

# When, Where, and What? A New Dataset for Anomaly Detection in Driving Videos

Yu Yao<sup>1</sup>Xizi Wang<sup>2</sup>Mingze Xu<sup>3\*</sup>Zelin Pu<sup>1</sup>Ella M. Atkins<sup>1</sup>David J. Crandall<sup>2</sup><sup>1</sup>University of Michigan    <sup>2</sup>Indiana University    <sup>3</sup>Amazon Rekognition  
{brianyao, ematkins}@umich.edu, {xiziwang, djcran}@indiana.edu

**Abstract.** Video anomaly detection (VAD) has been extensively studied. However, research on egocentric traffic videos with dynamic scenes lacks large-scale benchmark datasets as well as effective evaluation metrics. This paper proposes traffic anomaly detection with a *when-where-what* pipeline to detect, localize, and recognize anomalous events from egocentric videos. We introduce a new dataset called Detection of Traffic Anomaly (DoTA) containing 4,677 videos with temporal, spatial, and categorical annotations. A new spatial-temporal area under curve (STAUC) evaluation metric is proposed and used with DoTA. State-of-the-art methods are benchmarked for two VAD-related tasks. Experimental results show STAUC is an effective VAD metric. To our knowledge, DoTA is the largest traffic anomaly dataset to-date and is the first supporting traffic anomaly studies across *when-where-what* perspectives. Our code and dataset can be found in: <https://github.com/MoonBlvd/Detection-of-Traffic-Anomaly>

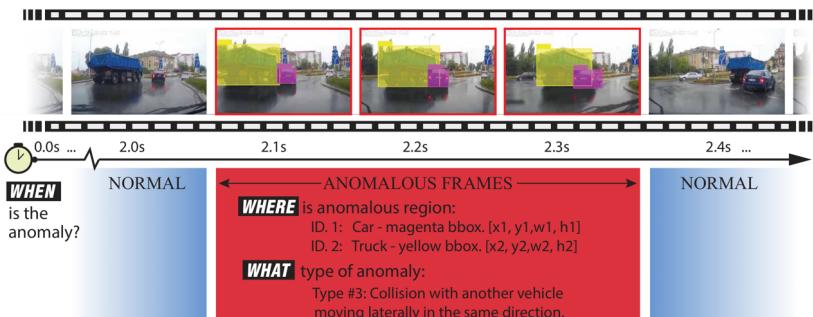


Fig. 1: Overview of DoTA dataset. Annotations are provided to answer the **When**, **Where** and **What** questions for driving video anomalies.

## 1 Introduction

Accurate perception of accident scenarios is a key challenge for advanced driver assistance systems (ADAS). Is an accident going to happen? Who will be involved? What type of accident is it? These critical questions demand detection,

\* This work was done while the author was at Indiana University.

localization, and classification of on-road anomalies for proper reaction and event data recording [42]. We propose accident perception as a **When-Where-What** pipeline: **When** the anomalous event starts and ends, **Where** the anomalous regions are in each video frame, and **What** the anomaly type is. This pipeline can be mapped to two computer vision tasks: video anomaly detection (VAD) and video action recognition (VAR). VAD predicts per-frame anomaly scores to answer the **When** question, and computes per-pixel or per-object anomaly scores as intermediate step to implicitly answer the **Where** question. VAR classifies video type to answer the **What** question.

Training deep learning-based methods for VAD and VAR has been made possible by large-scale labeled datasets. There are video datasets available for surveillance applications of VAD, including CUHK [27], ShanghaiTech Campus [26], and UCF-crime [34], and for human activity recognition for VAR, including Sports-1M [21] and Kinetics [22]. For traffic anomalies, recent first-person video datasets such as StreetAccident [4] and A3D [44] have annotations of anomaly start and end times, while DADA [9] provides human attention maps from video spectator eye-gaze. However, no large-scale dataset and benchmark yet covers the full **When-Where-What** pipeline.

This paper introduces Detection of Traffic Anomaly (DoTA), a large-scale benchmark dataset for traffic VAD and VAR. DoTA contains 4,677 videos with 18 anomaly categories [1] and multiple anomaly participants in different driving scenarios. DoTA provides rich annotation for each anomaly: type (category), temporal annotation, and anomalous object bounding box tracklets. Taking advantage of this large-scale dataset with rich anomalous object annotations, we propose a novel VAD evaluation metric called Spatio-temporal Area Under Curve (STAUC). STAUC is motivated by the popular frame-level Area Under Curve (AUC). While AUC uses a per-frame anomaly score which is usually averaged from a pixel-level or object-level score map, STAUC takes such score map and computes how much of it overlaps with the annotated anomalous region. This overlap ratio is used as a weighting factor for true positive predictions with STAUC. STAUC thus has AUC as its upper bound.

We benchmark existing VAD baselines and state-of-the-art methods on DoTA using both AUC and STAUC. We also propose a simple-but-effective ensemble method that improves the performance of any single approach, offering a new direction to explore. Extensive experiments show the importance of using this new metric in VAD research. To further complete the pipeline, we also benchmark recent VAR methods such as R(2+1)D [37] and SlowFast [10] on DoTA. Experiments show that applying generalized VAR methods to traffic anomaly understanding is far from perfect, motivating more research in this area.

This paper offers three contributions. First, we introduce DoTA, a large-scale ego-centric traffic video dataset to support VAD and VAR; to the best of our knowledge this is the largest traffic video anomaly dataset and the first containing detailed temporal, spatial, and categorical annotations. Second, we identify problems with the commonly-used AUC metric and propose a new spatio-temporal evaluation metric (STAUC) to address them. We benchmark state-of-

the-art VAD methods with both AUC and STAUC and show the effectiveness of our new metric. Finally, we provide benchmarks of state-of-the-art VAR algorithms on DoTA, which we hope will encourage further research to manage challenging ego-centric traffic video scenarios.

## 2 Related Work

**Existing Video Anomaly Detection (VAD) datasets** are generally from surveillance cameras. For example, UCSD Ped1/Ped2 [24], CUHK Avenue [27], and ShanghaiTech [26] were collected from campus surveillance cameras and include anomalies like prohibited objects and abnormal movements, while UCF-Crime [34] includes accidents, robbery, and theft. Anomaly detection in egocentric traffic videos has very recently attracted attention. Chan *et al.* [4] propose the StreetAccident dataset of on-road accidents with 620 video clips collected from dash cameras. The last ten frames of each clip are annotated as anomalous. Yao *et al.* [44] propose the A3D dataset containing 1,500 anomalous videos in which abnormal events are annotated with the start and end times. Fang *et al.* [9] introduce the DADA dataset for driver attention prediction in accidents, while Herzig *et al.* [17] extract a collision dataset with 803 videos from BDD100K [45]. In contrast, our DoTA dataset is much larger (nearly 5,000) but, much more importantly, contains richer annotations that support the whole **When-Where-What** anomaly analysis pipeline.

**Existing VAD models** mainly focus on the **When** problem but are also implicitly related to **Where**. Hasan *et al.* [14] propose a convolutional Auto-Encoder (ConvAE) to model the normality of video frames by reconstructing stacked input frames. Convolutional LSTM Auto-Encoder (ConvLSTMAE) is used in [30,6,28] capture regular visual and motion patterns. Luo *et al.* [29] propose a stacked RNN for temporally-coherent sparse coding (TSC-sRNN). Liu *et al.* [26] detect anomalies by looking for differences between predicted future frames and actual observations. Gong *et al.* [13] propose an MemAE network to query pre-saved memory units for reconstruction, while Wang *et al.* [38] design generalized one-class sub-spaces for discriminative regularity modeling. Other work has recently studied object-centric approaches. Ionescu *et al.* [20] propose K-means to cluster object features and train multiple support vector machine (SVM) classifiers with confidence as anomaly score. Morais *et al.* [31] model human skeleton regularity with local-global autoencoders and compute per-object anomaly scores. VAD in egocentric traffic scenarios is a new and challenging problem due to dynamic foreground and background, perspective projection, and complicated scenes. The most related work to ours is TAD [44], which predicts future object bounding boxes from past time steps with RNN encoder-decoders, where the standard deviation of predictions serves as the anomaly score. We benchmark stat-of-the-art VAD methods and their variants on DoTA dataset.

