# Traffic Accident Benchmark for Causality Recognition

Tackgeun You[1,2] and Bohyung Han[2]

[1] Dept. of CSE, POSTECH, Korea
[2] Dept. of ECE & ASRI, Seoul National University, Korea
tackgeun.you@postech.ac.kr bhhan@snu.ac.kr

**Abstract.** We propose a brand new benchmark for analyzing causality in traffic accident videos by decomposing an accident into a pair of events, cause and effect. We collect videos containing traffic accident scenes and annotate cause and effect events for each accident with their temporal intervals and semantic labels; such annotations are not available in existing datasets for accident anticipation task. Our dataset has the following two advantages over the existing ones, which would facilitate practical research for causality analysis. First, the decomposition of an accident into cause and effect events provides atomic cues for reasoning on a complex environment and planning future actions. Second, the prediction of cause and effect in an accident makes a system more interpretable to humans, which mitigates the ambiguity of legal liabilities among agents engaged in the accident. Using the proposed dataset, we analyze accidents by localizing the temporal intervals of their causes and effects and classifying the semantic labels of the accidents. The dataset as well as the implementations of baseline models are available in the code repository [3].

## 1 Introduction

Developing an autonomous driving system is one of the major problems in artificial intelligence. This problem has long been viewed as an extremely challenging task since it requires high-level scene understanding in addition to various low-level recognitions. Despite such challenges, autonomous driving has drawn wide attention, and significant improvement has been achieved over the past few years by virtue of advances in computer vision technologies.

Autonomous driving provides convenience to drivers, however, it also raises concerns about traffic accidents, creating the following needs. First, autonomous driving systems should be able to anticipate accidents, take a series of actions to mitigate fatalities, and help drivers escape from the accident. Second, they need to provide an interpretable reasoning process for an accident and deal with liability issues between self-driving vehicles, their manufacturers, passengers, and insurance companies.

[3] https://github.com/tackgeun/CausalityInTrafficAccident

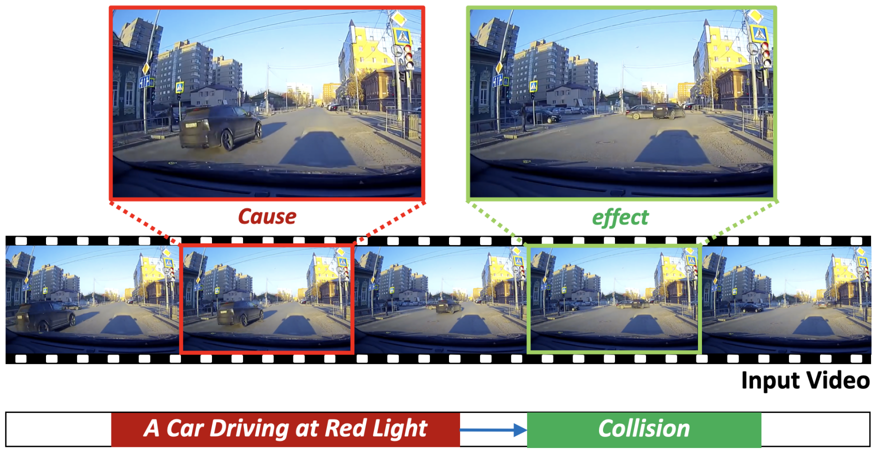


**Fig. 1.** An example of traffic accident video in our dataset, which is associated with a cause (in red) and an effect (in green) events. Both the cause and the effects have their semantic labels: *a car driving at red light* in the cause and *a collision between two vehicles* in the effect.

Despite various issues in autonomous driving systems, the research related to traffic accident analysis is rarely explored due to the following two reasons. First, it is challenging to construct a comprehensive video dataset with traffic accidents due to huge variations in the characteristics of accidents and the environment of traffic scenes. Second, the categories of traffic accidents are ill-defined while the diversity of dataset is crucial to learn robust models for accident recognition.

With the challenges, a few methods [6,13,22,28] mainly focus on the accident anticipation task that aims at forecasting accidents a few seconds earlier. However, the methods simply predict accidents without sophisticated analysis and potential to be extended to accident avoidance systems. On the other hand, Najm *et al.* [19] analyze a traffic accident as a composition of a *cause* and an *effect* event, based on the real-world traffic accident statistics reported by the police. According to [19], an effect event corresponds to the time interval that a vehicle is engaged to an accident, while a cause event means a precrash behavior of the vehicle that potentially leads to an accident. For example, at a road junction, a cause of *'a car driving at red light'* may result in an effect of *'a collision between two vehicles'* as illustrated in Fig. 1. Decomposing a traffic accident scene into a cause and an effect has advantages beyond simple accident anticipation in autonomous driving. First, identifying semantic labels for cause or effect in an accident provides atomic cues for accident analysis and future action planning. Second, the interpretability given by predicting cause and effect events can deal with liability issues between multiple agents.

