Philippine Journal of Science
150 (2): 515-525, April 2021
ISSN 0031 - 7683
Date Received: 28 Jul 2020

Investigating Visual Attention-based Traffic Accident Detection Model

Caitlienne Diane C. Juan[1], **Jaira Rose A. Bat-og**[1], **Kimberly K. Wan**[1], and **Macario O. Cordel II**[2*]

[1]College of Computer Studies, De La Salle University
Manila City, Metro Manila 0922 Philippines
[2]Data Science Institute, De La Salle University
Manila City, Metro Manila 0922 Philippines

Spotting abnormal or anomalous events using street and road cameras relies heavily on human observers which are subject to fatigue, distractions, and simultaneous attention limit. There are several proposed anomalous event detection systems based on complex computer vision algorithms and deep learning architectures. However, these systems are objective-agnostic, resulting in high false-negative cases in task-driven abnormal event detection. A straightforward solution is to use visual attention models. However, these are based on low-level features integrated with object detectors and scene context, rather than on the observers' object of gaze. In this paper, we explore a task-driven visual attention-based traffic accident detection system. We first examine the human fixations in free-viewing and task-driven goals using our proposed, first task-driven, fixation dataset of traffic incidents from different road cameras called TaskFix. We then used TaskFix to fine-tune the visual attention model, in this work called TaskNet. We evaluated the proposed fine-tuned model with quantitative and qualitative tests and compared it with other visual attention prediction architectures. The results indicate the potential of the visual attention models in abnormal event detection. The dataset is available here: https://bit.ly/TaskFixDataset

Keywords: anomalous event detection, human attention model, saliency models

INTRODUCTION

Catching abnormal events related to security in public and private spaces has advanced from manual roving to remote monitoring through multiple cameras strategically located in the vicinity. In traffic monitoring, in particular, control center personnel spot traffic accidents through road cameras rather than having traffic enforcers deployed on the roads. Recently, the task of simultaneously monitoring road cameras has become more challenging due to the significant increase in their number.

A research area in computer science and behavior understanding that could manage this challenge is anomaly detection. Anomaly detection aims to identify patterns in video feeds that are different from normal or expected behavior (Chandola *et al.* 2009). In the traffic monitoring context, anomalous or abnormal events could include gathering of people on the road, traffic violations, jaywalking, illegal parking, and traffic accidents. The typical approach is to use trajectory-based features (Ahmed *et al.* 2019) together with a classification or clustering algorithm to detect abnormal events. However, a comprehensive survey of anomaly detection (Kumaran

*Corresponding Author: macario.cordel@dlsu.edu.ph

et al. 2020) identified key issues in employing hand-crafted features, *e.g.* model generalizability; instead, exploring deep neural network (DNN)-based approaches are suggested.

Sophisticated DNN-based anomaly detection systems for detecting traffic accidents are proposed in (Shine *et al.* 2019; Yao *et al.* 2019; Zhao *et al.* 2019). The work of Yao and colleagues (2019) is designed for vehicle-mounted cameras while the works of Shine *et al.* (2019) and Zhao *et al.* (2019) assume stationary vehicles as part of the abnormal scenes. These make these systems limited to specific camera setups and in locations where roadside parking is allowed. There are other proposed anomalous event detection algorithms [*e.g.* Dhole *et al.* (2019); Ionescu *et al.* (2019); Yang *et al.* (2019)], but these are only applicable to surveillance scenarios captured by the University of California San Diego (UCSD 2013), Avenue (Lu *et al.* 2013), and ShanghaiTech (Luo *et al.* 2017) anomaly detection datasets. These systems are also objective agnostic, which will result in high false positives when used in traffic accident detection.

In this work, we study human visual attention models in traffic accident detection. The following motivates us to use such an approach:

- First, using a visual attention model avoids the training samples problem in common object detection systems due to the sparse occurrence of anomalous events. Works that involved fine-tuning visual attention models (Cordel *et al.* 2019; He *et al.* 2019; Murabito *et al.* 2018) perform well in predicting human attention despite the number of fine-tuning samples used.
- Second, using a visual attention model avoids the problem with the imbalanced distribution of normal and abnormal data as it only needs fixation data in a scene to allow competition between its neurons and show the most salient location in the scene (Itti and Koch 2000).
- Finally, being DNN-based, the visual attention model avoids the problem of choosing the appropriate feature extraction, environmental variations, and camera movement.