**Action Recognition methods** address the **What** problem to classify traffic anomalies. Two-stream networks [33] and temporal segment networks (TSN) [39]

Table 1: Comparison of published driving video anomaly datasets.

Dataset	# videos	# frames	Annotations
StreetAccident [4]	620	62,000 (20fps)	temporal
A3D [44]	1,500	128,175 (10fps)	temporal
DADA [9]	2,000	648,476 (30fps)	temporal, spatial (eye-gaze)
<b>DoTA</b>	<b>4,677</b>	<b>731,932 (10fps)</b>	temporal, spatial (tracklets), categories

leverage RGB and optical flow data. Tran *et al.* [36] first proposed 3D convolutional networks (C3D) for spatiotemporal modeling, followed by an inflated model [3]. Recent work substitutes 3D convolution with 2D and 1D convolution blocks (R(2+1)D [37]) to improve effectiveness and efficiency. Feichtenhofer *et al.* [10] propose the SlowFast model to extract video features from low and high frame rate streams. Online action detection in untrimmed, streaming videos is addressed by De Geest *et al.* [8], while Gao *et al.* [11] propose a reinforce encoder-decoder(RED) to tackle action prediction and online action recognition. Shou *et al.* [32] model temporal consistency with a generative adversarial network (GAN). Xu *et al.* [41] propose a temporal recurrent network (TRN) leveraging future prediction to aid online action detection. Gao *et al.* [12] uses reinforcement learning to detect the start time of actions. We benchmark VAR methods on DoTA dataset, and discuss online action detection in supplement.

### 3 The Detection of Traffic Anomaly (DoTA) Dataset

We introduce DoTA, the first publicly-available **When-Where-What** pipeline dataset with temporal, spatial, and categorical annotations.<sup>1</sup> To build DoTA, we collected more than 6,000 video clips from YouTube channels and selected diverse dash camera accident videos from different countries under different weather and lighting conditions. We avoided videos with accidents that were not visible or camera fall-off from wind shield, resulting in 4,677 videos with  $1280 \times 720$  resolution. Though the original videos are at 30 fps, we extracted frames at 10 fps for annotations and experiments in this paper. Table 1 compares DoTA with other ego-centric traffic anomaly datasets.

We annotated the dataset using a custom tool based on Scalabel<sup>2</sup>. Labeling traffic anomalies is subjective, especially for properties like start and end times. To produce high quality annotations, each video was labeled by three annotators, and the temporal and spatial (categorical) annotations were merged by taking average (mode) to minimize individual biases. Our 12 human annotators had different levels of driving experience.

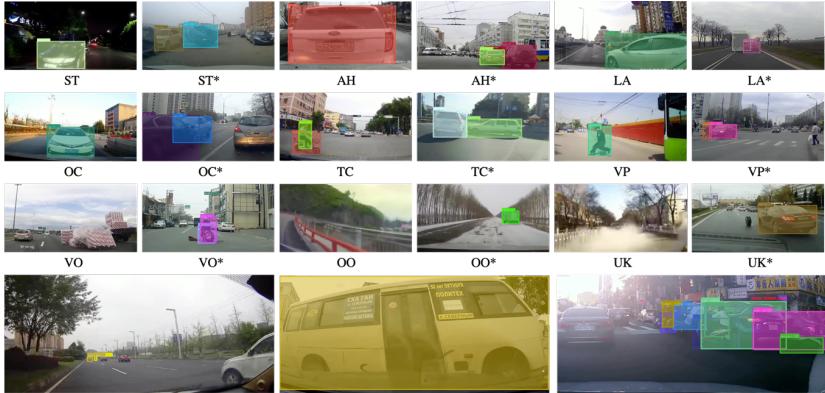
**Temporal Annotations.** Each DoTA video is annotated with anomaly start and end times, which separates it into three temporal partitions: precursor, which is normal video preceding the anomaly, the anomaly window, and post-anomaly, which is normal activity following the anomaly. Duration distributions are shown

<sup>1</sup> The dataset will be made publicly available upon publication.

<sup>2</sup> <https://scalabel.ai/>

Table 2: Traffic anomaly categories in the DoTA dataset

ID	Short	Anomaly Categories
1	ST	Collision with another vehicle which starts, stops, or is stationary
2	AH	Collision with another vehicle moving ahead or waiting
3	LA	Collision with another vehicle moving laterally in the same direction
4	OC	Collision with another oncoming vehicle
5	TC	Collision with another vehicle which turns into or crosses a road
6	VP	Collision between vehicle and pedestrian
7	VO	Collision with an obstacle in the roadway
8	OO	Out-of-control and leaving the roadway to the left or right
9	UK	Unknown



A distant VP\* anomaly, objects are extremely small

A nearby TC anomaly, object is extremely large

A complicated AH\* anomaly, 8 objects involved

Fig. 2: DoTA Samples. Spatial annotations are shown as shadowed bounding boxes. Short anomaly category labels with \* indicate non-ego anomalies.

in Fig. 3(a). Since early detection is essential for on-road anomalies [4,35], we asked the annotators to estimate the anomaly start as the time when the anomaly was inevitable. The anomaly end was meant to be the time when all anomalous objects are out of the field of view or are stationary. Our annotation is different from [9] where a frame is marked as anomaly start if half of the anomaly participant appears in the camera view; such a start time can be too early because anomaly participants often appear for a while before they start to behave abnormally. Our annotation is also distinct from [4] and [44] where the anomaly start is marked when a crash happens, which does not support early detection.

**Spatial Annotations.** DoTA is the first traffic anomaly dataset to provide detailed spatio-temporal annotation of anomalous objects. Each anomaly participant is assigned a unique track ID, and their bounding box is labeled from anomaly start to anomaly end or until the object is out of view. We consider seven common traffic participant categories: person, car, truck, bus, motorcycle, bicycle, and rider, following the BDD100K style [45]. Statistics of object categories and per-video anomalous object numbers are shown in Fig. 3(c) and 3(d). DADA [9] also provides spatial annotations by capturing video observers' eye-

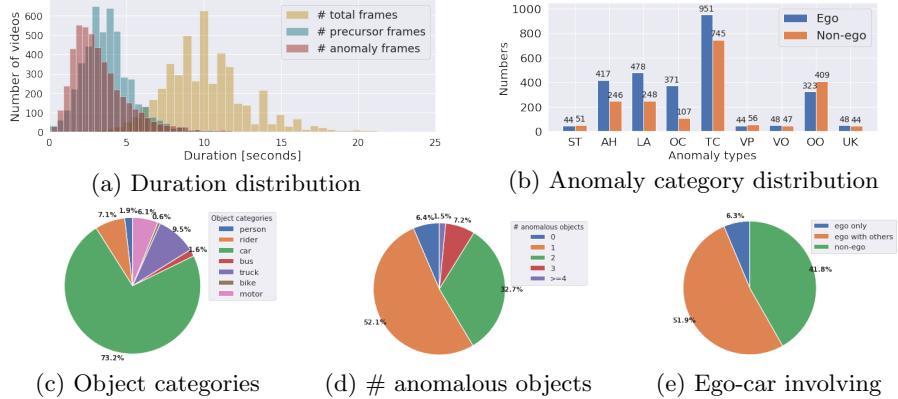


Fig. 3: DoTA dataset statistics.

gaze for driver attention studies. However, they have shown that eye-gaze does not always coincide with the anomalous region, and that gaze can have  $\sim 1$  to 2 seconds delay from anomaly start. Thus our tracklets provide improved annotation for spatio-temporal anomaly detection studies.

**Anomaly Categories.** Each DoTA video is assigned one of the 9 categories listed in Table 2, as defined in [1]. We have observed that the same anomaly category with different viewpoints are visually distinct, as shown in Fig. 2. Therefore we split each category to ego-involved and non-ego (marked with \*), resulting in 18 categories total. Sometimes the category can be ambiguous, particularly when one anomaly is followed by another. For example, an oncoming out-of-control (OO\*) vehicle might result in an oncoming collision (OC) with the ego vehicle. In such cases, we annotate the anomaly category as the dominant one in the video, i.e, the one that lasts longer during the anomaly period. The distribution of videos of each category is shown in Fig. 3(b).

## 4 Video Anomaly Detection (VAD) Methods

We benchmark both unsupervised and supervised VAD. Unsupervised VAD is divided into frame-level and object-centric methods according to different input and output types. Supervised VAD is similar to temporal action detection but outputs a binary label indicating anomaly or no-anomaly.

