinowski, Viorica Pătrăucean, Florent Altché, Michal Valko, Jean-Bastien Grill, Aäron van den Oord, and Andrew Zisserman. Broaden your views for self-supervised video learning. In *IEEE/CVF International Conference on Computer Vision*, pages 1255–1265, 2021. 7

[46] Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. Averitec: A dataset for real-world claim verification with evidence from the web. *Advances in Neural Information Processing Systems*, 36, 2024. 4

[47] Ankit Parag Shah, Jean-Bapstite Lamare, Tuan Nguyen-Anh, and Alexander Hauptmann. Cadp: A novel dataset for cctv traffic camera based accident analysis. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 1–9, 2018. 1, 2

[48] Aparajith Srinivasan, Anirudh Srikanth, Haresh Indrajit, and Venkateswaran Narasimhan. A novel approach for road accident detection using detr algorithm. In *International Conference on Intelligent Data Science Technologies and Applications*, pages 75–80, 2020. 2, 3

[49] Jillian Strauss, Luis F Miranda-Moreno, and Patrick Morency. Mapping cyclist activity and injury risk in a network combining smartphone gps data and bicycle counts. *Accident Analysis & Prevention*, 83:132–142, 2015. 1

[50] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6479–6488, 2018. 2

[51] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *IEEE/CVF conference on computer vision and pattern recognition*, 2020. 2

[52] Tomoyuki Suzuki, Hirokatsu Kataoka, Yoshimitsu Aoki, and Yutaka Satoh. Anticipating traffic accidents with adaptive loss and large-scale incident db. In *IEEE/CVF conference on computer vision and pattern recognition*, pages 3521–3529, 2018. 3, 4

[53] TikTok. TikTok. Retrieved August 12, 2023, from https://www.tiktok.com/. 3

[54] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in Neural Information Processing Systems*, 35:10078–10093, 2022. 2

[55] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 2, 7, 8

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[57] Vimeo. Vimeo. Retrieved August 12, 2023, from https://vimeo.com. 3

[58] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14549–14560, 2023. 2

[59] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems*, 34:22419–22430, 2021. 6

[60] X. X. Retrieved August 12, 2023, from https://twitter.com/. 3

[61] Li Xu, He Huang, and Jun Liu. Sutd-trafficqa: A question answering benchmark and an efficient network for video reasoning over traffic events. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9878–9888, 2021. 1, 2, 4

[62] Biao Yang, Weiqin Zhan, Pin Wang, Chingyao Chan, Yingfeng Cai, and Nan Wang. Crossing or not? context-based recognition of pedestrian crossing intention in the urban environment. *IEEE Transactions on Intelligent Transportation systems*, 23(6):5338–5349, 2021. 2

[63] Yu Yao, Xizi Wang, Mingze Xu, Zelin Pu, Ella Atkins, and David Crandall. When, where, and what? a new dataset for anomaly detection in driving videos. *arXiv preprint arXiv:2004.03044*, 2020. 1, 2, 3, 4, 8

[64] Yu Yao, Mingze Xu, Yuchen Wang, David J. Crandall, and Ella M. Atkins. Unsupervised traffic accident detection in first-person videos. In *IROS*, pages 273–280, 2019. 1, 2, 3

[65] YouTube. YouTube. Retrieved August 12, 2023, from https://www.youtube.com/. 3

[66] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2

[67] Yong Zhang, Junjie Li, Yaohua Guo, Chaonan Xu, Jie Bao, and Yunpeng Song. Vehicle driving behavior recognition-based on multi-view convolutional neural network with joint data augmentation. *IEEE Transactions on Vehicular Technology*, 68(5):4223–4234, 2019. 2

[68] Zhen Zhang, Yongming Han, Bo Ma, Min Liu, and Zhiqiang Geng. Temporal chain network with intuitive attention mechanism for long-term series forecasting. *IEEE Transactions on Instrumentation and Measurement*, 72:1–13, 2023. 6

[69] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International Conference on Machine Learning*, pages 27268–27286, 2022. 6