Motivated by such advantages, we constructed a novel video dataset for causality analysis in traffic accident scenes, which is referred to as CTA (Causality in Traffic Accident). We collected 1,935 videos of traffic accidents, which are captured by dashcams or monitoring cameras, from video repositories on the web. We annotate the semantic labels of cause and effect and their temporal intervals in each accident video. The detailed information of the semantic labels, including their kinds and distributions, is presented in Fig. 2.

Based on the traffic accident analysis dataset, we propose a novel task, temporal cause and effect event localization. As illustrated in Fig. 1, given a video including a traffic accident, the task aims to localize temporal intervals of cause and effect events as well as to identify their semantic labels, simultaneously. To deal with the problem, we adopt several action recognition algorithms—action detection and segmentation—as baseline methods. Experimental results show that modeling long-range contextual information is critical to achieve competitive performance for the localization of cause and effect events.

The main contributions of this paper are summarized below.

- We introduce a traffic accident analysis benchmark, denoted by CTA, which contains temporal intervals of a cause and an effect in each accident and their semantic labels provided by [19].
- We construct the dataset based on the semantic taxonomy in crash avoidance research [19], which makes the distribution of the benchmark coherent to the semantic taxonomy and the real-world statistics.
- We analyze traffic accident tasks by comparing multiple algorithms for temporal cause and effect event localization.

The rest of the paper is organized as follows. We first discuss the related work about traffic accident analysis in Section 2. Section 3 describes the procedure of dataset construction and the statistics of the collected dataset. Section 4 presents the analysis of our dataset using cause and effect event localization algorithms. We summarize the paper and discuss future works in Section 5.

## 2 Related Work

### 2.1 Traffic Accident Anticipation

Chan *et al.* [6] introduce the accident anticipation task with the Street Accident dataset that contain videos captured by dashcams. They propose an LSTM-based model with spatial attention module to estimate the likelihood of accident occurrence in the near future for each frame. Zeng *et al.* [28] propose a multi-task learning approach to improve accident anticipation accuracy, which also localizes risky regions associated with accidents. Herzig *et al.* [13] present the Collision dataset, which includes near-miss incident scenes in addition to accident videos. They propose spatio-temporal action graphs that effectively model the relationship between objects associated with an accident. Kataoka *et al.* [14,22] introduce a large-scale dataset for accident anticipation, referred to as near-miss incident database (NIDB), and propose an adaptive loss function to facilitate the earliest anticipation of an accident.

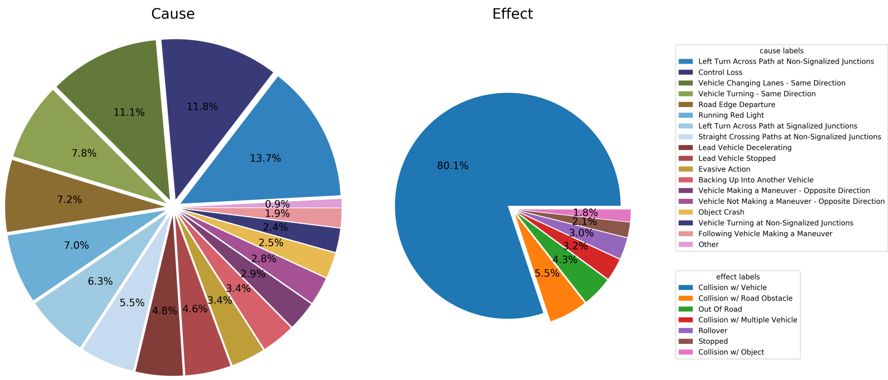


**Fig. 2.** The distributions of semantic labels for cause (left) and effect (right) events. According to the figure, 80.1% of accidents in this dataset are related with the collision of two vehicles.

On the other hand, Yao *et al.* [27] propose a traffic accident detection method based on self-supervision about the future location of vehicles in a scene. By exploiting an additional dataset, their approach manages to outperform [6] without the temporal location supervision of accidents.

There exist a few datasets based on synthetic videos obtained from the GTA5 game to reduce accident video collection cost. Aliakbarian *et al.* [3] collect synthetic driving videos with several scenarios including traffic accidents at scale by constructing a simulator. Kim *et al.* [15] introduce a domain adaptation benchmark for accident anticipation by collecting real and synthetic traffic accident videos. Our traffic accident analysis could take advantage of synthetic datasets by generating videos at a lower cost. However, it is not straightforward to simulate the real distribution of accident cause and effect and generate diverse videos relevant to our objective without sophisticated curation during the dataset construction process.