### 4.1 Frame-level Unsupervised VAD

Frame-level unsupervised VAD methods detect anomalies by either reconstructing past frames or predicting future frames and computing the reconstruction or prediction error. We benchmark three methods and their variants in this paper.

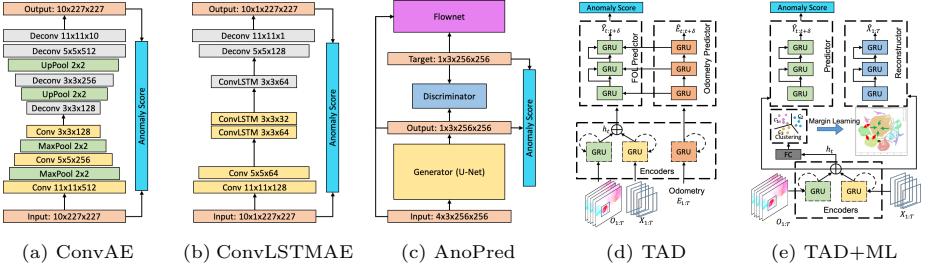


Fig. 4: Network architecture of unsupervised VAD methods.

**ConvAE** [14] is a spatio-temporal autoencoder model which encodes temporally stacked images with 2D convolutional encoders and decodes with deconvolutional layers to reconstruct the input (Fig. 4(a)). The per-pixel reconstruction error forms an anomaly score map  $\Delta I$  and the mean squared error (MSE) is computed as a frame-level anomaly score,

$$MSE = \frac{1}{|M|} \sum_{i \in M} \Delta I(i), \text{ and } \Delta I(i) = \|I(i) - \hat{I}(i)\|^2, \quad (1)$$

where  $I$  and  $\hat{I}$  are the ground truth and reconstructed/predicted frames,  $M$  represents all frame pixels, and  $\Delta I$  is also called anomaly score map. To further compare the effectiveness of image and motion features, we implement **ConvAE(gray)** and **ConvAE(flow)** to reconstruct the grayscale image and the dense optical flow, respectively. The input to ConvAE(flow) is a stacked historical flow map with size  $20 \times 227 \times 227$ , acquired from pre-trained FlowNet2 [19].

**ConvLSTMAE** [6] is similar to ConvAE but models spatial and temporal features separately. A 2D CNN encoder first captures spatial information from each frame, then a multi-layer ConvLSTM recurrently encodes temporal features. Another 2D CNN decoder then reconstructs input video clips (Fig. 4(b)). We also implemented **ConvLSTMAE(gray)** and **ConvLSTMAE(flow)**.

**AnoPred** [26] is a frame-level VAD method taking four continuous previous RGB frames as input and applying UNet to predict a future RGB frame (Fig. 4(c)). AnoPred boosts prediction accuracy with a multi-task loss incorporating image intensity, optical flow, gradient, and adversarial losses. AnoPred was proposed for surveillance cameras. However, traffic videos are much more dynamic, making future frame prediction difficult. Therefore we also benchmark a variant of AnoPred to focus on video foreground. We use Mask-RCNN [15] pre-trained on Cityscapes [7] to acquire object instance masks for each frame, and apply instance masks to input and target images, resulting in a **AnoPred+Mask** method that only predicts foreground objects and ignores noisy backgrounds such as trees and billboards. In contrast to [14,6], AnoPred uses Peak Signal to Noise Ratio,  $PSNR = 10 \log_{10} MSE^{-1}$  as anomaly score with better results.

## 4.2 Object-centric Unsupervised VAD

**TAD** [44] models normal bounding box trajectories in traffic scenes with a multi-stream RNN encoder-decoder [43] (Fig. 4(d)) to encode past trajectories and ego motion and to predict future object bounding boxes. Prediction results are collected; prediction consistency instead of accuracy is used to compute per-object anomaly scores. Per-object scores are averaged to form a per-frame score.

Ionescu *et al.* [20] propose to treat object normality as multi-modal and use k-means to find the normality clusters in hidden space. Liu *et al.* [25] use margin learning (ML) to enforce large distances between normal and abnormal features. We combine these ideas and propose **TAD+ML**, as shown in Fig. 4(e). We adopt k-means to cluster encoder hidden features. Each cluster is considered one normality, *i.e.* one type of normal motion, so that each training sample is initialized with a cluster ID as its normality label. Then we used a center loss [40] to enforce tight distribution of samples from the same normality and to enforce samples from different normalities to be distinguishable. Center loss is more efficient than triplet loss [25] in large batch training. Fig 4(e) shows an example of visualized hidden features after ML. Note that we removed the ego motion branch in TAD+ML for simplicity as it does not affect results.

**Ensemble.** Frame-level VAD methods focus on appearance while object-centric methods focus more on object motion. We are not aware of any method combining the two. Appearance-only methods may fail with drastic variance in lighting conditions and motion-only methods may fail when trajectory prediction is imperfect. In this paper, we combine AnoPred+Mask and TAD+ML, into an ensemble method. We trained each method independently and fused their output anomaly scores by average pooling. We have observed that such a late fusion is better than fusing hidden features in an early stage and training the two models together, since their hidden features are scaled differently. AnoPred+Mask encodes one feature per frame, while TAD+ML has one feature per object.

## 4.3 Supervised VAD as Online Action Detection

VAD can also be interpreted as binary action detection with normal and abnormal classes. We benchmark multiple video action detection methods on DoTA to provide insight in supervised VAD. We use an ImageNet pre-trained ResNet50 [16] model to collect frame features and train different classifiers: 1) **FC**, a three-layer fully-connected network for image classification; 2) **LSTM**, a one-layer LSTM classifier for sequential image classification; and 3) **Encoder-Decoder**, an LSTM model with an encoder classifying current frames and a decoder predicting future classes. We also train the temporal recurrent network (**TRN**) [41] which is built upon encoder-decoder except predictions are fed back to the encoder to improve performance.

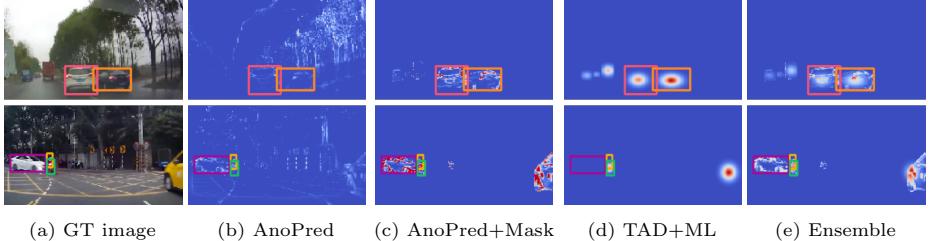


Fig. 5: Anomaly score maps computed by four methods. Ground truth anomalous regions are labeled by bounding boxes. Brighter color indicates higher score.

## 5 A New Evaluation Metric

### 5.1 Critique of Current VAD Evaluation

Most VAD methods compute an anomaly score for each frame by averaging scores over all pixels or objects. Current evaluation method plots receiver operating characteristic (ROC) curves using temporally concatenated scores and computes an area under curve (AUC) metric. AUC measures how well a VAD method answers the **When** question but ignores **Where** since averaged anomaly score lacks spatial information. We argue AUC is insufficient to fully evaluate VAD performance. In computing AUC, a true positive is a prediction where the model predicts high anomaly score for a positive frame. Fig. 5 shows two positive frames and their corresponding score maps computed by the four benchmarked VAD methods. Although the maps are different, the anomaly scores averaged from these maps are similar, meaning they are treated similarly in AUC evaluation. This results in similar AUCs among all methods, which leads to a conclusion that all perform similarly. However, AnoPred (Fig. 5(b)) predicts high scores for trees and other noise. AnoPred+Mask and TAD+ML (Fig. 5(c) and 5(d)) predict high scores for unrelated vehicles. Ensemble (Fig. 5(e)) alleviates these problems but still has high anomaly scores outside the labeled anomalous regions. Note that score maps of TAD+ML and Ensemble are pseudo-maps introduced in Section 5.2. Although these methods yield similar AUCs, VAD methods should be distinguished by their abilities to localize anomalous regions. Anomalous region localization is essential because it improves reaction to anomalies, e.g. collision avoidance, and aids in model explanation, e.g. a model predicts a car-to-car collision because it finds anomalous cars, not trees or noise. This motivates a new spatio-temporal metric to better address both **When** and **Where** questions.