However, current visual attention models or saliency models [*e.g.* in Huang *et al.* (2015), Itti and Koch (2000), Kruthiventi *et al.* (2015), Tang *et al.* (2016)] are based on the free-viewing or bottom-up visual attention mechanism, which combines low-level features with object detectors and scene context. Detecting traffic accidents in a scene requires a saliency model that is task-driven or based on the top-down mechanism of visual attention.

In this paper, we present the TaskFix dataset – a task-driven human fixation dataset collected from human observers performing traffic accident detection exercise (see Figure 1a for example collected data). We perform an analysis on TaskFix and found a significant statistical difference between the human fixations in normal and abnormal scenes. Based on these findings, we investigate further if a saliency model could be modulated to predict human attention on performing traffic accident detection tasks. TaskFix is used to fine-tune our proposed model named TaskNet. Our main contributions are as follows:

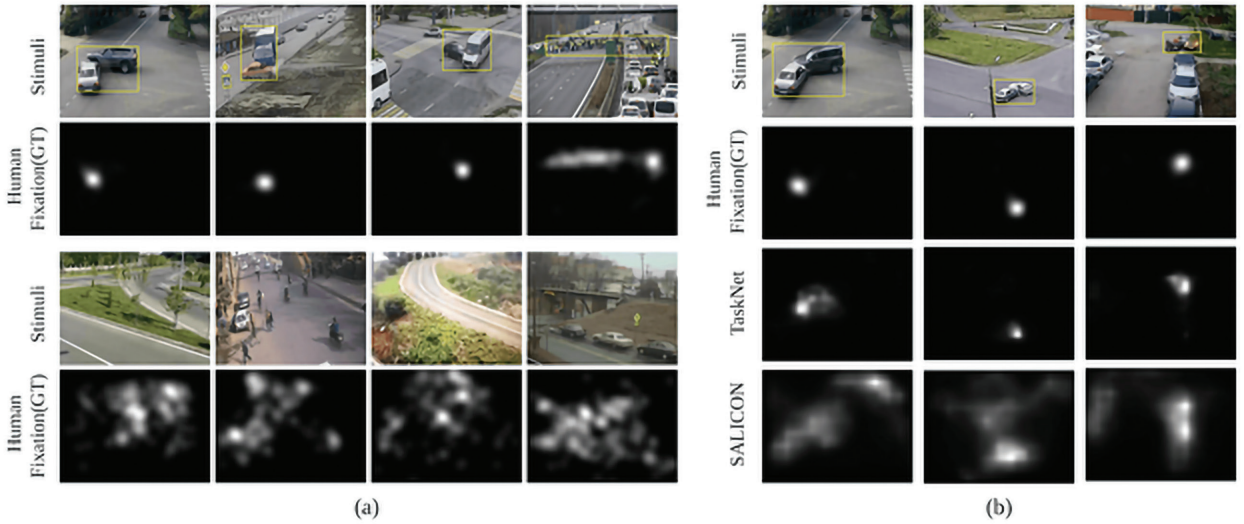

Figure 1. Using TaskFix, we showed that the fixations for abnormal traffic scenes, first four examples in (a), are statistically different from the fixations in normal traffic scenes, second four examples in (a). We use this observation to design a visual attention-based traffic incident detection model. Shown in (b) are the ground truth, the predicted locations of traffic accidents using TaskNet, and the outputs from a free-viewing model.

1. We provide a novel fixation dataset, called TaskFix, whose fixation data were collected from human observers performing traffic accident detection exercise. TaskFix consists of scenes with 718 abnormal traffic events (vehicular accidents) and 718 normal traffic events, for a total of 1436 images. This allows research on top-down human attention and anomalous event detection.
2. We discover a significant statistical difference in the fixation data of observers under the detection task when the event being spotted is in the scene and not in the scene.
3. We show that a visual attention model could encode the detection task performed in the fixation collection. The visual attention model shows better performance in detecting traffic accidents than other free-viewing-based saliency models (in qualitative and quantitative evaluations, refer to Figure 1b).