The accident anticipation task is limited to predicting the occurrence of an accident without its semantic understanding. In contrast, we focus on more challenging tasks—localizing cause and effect events of an accident and estimating semantic labels of an accident—and expect our research to facilitate in-depth analysis of traffic accident scenes.

### 2.2 Causality in Visual Domain

Causality indicates influence by which an event contributes to the generation of another one, where the former event is referred to as the cause and the latter one is referred to as the effect. The simplest mathematical expression for causality is a bivariate model, which consists of a single cause variable, a single effect

**Table 1.** Comparison of traffic accident datasets. The asterisk (*) indicates averaged duration. The triangle (△) means that only effect type is provided.

| Dataset | # of accidents | Causality | Semantic labels | Duration (sec) | Accident type | Video source |
|---|---|---|---|---|---|---|
| VIENA$^2$ [3] | ∼1,200 | - | △ | 5 | synthetic | GTA5 game |
| GTACrash [15] | 7,720 | - | - | 2 | synthetic | GTA5 game |
| Street Accident [6] | 678 | - | △ | 5 | real | Youtube (dashcam) |
| NIDB [22,14] | 4,595 | - | - | 10–15 | near-miss | Mounted on taxi |
| Collision [13] | 803 | - | △ | *40 | real+near-miss | Dashcam |
| YouTubeCrash [15] | 122 | - | - | 2 | real | Youtube (dashcam) |
| CTA (ours) | 1,935 | ✓ | ✓ | *17.7 | real | Youtube |

variable, and a directed edge from the cause to the effect. Research on causality often addresses properties of the directed edge, which describe the causal relation between cause and effect variables.

Lopez-Paz *et al.* [18] propose a binary classifier that identifies whether given two variables $X$ and $Y$ have a causal ($X \rightarrow Y$) or an anti-causal ($Y \rightarrow X$) relation. Based on the binary classifier, they reveal causal relationships between object presence and visual features. Causality in videos is explored in [20,25], where they both aim to classify whether a video is played in a forward or a backward direction. Lebeda *et al.* [17] propose a statistical tool to analyze causality by separating camera motion from the observed one in a scene.

In contrast to the prior works exploring causal relationships, our novelty lies in addressing causality to represent and analyze traffic accident videos—how videos are decomposed, what types of accidents happen, and which prior events trigger the accidents.

### 2.3 Action Understanding

Action understanding algorithm is a core component for video understanding, visual surveillance. and autonomous driving. We review action classification and localization tasks in this subsection.

**Action classification** This task, also referred to as action recognition, categorizes an input video into one or more semantic action classes. There have been a lot of works for this problem, which are often related to video representation learning. Primitive video representation learning methods for action classification include two-stream networks [21,10], C3D [23], 3D-ResNet [11], and I3D [5] while TSN [24] learns augmented representations on top of the standard methods by sparse and uniform sampling of video segments.

**Action localization** The objective of action localization is to identify action class labels and their temporal intervals in a video. There are three kinds of mainstream algorithms—proposal-based action detection, single-stage action detection, and temporal action segmentation; they commonly follow the successful design practicess in the image domain.

Proposal-based action detection [29,30,26,7] first extracts proposals for the temporal regions that are likely to have action instances, and then classifies the individual proposals. Single-stage action detection [4] estimates action intervals from the predefined temporal anchors by regression. Contrary to the two detection-based methods predicting an action label per temporal interval, action segmentation methods [9,16] perform frame-level prediction and obtain temporal information of actions. Note that we tested all three types of action localization methods on our traffic accident analysis benchmark; action segmentation methods are often designed to capture long-range temporal dependency effectively, and results in superior performance in our dataset.

## 3 Traffic Accident Dataset for Causality Understanding

This section describes how we collected the traffic accident dataset for causality understanding, and presents its statistics.

### 3.1 Semantic Taxonomy of Traffic Accident

We constructed a unique dataset, CTA, based on the semantic taxonomy from the report of precrash typology [19], which specifies causes and effect events of accidents observed in the real-world[4]. The prior works related to traffic accident [6,22,13,3,15,27] have little consideration about semantic taxonomy of accidents, and often suffer from intrinsic biases in datasets.

We decompose a traffic accident into a matching pair of events—a cause and an effect. Following the concept of causality, the cause event of an accident corresponds to risky behavior of an agent, such as a vehicle and a pedestrian, that may lead to the accident. On the other hand, the effect event of an accident is related to physical damage of the agents involved in the accident. In principle, a single accident may have multiple causes because many agents can contribute to the accident, but all videos in our dataset contain only a single pair of cause and event.

For each cause and effect event, we assign semantic labels, which correspond to the specific activities of agents that eventually result in accidents. The semantic labels are obtained from the real-world statistics [19].