### 5.2 The Spatial-Temporal Area Under Curve (STAUC) Metric

First, calculate the true anomalous region rate ( $TARR$ ) for each positive frame,

$$TARR_t = \frac{\sum_{i \in m_t} \Delta I(i)}{\sum_{i \in M} \Delta I(i)}, \quad (2)$$

where  $\Delta I$  is the anomaly score map from Eq. (1),  $M$  represents all frame pixels,  $m_t$  is the annotated anomalous frame region (i.e., the union of all annotated bounding boxes).  $TARR \in [0, 1]$  is a scalar describing how much of the anomaly score is located within the true anomalous region.  $TARR$  is inspired by anomaly segmentation tasks where the overlap between prediction and annotation is computed [2]. Next, calculate the spatio-temporal true positive rate ( $STTPR$ ),

$$STTPR = \frac{\sum_{t \in TP} TARR_t}{|P|}, \quad (3)$$

where  $TP$  represents all true positive predictions and  $P$  represents all ground truth positive frames.  $STTPR$  is a weighted TPR where each true positive is weighted by its  $TARR$ . We then use  $STTPR$  and FPR to plot a spatio-temporal ROC (STROC) curve and then calculate the STAUC. Note that  $STAUC \leq AUC$  and the two are equal in the best case where  $TARR_t = 1 \forall t$ .

Object-centric VAD [44, 20, 31] computes per-object anomaly scores  $s_k$  instead of an anomaly score map  $\Delta I$ . To generalize the STAUC metric to object-centric methods, we first create pseudo-anomaly score maps per Fig. 5(d). Each object has a 2D Gaussian distribution centered in its bounding box. Pixel score is then computed as the sum of the scores calculated from all boxes it occupies,

$$\Delta I_{pseudo}(i) = \sum_{\forall k, i \in B_k} s_k e^{-\frac{|i_x - x_k|^2}{2w_k} - \frac{|i_y - y_k|^2}{2h_k}}, \quad (4)$$

where  $i_x$  and  $i_y$  are coordinates of pixel  $i$  and  $[x_k, y_k, w_k, h_k]$  is center location, width, and height of object bounding box  $B_k$ . For Ensemble method, we take the average of  $\Delta I$  and  $\Delta I_{pseudo}$  as the anomaly score map in Fig. 5(e). This map is used like  $\Delta I$  in Eq. (2) to compute  $TARR$  and STAUC.

$TARR$  is not robust to anomalous region size  $m_t$ . When  $m_t \ll M$ ,  $TARR$  could be small even though all anomaly scores are high in  $m_t$ . We thus propose selecting the top  $N\%$  of pixels with the largest anomaly scores as candidates, and compute  $TARR$  from these candidates instead of all pixels. Selecting a constant  $N$  can be arbitrary. An extremely small  $N$  such as 0.01 may result in a biased candidate set dominated by false or true detections such that  $TARR = 0$  or 1. To address this issue, we compute an adaptive  $N$  for each frame based on the size of its annotated anomalous region as given by

$$N_{adaptive} = \frac{\text{number of pixels in annotated anomalous region}}{\text{Total number of pixels}} \times 100. \quad (5)$$

The average  $N_{adaptive}$  of DoTA is 11.12 with a standard deviation 13.09. The minimum and maximum  $N_{adaptive}$  values are 0.005 and 95.8, showing extreme cases where the anomalous object is very small (far away) or large (nearby).

A critical consideration for any new metric is its robustness to hyper parameters. We have tested STAUC with  $N = [1, 5, 10, 20, 50, 100, N_{adaptive}]$  for different VAD methods per Fig. 6(a), STAUC slightly decreases with  $N$  increasing but stabilizes when  $N$  is large indicating STAUC is robust. Fig. 6(b) shows

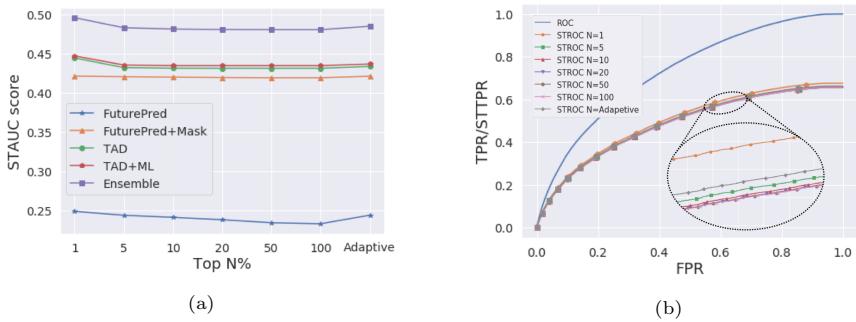


Fig. 6: (a) STAUC values of different methods using different top  $N\%$ ; (b) ROC curve and STROC curves of the Ensemble method with different top  $N\%$ .

that STROC curves with different  $N$  are close, especially when  $N \geq 5$ , and their upper bound is the traditional ROC.  $N_{adaptive}$  is selected for our benchmarks based on each frame’s annotation and its corresponding mid-range STAUC value.

## 6 Experiments

We benchmarked VAD and VAR with the **When-Where-What** pipeline. We randomly partitioned DoTA into 3,275 training and 1,402 test videos and use these splits for both tasks. Unsupervised VAD models must be trained only with normal data, so we extract precursor frames from each video for training. Supervised VAD and VAR models are trained using all training data.

### 6.1 Task 1: Video Anomaly Detection (VAD)

**Implementation Details.** We trained all ConvAE and ConvLSTMAE variants using AdaGrad with learning rate 0.01 and batch size 24. AnoPred, TAD, and their variants are trained per the original papers. TAD+ML uses k-means ( $k = 10$ ) and center loss weight 1 [20,25]. To train supervised methods, we first extract image features using ImageNet pre-trained ResNet50, then train each model with learning rate 0.0005, batch size 16. All models are trained on NVIDIA TITAN XP GPUs. To fairly compare frame- and object-based methods, we ignore videos with unknown category or without objects, resulting in 1,305 test videos.

**Overall Results.** The top four rows of Table 3 show performance of ConvAE and ConvLSTMAE with grayscale or optical flow inputs. Generally, using optical flow achieves better AUC, indicating motion is an informative feature for this task. However, all baselines achieve low STAUC, meaning that they cannot localize anomalous regions well. AnoPred achieves 67.5 AUC but only 24.4 STAUC, while AnoPred+mask has 2.7 lower AUC but 17.7 higher STAUC. By applying instance masks, the model focuses on foreground objects to avoid computing high scores for background, resulting in slightly lower AUC but much

Table 3: Benchmarks of VAD methods on the DoTA dataset.

Method	Type	Input	AUC $\uparrow$	STAUC $\uparrow$
ConvAE (gray) [14]	Unsupervised	Gray	64.3	7.4
ConvAE (flow) [14]		Flow	66.3	7.9
ConvLSTMMAE (gray) [6]		Gray	53.8	12.7
ConvLSTMMAE (flow) [6]		Flow	62.5	12.2
AnoPred [26]		RGB	67.5	24.4
AnoPred [26] + Mask		Masked RGB	64.8	42.1
TAD [44]		Box + Flow	69.2	43.3
TAD [44] + ML [20,25]		Box + Flow	69.7	43.7
Ensemble		RGB + Box + Flow	<b>73.0</b>	<b>48.5</b>
FC			61.7	-
LSTM [18]	Supervised	RGB	63.7	-
Encoder-Decoder [5]		RGB	73.6	-
TRN [41]			<b>78.0</b>	-

higher STAUC. This supports our hypothesis that *higher AUC does not imply a better VAD model, while STAUC reveals its ability to localize anomalous regions*. TAD outperforms AnoPred on both metrics by specifically focusing on object motion and location, both of which are important indicators of traffic anomalies. The margin learning (ML) module further improves TAD by a small margin. Our Ensemble method achieves the best AUC and STAUC among all methods, indicating that combining frame-level appearance and motion features is a direction worth investigating in future VAD research.

State-of-the-art supervised methods such as TRN achieve higher AUC than unsupervised methods. These methods focus on temporal modeling and simplify spatial modeling by using pre-trained features. We believe that exploring spatial modeling could further boost the performance of supervised methods. However, since these models directly predict an anomaly score for each frame rather than computing an anomaly score map, it is not straightforward to compute STAUC for them. Other ways such as a soft attention or class activation map might help model explainability in the future [23,46].