RELATED WORK

One of the popular approaches in anomaly detection is trajectory analysis, as detailed in a survey article by Ahmed *et al.* (2019). In trajectory-based approaches, the system uses object trajectory features to represent the positions of an object over time. These features are then used for event detection *via* Hidden-Markov Model [*e.g.* in Suk *et al.* (2019)], Bayesian framework [*e.g.* in San Miguel and Martinez (2012)], or Dirichlet process [*e.g.* in Bastani *et al.* (2016)]. In traffic surveillance, Ahmed and colleagues (2018) proposed to cluster the trajectory of a moving object by assigning a normality score to an object path based on a *t*-norm of fuzzy sets. The trajectory features used are high-level features including origin, destination, path, speed, and object size. Another trajectory-based work on traffic surveillance (Santhosh *et al.* 2019) focuses on identifying the moving object at the pixel-level through the use of optical flow and Bayesian algorithm. It resulted in a faster algorithm but only works for spatially separated objects and specific camera positions. Trajectory-based approaches, however – as noted by Ahmed *et al.* (2019) – require long-duration tracking, long-duration video datasets, and trajectory clustering, which highly depends on the track quality. Our work proposes a prediction of abnormal events at the scene level, thus removing the dependence on long-duration video datasets.

Kumaran *et al.* (2020) conducted a comprehensive survey on anomaly detection, which summarized key issues, including i) the limitations of benchmark comparisons in representing all real-life situations, and ii) the lack of generic techniques applicable to all datasets. The team also identified a possible solution through the use of DNN architecture. Some of these DNN systems that are applied in road anomaly include the work of Xu *et al.* (2017) that uses DNN and Autoencoder as the main feature extractor and SVM for identifying non-pedestrians appearing on walkways, Zhou *et al.* (2016) that uses CNN to detect U-turn movement of vehicles and unexpected presence of vehicles on walkways, and Vishnu *et al.* (2018) that detects congestions, ambulances, and accidents. In training their computational models, Xu *et al.* (2017) used the UCSD datasets, Zhou *et al.* (2016) used the UCSD, UMN, and U-turn datasets, and Vishnu *et al.* (2018) used their own local dataset.

More recently, traffic accident detection *via* DNN-based anomaly detection systems (Shine *et al.* 2019; Yao *et al.* 2019; Zhao *et al.* 2019) has been proposed. The work of Yao *et al.* (2019) presented an unsupervised model for traffic accident detection using a vehicle-mounted camera. The works of Shine *et al.* (2019) and Zhao *et al.* (2019) used region proposal-based DNN models to detect anomalies in road cameras. Both works assumed that normal vehicles never stay on the road except for unusual events. The former addressed the parked vehicle problem using a rule-based decision module while the latter used multi-object track algorithms to detect moving vehicles. These works are insightful, however, the datasets used to train these systems are formed using single or homogeneous scenes and, thus, cannot be applied to traffic analysis and monitoring since data obtained are from multiple camera feeds.

Visual attention models or saliency models are trained to determine and predict the visual attention of humans. It produces saliency maps that simulate the gaze movement of humans in an image. Huang *et al.* (2015) introduced a saliency prediction model that integrates a pre-trained DNN architecture. It takes advantage of the semantic-rich features to reduce the semantic gap between the saliency model and human eye fixation. Kruthiventi *et al.* (2015) also constructed a fully connected DNN to predict human visual attention. It incorporates a location biased convolution layer to model location-dependent patterns. Cornia *et al.* (2016) used a convolutional neural network to extract features of an image to predict the saliency map. It incorporates a loss function to train the model and to address the center-bias problem of saliency maps. These models, however, mimic the free-viewing or the bottom-up mechanism of human attention rather than the task-driven or top-down mechanism of human attention.