Our dataset is constructed based on the semantic taxonomy described above; each video has annotation for a cause and an effect, which are associated with semantic labels. Fig. 1 illustrates the relationship between a video, a pair of cause and effect events, and semantic labels. The list of semantic labels of cause and effect events of our benchmark is shown in Figure 4 and Table 4.

[4] 2004 General Estimates System (GES) crash database [1] contains a nationally representative sample of police reports dealing with all types of a vehicle crash.

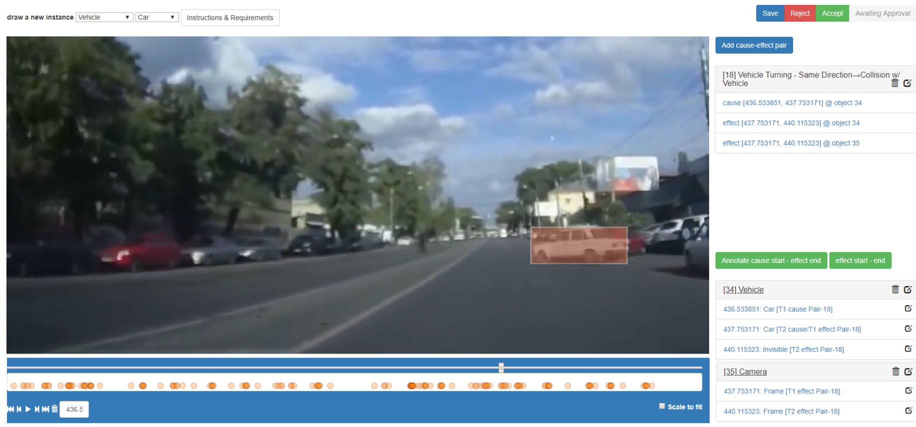


**Fig. 3.** An example view of our annotation tool that supports spatio-temporal annotations. We maintain two kinds of annotations: one for cause and effect events (upper right) and the other for object instances (lower right). The bounding box for the 34$^{\text{th}}$ instance denotes the spatio-temporal start position of the cause of the 18$^{\text{th}}$ accident, which has a cause semantic label *'vehicle turning'* and an effect label *'Collision with Vehicle'*.

### 3.2 Construction of Dataset

**Collecting accident videos** To acquire diverse types of traffic accident scenes, we collect traffic accident videos downloaded from several Youtube channels. Because a single Youtube video may contain multiple traffic accidents, we split the video into distinct sub-clips without shot changes and make each sub-clip associated with only a single accident. The sub-clip split process consists of the following two steps; 1) initial shot boundaries are obtained using a built-in shot boundary detector in FFmpeg[5], and 2) wrong shot boundaries are eliminated and re-annotated manually. Given the sub-clips, we perform an additional filtering step to exclude 1) videos in low resolutions, 2) videos zoomed in or out near the moment of the accident, 3) videos that have ambiguity in determining semantic labels, and 4) too complex videos having multiple cause and effect events. By applying the procedures described above, 59.8% of accident videos are survived. Eventually, 1,935 videos (corresponding to 9.53 hours) with only a single traffic accident remain in our dataset. They are split into 1,355 (70%), 290 (15%) and 290 (15%) videos for train, validation and testing, respectively. Most of the videos in our dataset are captured by dashcam while there exists a small fraction (∼12%) of videos from monitoring cameras.

**Annotating videos** Fig. 3 illustrates our annotation tool. We annotate temporal intervals and semantic labels of cause and effect in traffic accidents via the

[5] https://ffmpeg.org/

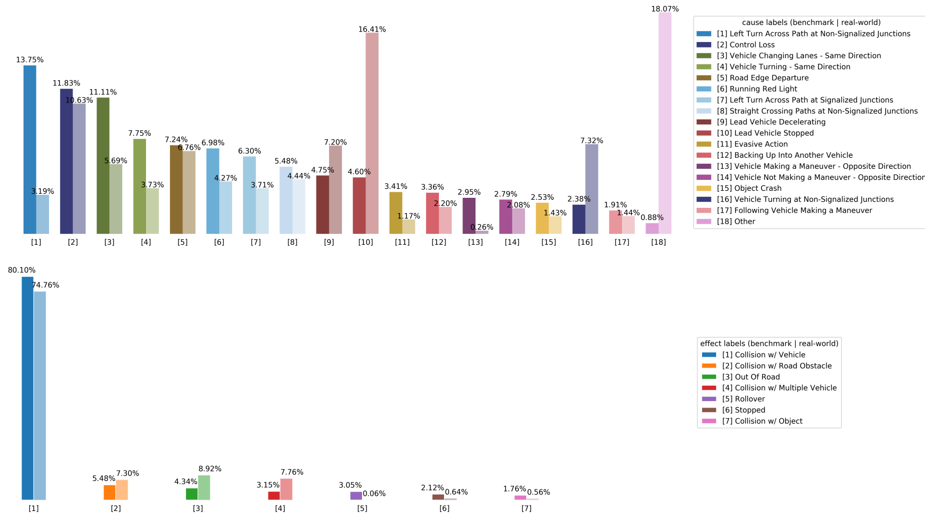


**Fig. 4.** The distribution of semantic labels for cause (top) and effect (bottom) observed in the proposed dataset and the real-world statistics. For each semantic label, the darker bar denotes our benchmark dataset while the lighter one indicates the real-world.