**Per-class Results.** Table 4 shows per-class results of unsupervised methods: AnoPred, AnoPred+Mask, TAD+ML and Ensemble. We observe that STAUC (unlike AUC) distinguishes performance by anomaly type, offering guidance as researchers seek to improve their methods. For example, Ensemble has comparable AUCs on OC and VP anomalies (73.4 vs 70.1) but significantly different STAUCs (56.6 vs 35.2), showing that anomalous region localization is harder on VP. Similar trends exist for the AH\*, LA\*, VP\* and VO\* columns. Second, frame-level and object-centric methods compensate each other in VAD as shown by the Ensemble method’s highest AUC and STAUC values in most columns. Third, localizing anomalous regions in non-ego anomalies is more difficult, as STAUCs on ego-involved anomalies are generally higher. One reason is that ego-involved anomalies have better dashcam visibility and larger anomalous regions, making them easier to detect. Table 4 also shows the difficulties of detecting different categories, with AH\*, VP, VP\*, VO\* and LA\* especially challenging for all methods. We observed that pedestrians in VP and VP\* videos become occluded

Table 4: Evaluation metrics of each individual anomaly class. Ego-involved and non-ego (\*) anomalies are shown separately. VO and OO columns are not shown because they do not contain anomalous traffic participants.

Method	ST	AH	LA	OC	TC	VP	ST*	AH*	LA*	OC*	TC*	VP*	VO*	OO*
<b>Individual Anomaly Class AUC:</b>														
AnoPred	69.9	73.6	<b>75.2</b>	69.7	73.5	66.3	70.9	62.6	60.1	65.6	65.4	64.9	64.2	57.8
AnoPred+Mask	66.3	72.2	64.2	65.4	65.6	66.6	72.9	63.7	60.6	66.9	65.7	64.0	58.8	59.9
TAD+ML	67.3	77.4	71.1	68.6	69.2	65.1	75.1	66.2	66.8	74.1	72.0	69.7	63.8	69.2
Ensemble	<b>73.3</b>	<b>81.2</b>	74.0	<b>73.4</b>	<b>75.1</b>	<b>70.1</b>	<b>77.5</b>	<b>69.8</b>	<b>68.1</b>	<b>76.7</b>	<b>73.9</b>	<b>71.2</b>	<b>65.2</b>	<b>69.6</b>
<b>Individual Anomaly Class STAUC:</b>														
AnoPred	37.4	31.5	32.8	34.3	33.6	24.9	25.9	15.0	12.5	13.0	20.9	14.0	8.2	8.8
AnoPred+Mask	51.8	51.9	45.1	50.3	47.5	<b>41.0</b>	45.3	31.1	33.8	42.5	40.3	25.3	22.9	33.8
TAD+ML	47.4	55.6	46.3	52.2	47.2	26.6	45.1	33.6	38.5	46.9	39.3	25.6	<b>29.0</b>	<b>44.4</b>
Ensemble	<b>54.4</b>	<b>60.3</b>	<b>53.8</b>	<b>56.5</b>	<b>54.9</b>	35.2	<b>52.4</b>	<b>36.4</b>	<b>40.8</b>	<b>51.9</b>	<b>44.7</b>	<b>28.6</b>	28.6	43.5

or disappear quickly after an anomaly happens, making it hard to detect the full anomaly event. AH\* has a similar issue since sometimes the vehicle ahead is largely occluded by the vehicle it impacts. VO\* is a rarer case in which a vehicle hits obstacles such as bumpers or traffic cones which are typically not detected and are sometimes occluded by the anomalous vehicle. Vehicles involved in LA\* usually move towards each other slowly until they collide and stop, making the anomaly subtle thus hard to distinguish.

**Qualitative Results.** Fig. 7 shows per-frame anomaly scores and *TARRs* of three methods on a video where they all achieve high AUCs. AnoPred+Mask has low *TARR* along the video, indicating failure of correctly localizing anomalous regions. TAD+ML computes high anomaly scores but low *TARR* in the left example due to inaccurate trajectory prediction for the left car. In the right image, it finds one of the anomalous cars but also marks an unrelated car by mistake. Ensemble combines the benefits of both with anomaly scores for 20-30th anomaly frames always higher than normal frames. It computes high *TARR* during 10-20th anomaly frames as shown in the left score map. The right map shows a failure case combining the failure of AnoPred+Mask and TAD+ML. Although these methods achieve high AUC, their spatial localization is limited per *TARR*. More qualitative results are shown in our supplement.

## 6.2 Task 2: Video Action Recognition (VAR)

The goal of VAR is to assign each video clip to one anomaly category. We benchmark seven VAR methods on DoTA: C3D [36], I3D [3], R3D [37], MC3 [37], R(2+1)D [37], TSN [39] and SlowFast [10]. The previous training/test split is used. Unknown UK(\*) anomalies are ignored, yielding 3216 training and 1369 test videos. We trained all models with SGD, learning rate 0.01 and batch size 16 on NVIDIA TITAN XP GPUs. Models are initialized with Sports-1M [21] (C3D) or Kinetics [22] (rest) pre-trained weights; 0.5 probability random horizontal flip offers data augmentation. For evaluation, we randomly select ten clips from each test video per [10] except TSN which uses 25 frames per video.

Table 5: VAR method per-class and mean top-1 accuracy with the DoTA dataset.

Method	backbone	Anomaly Class																	
		ST	AH	LA	OC	TC	VP	VO	OO	ST*	AH*	LA*	OC*	TC*	VP*	VO*	OO*	AVG	
TSN	ResNet50	18.2	67.2	<b>52.9</b>	<b>53.8</b>	<b>71.0</b>	0.0	0.0	61.6	0.0	14.7	25.3	6.7	48.1	9.5	0.0	<b>53.4</b>	30.2	
C3D	VGG16	25.5	61.8	43.9	47.8	57.9	3.3	4.4	52.9	1.2	18.4	36.0	6.7	55.9	8.6	6.0	33.2	29.0	
13D	InceptionV1	10.0	62.4	45.8	45.8	62.2	2.8	6.9	66.6	2.4	<b>28.1</b>	24.5	4.7	60.3	9.5	5.0	37.6	29.7	
R3D	ResNet18	0.0	56.5	49.6	49.8	66.6	4.4	6.2	47.7	1.8	17.6	32.2	1.0	48.3	<b>15.2</b>	6.5	48.0	28.2	
MC3D	ResNet18	6.4	62.9	40.1	57.7	64.5	16.7	0.0	61.5	2.4	18.1	20.2	4.0	62.2	4.8	<b>6.5</b>	45.6	29.6	
R(2+1)D	ResNet18	4.5	64.7	42.8	47.6	68.7	<b>25.6</b>	5.6	64.4	<b>9.4</b>	14.3	24.3	2.3	<b>64.7</b>	9.5	0.0	47.8	<b>31.0</b>	
SlowFast	ResNet50	0.0	<b>70.0</b>	46.0	48.9	67.2	5.6	<b>13.1</b>	<b>68.3</b>	5.9	24.9	<b>37.2</b>	3.3	64.0	0.0	0.0	41.3	<b>31.0</b>	

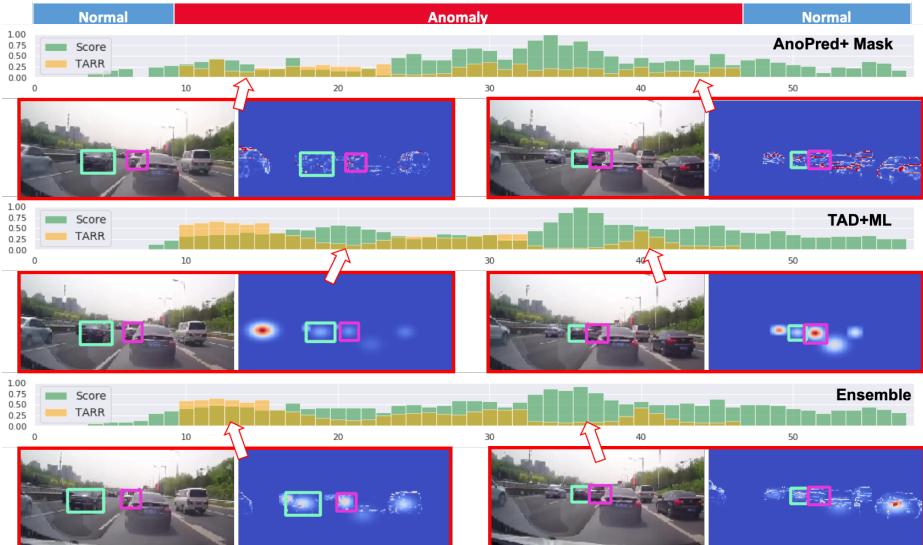


Fig. 7: Per-frame anomaly scores and TARRs of three methods. Selected RGB frame and score maps are shown. Note that TARR only exists in positive frames.

Table 5 lists the backbone network of each model and its per-class accuracy. Although newer methods R(2+1)D and SlowFast achieve higher average accuracy, all candidates suffer from low accuracy on DoTA, indicating that traffic anomaly classification is challenging. First, distant anomalies and occluded objects have low visibility thus are hard to classify. For example, VO(\*) are hard to classify due to low visibility and diverse obstacle types per Section 6.1. AH\* and OC\* are also difficult since the front or oncoming vehicles are often occluded. Second, some anomalies are visually similar to others. For example, ST(\*) are rare and look similar to AH(\*) or LA(\*) (Fig.2) since the only difference is whether the collided vehicle is starting, stopping, or stationary. Third, anomaly category is usually determined by the frames around anomaly start time, while the later frames do not reveal this category clearly. We have observed 2-4% accuracy improvement when testing models only on first half of each clip. Additional benchmarks are available in our supplement.

## 7 Conclusion and Future Work

This paper investigated a **When-When-What** pipeline for traffic anomaly detection. We introduced a large-scale dataset containing temporal, spatial, and categorical annotations and benchmarked state-of-the-art VAD and VAR methods. We proposed a new spatial-temporal area under curve (STAUC) metric to better evaluate VAD performance. Experiments showed STAUC outperforms AUC but that traffic video anomaly detection and classification problems are far from solved. DoTA offers the community new data for further VAD and VAR research and also can be used to study important object (visual saliency) detection, online detection of traffic anomaly, and validation and verification of autonomous driving efforts.

## 8 Acknowledgement

This research has been supported by the National Science Foundation under awards CNS 1544844 and CAREER IIS-1253549, and by the IU Office of the Vice Provost for Research, the IU College of Arts and Sciences, and the IU Luddy School of Informatics, Computing, and Engineering through the Emerging Areas of Research Project “Learning: Brains, Machines, and Children.” We also thank Derek Lukacs and RedBrickAI<sup>3</sup> for supporting our data annotation work. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing the official policies, either expressly or implied, of the U.S. Government, or any sponsor.

## References

1. Bakker, J., Jeppsson, H., Hannawald, L., Spitzhüttl, F., Longton, A., Tomasch, E.: Ig glad-international harmonized in-depth accident data. In: ESV (2017)
2. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: Mvtec ad—a comprehensive real-world dataset for unsupervised anomaly detection. In: CVPR (2019)
3. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR (2017)
4. Chan, F.H., Chen, Y.T., Xiang, Y., Sun, M.: Anticipating accidents in dashcam videos. In: ACCV (2016)
5. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv:1406.1078 (2014)
6. Chong, Y.S., Tay, Y.H.: Abnormal event detection in videos using spatiotemporal autoencoder. In: ISNN (2017)
7. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016)
8. De Geest, R., Gavves, E., Ghodrati, A., Li, Z., Snoek, C., Tuytelaars, T.: Online action detection. In: ECCV (2016)

---

<sup>3</sup> <https://www.redbrickai.com/>

9. Fang, J., Yan, D., Qiao, J., Xue, J.: Dada: A large-scale benchmark and model for driver attention prediction in accidental scenarios. arXiv:1912.12148 (2019)
10. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: ICCV (2019)
11. Gao, J., Yang, Z., Nevatia, R.: Red: Reinforced encoder-decoder networks for action anticipation. BMVC (2017)
12. Gao, M., Xu, M., Davis, L.S., Socher, R., Xiong, C.: StartNet: Online detection of action start in untrimmed videos. In: ICCV (2019)
13. Gong, D., Liu, L., Le, V., Saha, B., Mansour, M.R., Venkatesh, S., Hengel, A.v.d.: Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: ICCV (2019)
14. Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, L.S.: Learning temporal regularity in video sequences. In: CVPR (2016)
15. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: ICCV (2017)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
17. Herzig, R., Levi, E., Xu, H., Gao, H., Brosh, E., Wang, X., Globerson, A., Darrell, T.: Spatio-temporal action graph networks. In: CVPRW (2019)
18. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation (1997)
19. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: Evolution of optical flow estimation with deep networks. In: CVPR (2017)
20. Ionescu, R.T., Khan, F.S., Georgescu, M.I., Shao, L.: Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In: CVPR (2019)
21. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: CVPR (2014)
22. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv:1705.06950 (2017)
23. Kim, J., Rohrbach, A., Darrell, T., Canny, J., Akata, Z.: Textual explanations for self-driving vehicles. In: ECCV (2018)
24. Li, W., Mahadevan, V., Vasconcelos, N.: Anomaly detection and localization in crowded scenes. TPAMI (2013)
25. Liu, W., Luo, W., Li, Z., Zhao, P., Gao, S.: Margin learning embedded prediction for video anomaly detection with a few anomalies. In: IJCAI (2019)
26. Liu, W., Luo, W., Lian, D., Gao, S.: Future frame prediction for anomaly detection—a new baseline. In: CVPR (2018)
27. Lu, C., Shi, J., Jia, J.: Abnormal event detection at 150 fps in matlab. In: ICCV (2013)
28. Luo, W., Liu, W., Gao, S.: Remembering history with convolutional lstm for anomaly detection. In: ICME (2017)
29. Luo, W., Liu, W., Gao, S.: A revisit of sparse coding based anomaly detection in stacked rnns framework. In: ICCV (2017)
30. Medel, J.R., Savakis, A.: Anomaly detection in video using predictive convolutional long short-term memory networks. arXiv:1612.00390 (2016)
31. Morais, R., Le, V., Tran, T., Saha, B., Mansour, M., Venkatesh, S.: Learning regularity in skeleton trajectories for anomaly detection in videos. In: CVPR (2019)
32. Shou, Z., Pan, J., Chan, J., Miyazawa, K., Mansour, H., Vetro, A., Giro-i Nieto, X., Chang, S.F.: Online detection of action start in untrimmed, streaming videos. In: ECCV (2018)

33. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: NeurIPS (2014)
34. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: CVPR (2018)
35. Suzuki, T., Kataoka, H., Aoki, Y., Sato, Y.: Anticipating traffic accidents with adaptive loss and large-scale incident db. In: CVPR (2018)
36. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV (2015)
37. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: CVPR (2018)
38. Wang, J., Cherian, A.: Gods: Generalized one-class discriminative subspaces for anomaly detection. In: ICCV (2019)
39. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: ECCV (2016)
40. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: ECCV (2016)
41. Xu, M., Gao, M., Chen, Y.T., Davis, L.S., Crandall, D.J.: Temporal recurrent networks for online action detection. In: ICCV (2019)
42. Yao, Y., Atkins, E.: The smart black box: A value-driven high-bandwidth automotive event data recorder. TITS (2020)
43. Yao, Y., Xu, M., Choi, C., Crandall, D.J., Atkins, E.M., Dariush, B.: Egocentric vision-based future vehicle localization for intelligent driving assistance systems. In: ICRA (2019)
44. Yao, Y., Xu, M., Wang, Y., Crandall, D.J., Atkins, E.M.: Unsupervised traffic accident detection in first-person videos. In: IROS (2019)
45. Yu, F., Xian, W., Chen, Y., Liu, F., Liao, M., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving video database with scalable annotation tooling. arXiv:1805.04687 (2018)
46. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR (2016)

# When, Where, and What? A New Dataset for Anomaly Detection in Driving Videos

## – Supplementary Material

Yu Yao<sup>1</sup>

Xizi Wang<sup>2</sup>

Mingze Xu<sup>3\*</sup>

Zelin Pu<sup>1</sup>

Ella M. Atkins<sup>1</sup>

David J. Crandall<sup>2</sup>

<sup>1</sup>University of Michigan

<sup>2</sup>Indiana University

<sup>3</sup>Amazon Rekognition

{brianyao,ematkins}@umich.edu, {xiziwang,djcran}@indiana.edu

## 1 Additional DoTA Dataset Example

Fig. 1 shows one sampled frame sequence for each anomaly category in our DoTA dataset. Each row shows five frames sampled from one DoTA video, including two frames from the normal precursor, two frames from the anomaly window (marked by a red boundary), and one frame from the post-anomaly. The annotated bounding boxes of anomalous objects are shown by shadowed rectangles, and objects across frames are in consistent colors. Anomaly category abbreviations are listed to the left, where “\*” indicating non-ego anomalies.