following two steps; 1) we determine the semantic labels of cause and effect in each candidate video, and 2) we annotate the temporal intervals for cause and effect events. After that, the temporal intervals are adjusted for their consistency with the semantic labels. Note that the semantic label estimation solely depends on visual information because no other information is accessible.

We annotate an effect event first and the label of a cause event is conditioned on the one corresponding to the matching effect. The start time of an effect is the moment that a vehicle begins to suffer any physical damage while its end time corresponds to the frame at which there is no more event happening to all the involved objects. On the other hand, the duration a cause is from the frame that a vehicle starts any abnormal movements or wrongdoings to the moment that such atypical activities end. In practice, the end time of a cause is often ambiguous and annotated as the same time with the start time of an effect.

### 3.3 Statistics of Our Dataset

Fig. 4 demonstrates the distribution of semantic labels of cause in both our benchmark dataset and the real-world statistics. Our dataset covers 18 semantic labels in cause and 7 semantic labels in effect while the distributions of semantic labels in the benchmark and the real-world are roughly consistent. However, the distributions are particularly different for *other* class. This is partly because we use a subset of semantic classes in the real-world data for the construction of our benchmark dataset and all the semantic classes missing in our dataset now belong to *other* class. The missing classes are mostly related to violent accidents such as collisions with pedestrians or animals; they are removed from YouTube. Although they may induce unwanted bias in the dataset, our dataset is constructed based on the semantic taxonomy with the real-world statistics.

## 4 Traffic Accident Benchmark

We demonstrate the task—temporal cause and effect events classification and localization—for traffic accident benchmark and introduce the evaluation method with simulating real-world which exploits real-world distribution for performance evaluation.

### 4.1 Temporal Cause and Effect Events Recognition

The main target task of our dataset is temporal cause and effect event recognition, which consists of two subtasks, classification and localization. The classification task aims to identify semantic labels for each cause and effect event while the objective of the localization task to estimate the temporal interval for each cause and effect event. Compared to the standard action recognition task, where each action or its instance is predicted independently, our problem need to consider temporal constraints of cause and effect—the cause event always precedes the effect event—and understand causal relation of the two events—the dynamics of vehicles is consistent with the causality of the accident.

### 4.2 Baselines

We adopt temporal segment networks (TSN) [24] as the baseline algorithm for action classification, where two consensus functions, average and linear function, are utilized for evaluation. For action localization, three baselines with unique characteristics are tested. The first baseline is Single-Stream Temporal Action Proposals (SST) [4], which employs a Gated Recurrent Units (GRU) [8] to classify a label for each proposal corresponding to a video segment. For this baseline, we train two additional variant models by replacing the forward GRU with a backward GRU (Backward SST) and bi-directional GRU (Bi-SST). The second option is R-C3D [26], which is a simple extension of R-CNN for object detection; it detects actions by proposal generation followed by classification. The third one is Multi-Stage Temporal Convolutional Network (MS-TCN) [9], which consists of repeated building blocks of Single Stage Temporal Convolutional Network (SS-TCN). SS-TCN consists of 1D dilated convolutions to model long-range dependencies and perform frame-level dense predictions.

We use I3D [5] RGB stream for our video representation of all baselines. The detailed architectures of all baselines and their training details (*e.g.*, learning rate, hyper-parameters, etc.) are described in the code repository.

**Table 2.** Performance comparisons of action classification methods.

| Method | Top-1 mean accuracy (%) | | | Top-2 mean accuracy (%) | | |
|---|---|---|---|---|---|---|
| | cause | effect | mean | cause | effect | mean |
| Trivial Prediction | 13.7 | 80.1 | 46.9 | 25.5 | 85.6 | 55.6 |
| TSN [24] (average) | 18.8 | 43.8 | 31.3 | 31.3 | 87.5 | 59.4 |
| TSN (linear) | 31.3 | 87.5 | 59.4 | 37.5 | 93.8 | 65.7 |

### 4.3 Evaluation Metrics

**Classification accuracy** We use the standard metric for the evaluation of classification methods. Note that we perform classification over semantic labels and the accuracies of individual classes are averaged to report the final score.