This figure illustrates that some samples from different categories look similar, for example ST (row 1) is similar to both AH (row 3) and OC (row 7) except that in ST the front car is stationary. The AH\* sample is similar to the OC\* sample since it is difficult to distinguish front and rear vehicle views. The VP sample is close to the TC sample due to the similarity between a pedestrian and a rider. Moreover, some non-ego anomalies can have low visibility due to their distance from the camera, such as the VP\* and the OO\* example in Fig. 1. VO and VO\* are anomalies where vehicles collide with unexpected/auxiliary obstacles such as dropped cargo and traffic cones. Note that VO and OO are two anomaly categories with no bounding box label typically provided; by definition, VO and OO do not involve traffic participants.

## 2 Additional VAD (Task 1) Results

We present more qualitative results of AnoPred+Mask, TAD+ML and Ensemble methods in this supplement section. Fig. 2(a) shows an ego-involved ahead collision (AH). AnoPred+Mask computes a high anomaly score in the early frames by mistake since the prediction of the left car is inaccurate, as shown in the score map. TAD+ML computes a low anomaly score for this frame and therefore the Ensemble method benefits. The right example shows the TAD+ML method correctly computing a high score for the ahead car but also another high score for the bus on the right. The ensemble benefits from AnoPred+Mask so that

---

\* This work was done while the author was at Indiana University.

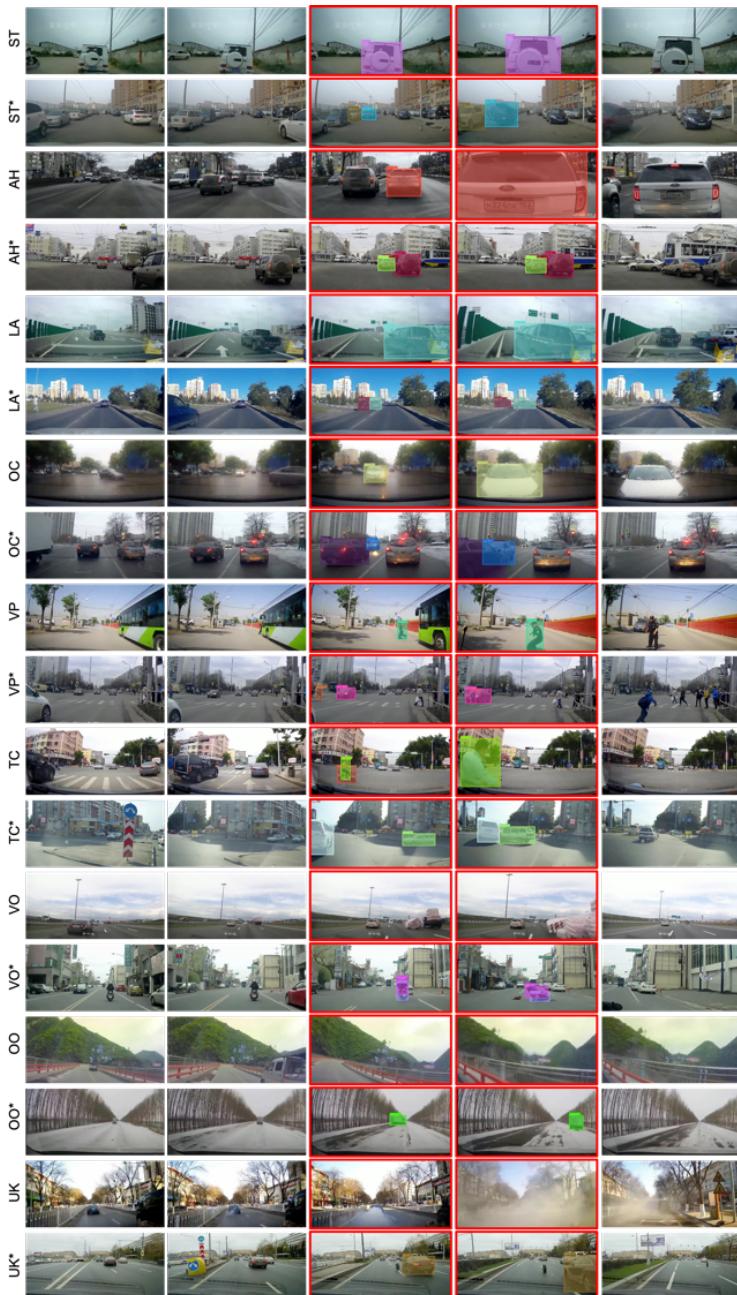
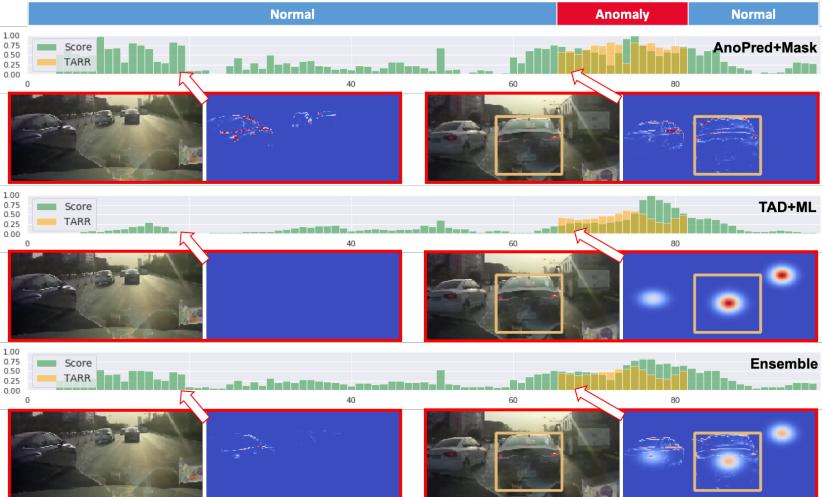
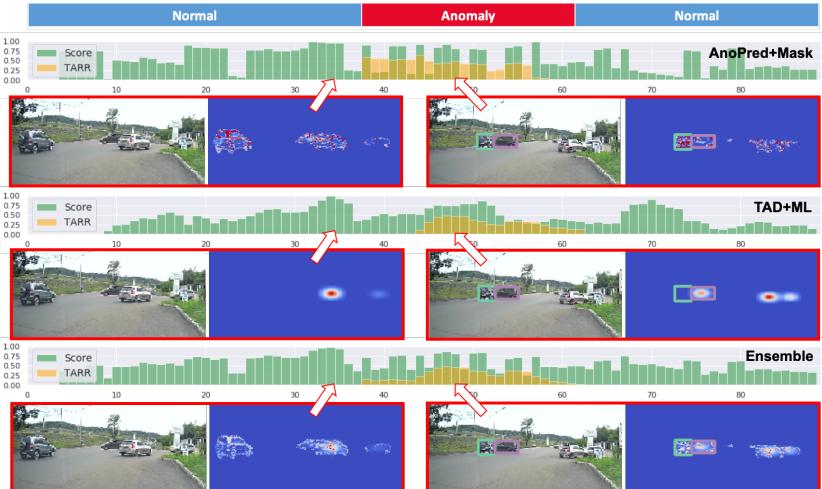


Fig. 1: Sampled sequences from the DoTA dataset. Frames with a red boundary are anomalous frames. Spatial annotations are shown as shadowed bounding boxes. Short anomaly category labels with \* indicate non-ego anomalies.



(a) An example with Ensemble method outperforming each individual method.



(b) A failure case where both spatial and temporal performance is bad.

Fig. 2: Additional qualitative results.

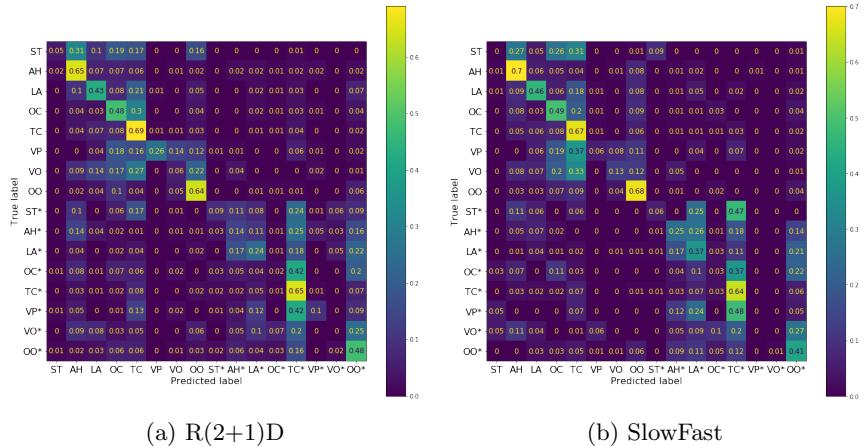


Fig. 3: Confusion matrix of two state-of-the-art VAR methods on DoTA.

it focuses more attention on the ahead car instead of the bus. Fig. 2(b) shows a failure case where all methods perform poorly in detecting a non-ego turning/crossing accident (TC\*). The left example shows that all methods compute high anomaly scores for normal frames, where the silver car had to brake before turning to the right to avoid the black car which is turning into its lane. This example reveals that the tested unsupervised methods predict false alarms for near-incidence events. The right example shows that TAD+ML misses one of the anomalous cars, which is captured by AnoPred+Mask. This can be caused by the failure of object tracking in collision scenarios.

### 3 Additional VAR (Task 2) Results

In our submitted paper, we benchmarked several state-of-the-art video action recognition (VAR) models on DoTA dataset. Fig. 3 show the confusion matrices of R(2+1)D and SlowFast, two of the best models evaluated in our experiments. In addition to Table 5 in the paper, the confusion matrix shows the most confusing categories to help us understand challenging scenarios provided in the DoTA dataset. We make three observations from Fig. 3. First, both models have similar confusion matrices, indicating that they perform similarly on DoTA dataset. Second, some categories are confused with other specific categories due to their similarities. Among all categories, TC, TC\*, OC and OO\* are four classes for which many categories are confused. One reason is that there are a large number of samples for these categories in DoTA. Another reason is the similarities among categories. For example OO\* is usually an out-of-control vehicle swerving on the road and finally leaving the roadway. Other non-ego anomalies, while having their own features, often result in similar irregular motions, resulting in confusion with OO\*. Third, ego-involved categories are usually not confused

Table 1: Online Video Action Detection on our DoTA dataset. “\*” indicates non-ego anomaly categories.

Method	Anomaly Category																	
	ST	AH	LA	OC	TC	VP	VO	OO	ST*	AH*	LA*	OC*	TC*	VP*	VO*	OO*	mAP	
FC	<b>2.5</b>	13.9	10.6	6.2	16.3	0.8	<b>1.2</b>	21.0	0.6	2.9	3.0	0.6	8.0	<b>1.2</b>	0.7	7.6	9.9	
LSTM	0.6	19.9	15.1	9.2	25.3	2.4	0.6	34.3	0.6	3.8	5.0	1.5	11.0	1.2	0.5	13.3	12.9	
Encoder-Decoder	0.5	20.1	15.6	10.4	28.1	<b>2.9</b>	0.7	<b>39.9</b>	<b>0.8</b>	3.7	7.4	2.5	14.7	1.2	0.5	13.2	14.5	
TRN	1.0	<b>22.8</b>	<b>20.6</b>	<b>15.5</b>	<b>30.0</b>	1.5	0.7	32.3	0.7	<b>4.0</b>	<b>10.2</b>	<b>2.9</b>	<b>17.0</b>	1.2	<b>0.7</b>	<b>13.8</b>	<b>15.3</b>	

with non-ego categories. This indicates that although the per-class recognition is difficult, current methods could capably distinguish ego-involved and non-ego anomalies.

## 4 Task 3: Online Action Detection

We provide benchmarks for online video action detection on DoTA dataset. Online action detection recognizes the anomaly type by only observing the current and past frames, making it suitable for autonomous driving applications. Since online action detection does not have a full observation of the whole video sequence, online action detection is considered a more difficult task than is traditional VAR. In this supplementary material, we provide benchmarks of several state-of-the-art online action detection methods on DoTA dataset. We use the same four online methods that have been used in supervised VAD: **FC**, **LSTM**, **Encoder-decoder** and **TRN**. The only difference is that the classifiers are designed to predict only one out of the 16 anomaly categories. We use the same training configurations to train these models. Table 1 shows the per-class average precision (AP) and the mean average prediction (mAP).

**Quantitative Results.** We observe that although TRN, a state-of-the-art method, achieves the highest mAP, all methods suffer from low precision on DoTA. Similar to what we have observed in the paper’s VAD and VAR experiments, online action detection is also difficult for ST, ST\*, VP, VP\*, VO and VO\*. AH\* and OC\* are also difficult due to the highly occluded front of a typical oncoming vehicle. We also observe that ego-involved anomalies are easier to recognize than non-ego anomalies due to their higher visibility.

**Qualitative Results.** Fig. 4 shows some examples of TRN results on our DoTA dataset. The bar plots show the classification confidences of each frame. Cyan colors represent anomalous frames while gray colors represent background (normal) frames. We make the following observations from this experiment: 1) Transition frames between normal and abnormal events are hard to classify. For example class confidences are low at the frames where color changes, i.e., anomaly start and end frames; 2) Subsequent frames after an anomaly begins can be hard to detect. For example confidence significantly decreases at around the 40th frame of first example and the 60th frame of the third example; 3) Visually

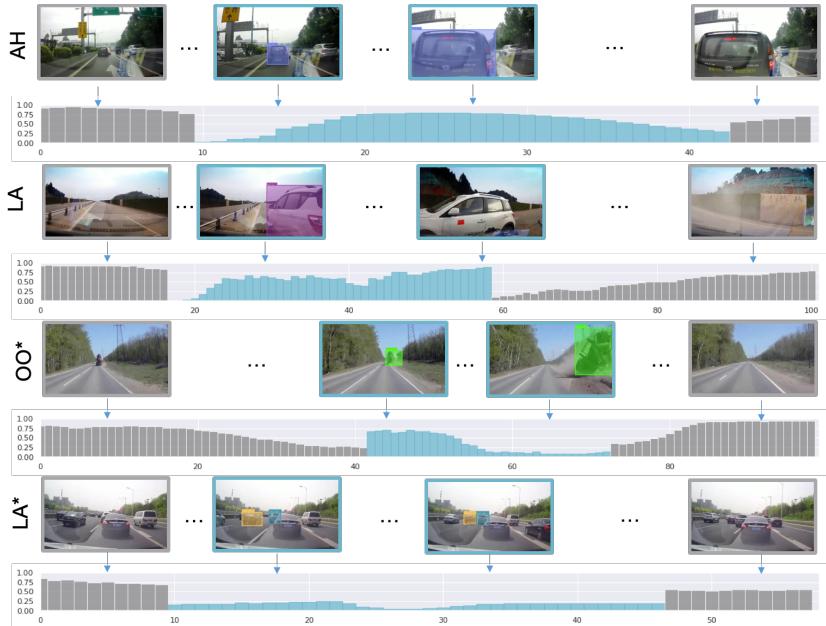


Fig. 4: Qualitative results of Temporal Recurrent Network (TRN) on our DoTA dataset. The bar plots show classification confidences of each video frame. Gray bars are confidences of "background" (or "normal") classes while cyan bars are confidences of ground truth anomaly classes. The top two rows are two ego-involved anomalies, while the 3rd row is a non-ego out-of-control anomaly. The 4th row is a case where TRN fails to detect a lateral collision.

similar anomalies and gentle anomalies are hard to detect. In the bottom failure case, the confidence of ground truth anomaly class  $LA^*$  is always low. These frames are either classified as background (normal) or  $AH^*$  due to the fact that this  $LA^*$  anomaly is visually similar to a typical  $AH^*$  anomaly since this collision is relatively gentle.