**Accuracy with temporal IoU** We adopt the "accuracy" at a temporal Intersection over Union (tIoU) threshold, which measures the percentage of the predictions that have tIoUs larger than the threshold. Given the tIoU threshold $\tau$, the accuracy is defined by

$$\text{accuracy}^{\tau} = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1} \left[ \frac{(\text{prediction}_n \cap \text{gt}_n)}{(\text{prediction}_n \cup \text{gt}_n)} > \tau \right], \tag{1}$$

where gt denotes ground-truth interval and $N$ is the number of examples, and $\mathbb{1}[\cdot]$ is an indicator function.

We use top-1 accuracy for evaluation, which takes the temporal interval with the highest score as the prediction.

**Evaluation with prior distribution** We also evaluate the algorithms based on the weighted accuracy, which leverages the real-world distribution of semantic classes as prior information. This is possible because our benchmark dataset is constructed based on the real-world accident distribution. The weighted mean accuracy is computed by

$$\text{weighted mean accuracy}^{\tau} = \sum_{c \in \mathcal{C}} w_c \cdot \text{accuracy}_c^{\tau}, \tag{2}$$

where $w_c$ indicates the frequency of a semantic class $c$ in the real-world. If $w_c$ is identical to all $c$'s, the metric is equivalent to (unweighted) mean accuracy. Note that the weighted mean accuracy penalizes predictions for labels that rarely happen in the real-world and this weighting scheme is crucial for planning actions to mitigate the fatalities.

### 4.4 Analysis of Action Classification Performance

**Action classification** We compared between TSN with two consensus functions and the trivial prediction, which outputs the most frequent semantic labels

**Table 3.** Performances of the baseline algorithms for temporal cause and effect event localization on the test set of our dataset (CTA).

| Algorithm type | Method | Accuracy (%) at a tIoU threshold | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | tIoU > 0.3 | | | tIoU > 0.5 | | | tIoU > 0.7 | | |
| | | cause | effect | mean | cause | effect | mean | cause | effect | mean |
| Trivial prediction | Un-normalized | 13.45 | 26.55 | 20.00 | 9.31 | 15.52 | 12.42 | 2.41 | 4.48 | 3.45 |
| | Normalized | 21.72 | 37.24 | 29.48 | 11.38 | 19.66 | 15.52 | 2.41 | 6.90 | 4.66 |
| Single-stage detection | SST [4] | 23.45 | 31.72 | 27.59 | 17.24 | 17.24 | 17.24 | 6.90 | 6.55 | 6.72 |
| | Backward SST | 30.00 | 44.83 | 37.41 | 17.93 | 24.83 | 21.38 | 5.17 | 6.21 | 5.69 |
| | Bi-SST | 29.66 | 54.48 | 42.07 | 17.24 | 27.24 | 22.24 | 5.17 | 10.00 | 7.59 |
| | SS-TCN + SST | 32.41 | 48.97 | 40.69 | 20.00 | 30.00 | 25.00 | 9.31 | 12.76 | 11.03 |
| Single-stage detection | R-C3D [26] | 36.21 | 58.62 | 47.41 | 22.07 | 38.28 | 30.17 | 8.62 | 13.10 | 10.86 |
| Segmentation | SS-TCN [9] | 38.28 | 54.97 | 46.62 | 23.86 | 36.48 | 30.17 | 10.55 | 17.10 | 13.83 |
| | MS-TCN [9] | 41.45 | 57.45 | 49.45 | 28.07 | 37.86 | 32.97 | 11.10 | 17.72 | 14.41 |

of cause and effect event. According to Table 2, TSN with average consensus function is worse than the trivial prediction. In contrast, TSN with linear consensus function outperforms the trivial prediction. This is because the average consensus function ignores the temporal order of video frames while the linear function preserves the temporal order of frame features in a video.

### 4.5 Analysis of Action Localization Performance

**Trivial prediction via averaging temporal intervals** To evaluate the performance of the baseline algorithms, we computed the accuracy of the trivial prediction, which is given by the average interval of all ground-truths for cause and effect in the training dataset. We compute the average intervals for both unnormalized and normalized videos, where the normalization means the equalization of video lengths. Table 3 presents that the trivial methods are not successful in most cases compared to the baselines.

**Variants of single-stage detection methods** Table 3 also presents that effect localization performance is sensitive to the choice of GRUs in SST while cause localization is relatively stable in the direction of GRU placement. We observe that the contextual information from future frames, which can be acquired by backward SST better, is crucial for recognizing effect events. Exploiting both contextual information from past and future frames as in Bi-SST delivers the best performance.

**Detection vs. segmentation** The methods for temporal segmentation such as SS-TCN and MS-TCN tend to achieve better performance than detection-based techniques, especially at high tIoU thresholds, although the proposal-based detection method, R-C3D is comparable to segmentation-based approaches. This is partly because action localization methods based on detection are designed to

**Table 4.** Localization performance of MS-TCN for individual semantic classes. The semantic labels of cause and effect are sorted in a descending order of frequency.

| | Semantic label | Accuracy (%) | | |
|---|---|---|---|---|
| | | tIoU > 0.3 | tIoU > 0.5 | tIoU > 0.7 |
| Cause | [1] Left turn across path at non-signalized junctions | 38.71 | 25.81 | 12.90 |
| | [2] Control Loss | 74.19 | 35.48 | 12.90 |
| | [3] Vehicle changing lanes: same direction | 17.24 | 13.79 | 6.90 |
| | [4] Vehicle turning: same direction | 44.00 | 36.00 | 16.00 |
| | [5] Road edge departure | 45.83 | 33.33 | 25.00 |
| | [6] Running red light | 31.58 | 26.32 | 21.05 |
| | [7] Left turn across path at signalized junctions | 56.25 | 50.00 | 25.00 |
| | [8] Straight crossing paths at non-signalized junctions | 50.00 | 31.82 | 4.55 |
| | [9] Lead vehicle decelerating | 35.29 | 35.29 | 29.41 |
| | [10] Lead vehicle stopped | 33.33 | 26.67 | 13.33 |
| | [11] Evasive action | 50.00 | 30.00 | 20.00 |
| | [12] Backing up into another vehicle | 50.00 | 25.00 | 12.50 |
| | [13] Vehicle making a maneuver: opposite direction | 37.50 | 25.00 | 25.00 |
| | [14] Vehicle not making a maneuver: opposite direction | 66.67 | 66.67 | 33.33 |
| | [15] Object crash | 33.33 | 8.33 | 8.33 |
| | [16] Vehicle turning at non-signalized junctions | 37.50 | 37.50 | 12.50 |
| | [17] Following vehicle making a maneuver | 40.00 | 20.00 | 20.00 |
| | [18] Other | 50.00 | 25.00 | 0.00 |
| | Mean accuracy | 43.97 | 30.67 | 16.59 |
| | Weighted mean accuracy (benchmark) | 43.45 | 30.00 | 15.86 |
| | Weighted mean accuracy (real-world) | 44.51 | 30.31 | 13.39 |
| Effect | [1] Collision with vehicle | 56.83 | 36.56 | 18.06 |
| | [2] Collision with road obstacle | 43.75 | 12.50 | 6.25 |
| | [3] Out of road | 80.00 | 66.67 | 20.00 |
| | [4] Collision with multiple vehicle | 44.44 | 22.22 | 11.11 |
| | [5] Rollover | 44.44 | 33.33 | 22.22 |
| | [6] Stopped | 66.67 | 50.00 | 33.33 |
| | [7] Collision with object | 75.00 | 37.50 | 25.00 |
| | Mean accuracy | 58.73 | 36.96 | 19.42 |
| | Weighted mean accuracy (benchmark) | 57.24 | 36.55 | 17.93 |
| | Weighted mean accuracy (real-world) | 57.14 | 36.47 | 16.97 |

identify multiple events in a video while the videos in our dataset contain only a single instance of cause and effect.

**GRU vs. stack of dilated convolutions** To verify the effectiveness of SS-TCN without the advantage of action segmentation over action detection, we tested the accuracy of SST after replacing GRU in SST by SS-TCN, which is denoted by SS-TCN+SST. SS-TCN+SST outperforms all variants of SST with large margins at all tIoU thresholds as presented in Table 3. Note that while Bi-directional GRU is capable of modeling long-range dependencies, it turns out that stacking of 1D dilated convolutions is more effective.

**Localization performance of individual semantic classes** Table 4 shows the localization performance of individual semantic classes, where the results from MS-TCN are reported. Note that, in addition to the naïve mean accuracy, we present two versions of weighted mean accuracy; they are differentiated by whether the weights are from the real-word distribution or the sampled distribution in our benchmark.

**Qualitative results** Fig. 5 illustrates the qualitative results of predictions given by the compared algorithms.

## 5 Discussion and Future Works

We introduced a traffic accident benchmark and demonstrated temporal cause and effect event classification and localization performance of several baseline approaches. Our benchmark annotates cause and effect events separately to facilitate research for causality understanding and takes advantage of real-world grounded semantic taxonomy and the associated distribution for building dataset. Our dataset contains 1,935 traffic accident videos, each of which is annotated with a pair of temporal intervals of cause and effect with their semantic labels.

Spatio-temporal cause and effect localization would be a straightforward extension of our work towards capturing object-level cause and effect information, but it requires additional annotations for individual objects in videos. In the current version of the dataset, we discard the traffic accident videos with the ambiguous semantic labels for cause and effect events. Also, there exists only a single semantic label for each cause and effect event, and additional efforts should be made for the construction of the more comprehensive dataset.

**Acknowledgement** This work was supported by Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) [2017-0-01780, 2017-0-01779] and Microsoft Research Asia. We also appreciate Jonghwan Mun and Ilchae Jung for valuable discussion.

## References

1. National automotive sampling system (nass) general estimates system (ges) analytical user's manual 1988-2004 (2005), `https://one.nhtsa.gov/Data/National-Automotive-Sampling-System-(NASS)`
2. American time use survey (2013), `https://www.bls.gov/tus/`
3. Aliakbarian, M.S., Saleh, F.S., Salzmann, M., Fernando, B., Petersson, L., Andersson, L.: Viena2: A driving anticipation dataset. In: ACCV (2018)
4. Buch, S., Escorcia, V., Shen, C., Ghanem, B., Niebles, J.C.: SST: Single-stream temporal action proposals. In: CVPR (2017)
5. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR (2017)
6. Chan, F.H., Chen, Y.T., Xiang, Y., Sun, M.: Anticipating accidents in dashcam videos. In: ACCV (2016)

**Fig. 5.** The qualitative localization results of the 7 baselines algorithms placed with same order in table 3 and the ground-truths. A bar indicates the duration of a cause or an effect while it has the following color codes: temporal interval of predicted cause, predicted effect, predicted cause and effect (overlapped) ground-truth cause and ground-truth effect.

7. Chao, Y.W., Vijayanarasimhan, S., Seybold, B., Ross, D.A., Deng, J., Sukthankar, R.: Rethinking the faster R-CNN architecture for temporal action localization. In: CVPR (2018)
8. Cho, K., van Merrienboer, B., Gulcehre, C., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. In: EMNLP (2014)
9. Farha, Y.A., Gall, J.: Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In: CVPR (2019)
10. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: CVPR (2016)
11. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: CVPR. pp. 6546–6555 (2018)
12. Heilbron, F.C., Escorcia, V., Ghanem, B., Niebles, J.C.: Activitynet: A large-scale video benchmark for human activity understanding. In: CVPR (2015)
13. Herzig, R., Levi, E., Xu, H., Gao, H., Brosh, E., Wang, X., Globerson, A., Darrell, T.: Spatio-temporal action graph networks. In: ICCVW (2019)
14. Kataoka, H., Suzuki, T., Oikawa, S., Matsui, Y., Satoh, Y.: Drive video analysis for the detection of traffic near-miss incidents. In: ICRA (2018)
15. Kim, H., Lee, K., Hwang, G., Suh, C.: Crash To Not Crash: Learn to identify dangerous vehicles using a simulator. In: AAAI (2019)
16. Lea, C., Flynn, M.D., Vidal, R., Reiter, A., Hager, G.D.: Temporal convolutional networks for action segmentation and detection. In: CVPR (2017)
17. Lebeda, K., Hadfield, S., Bowden, R.: Exploring causal relationships in visual object tracking. In: ICCV (2015)
18. Lopez-Paz, D., Nishihara, R., Chintala, S., Schölkopf, B., Bottou, L.: Discovering causal signals in images. In: CVPR (2017)
19. Najm, W.G., Smith, J.D., Yanagisawa, M.: Pre-crash scenario typology for crash avoidance research (2007), https://rosap.ntl.bts.gov/view/dot/6281
20. Pickup, L.C., Pan, Z., Wei, D., Shih, Y., Zhang, C., Zisserman, A., Schölkopf, B., Freeman, W.T.: Seeing the arrow of time. In: CVPR (2014)
21. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: NIPS (2014)
22. Suzuki, T., Kataoka, H., Aoki, Y., Satoh, Y.: Anticipating traffic accidents with adaptive loss and large-scale incident db. In: CVPR (2018)
23. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV (2015)
24. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Val Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: ECCV (2016)
25. Wei, D., Lim, J., Zisserman, A., Freeman, W.T.: Learning and using the arrow of time. In: CVPR (2018)
26. Xu, H., Das, A., Saenko, K.: R-c3d: region convolutional 3d network for temporal activity detection. In: ICCV (2017)
27. Yao, Y., Xu, M., Wang, Y., Crandall, D.J., Atkins, E.M.: Unsupervised traffic accident detection in first-person videos. In: IROS (2019)
28. Zeng, K.H., Chou, S.H., Chan, F.H., Niebles, J.C., Sun, M.: Agent-centric risk assessment: Accident anticipation and risky region localization. In: CVPR (2017)
29. Zeng, R., Huang, W., Tan, M., Rong, Y., Zhao, P., Huang, J., Gan, C.: Graph convolutional networks for temporal action localization. In: ICCV (2019)
30. Zhao, Y., Xiong, Y., Wang, L., Wu, Z., Tang, X., Lin, D.: Temporal action detection with structured segment networks. In: ICCV (2017)