Our work investigates human attention models for a top-down anomaly detection task, particularly traffic accident detection. Due to limited datasets (refer to Table 1 for a summary of related datasets) with traffic accidents as anomalous events, we built our own datasets and collected fixation data.

We first propose a task-driven dataset. We performed a statistical analysis on the dataset and used our observation to develop a visual attention-based traffic accident detection model. We then utilized this task-driven dataset to investigate a visual attention-based computational model for traffic accident detection.

METHODOLOGY

To facilitate the methodology discussion, we define the following terms used in this section. Anomalous or abnormal traffic scenes are used to describe scenes with traffic accidents; otherwise, the scene is described as a normal traffic scene. Fixation data or simply fixations refer to the eye gaze location (x,y) in a scene collected through an eye tracker device. The fixation map reflects the distribution of the fixation data in the scene derived by passing a low-pass Gaussian filter over the fixation data.

The Proposed TaskFix Dataset Collection

As summarized in Table 1, there are several datasets for anomaly detection used in previous works, but these do not contain traffic accidents and do not have fixation data. There are several datasets containing fixation data, but these reflect image saliency based on bottom-up attention mechanism rather than abnormal event based on top-down attention mechanism. In order to study anomaly detection in road traffic, a dataset containing traffic accidents – with fixation data – is created.

Normal and abnormal traffic scenes are collected from video frames of road cameras showing natural traffic flow and vehicular accidents, respectively. These scenes are compiled from diverse road locations with various camera angles, quality, and weather conditions for a full set of 1436 images, with 718 samples for each traffic scene type. The video frames from these cameras are read and resized to 1024 × 768.

Through the Tobii Eye Tracker 4C with a 90-Hz sampling rate, the eye fixation data for each image sample are collected from 18 observers, aged 18–27 yr old with normal or corrected-to-normal vision. The images are presented to the observer on a 15-in LCD monitor with a 1366 × 768 screen resolution. The images are scaled to the full height of the screen, with the image width fixed at 1024 pixels. For each image shown, the observers are instructed to spot the occurrence of traffic accidents in the scene. Each image is shown for 5 s followed by a drift correction that requires observers to fixate at the center. Separately, three undergraduate students are hired to draw a bounding box (bbox) on the location of traffic accidents. The intersection of these bboxes was used as the ground truth bbox.

Fixation Data Validation

The agreement of the observers' fixations on the occurrence and location of traffic accidents is computed to quantitatively determine the collected fixation quality. A variable $\overline{P}$, shown in Equation 1, of the Fleiss' Kappa (Nichols *et al.* 2010) is used to calculate the extent to which the observers agree in detecting abnormal incidents:

Table 1. Comparison of TaskFix dataset with other publicly available datasets for anomaly detection and for image saliency.

	Datasets	Object in the abn. events	Anomalous event examples	Scene	# of samples	# of abn. events	Fixation data
Anomaly detection dataset	CAVIAR	Person/people	One person walking, fainting, slumping, people meeting	Lobby			No
	UCSD 1 and 2	Person/people	People walking across the walkway or grass	Walkways	14,000 4,560	4,005 1,636	No
	U-turn	Vehicles	U-turn	Junction	–	–	No
	UMN	Person/people	Unusual crowd activity	Walkways	7,710	–	No
	Avenue	Person/people	Wrong direction, strange action, abnormal object	Building entrance	30,652	3,820	No
	Shanghai-Tech	Person/people/ object	Wrong direction, strange action, abnormal object	Walkways	317,398	17,090	No
Image saliency dataset	SALICON	NA	NA	NA	10,000	0	Yes
	OSIE	NA	NA	NA	700	0	Yes
	EMOd	NA	NA	NA	1,019	0	Yes
	MIT300	NA	NA	NA	300	0	Yes
	CAT2000	NA	NA	NA	2,000	0	Yes
Ours (TaskFix)		Vehicles	Accidents	Road	1,436	718	Yes

$$\bar{P} = \frac{1}{Nn(n-1)} \left(\sum_{i=1}^{N} \sum_{j=1}^{K} n_{ij}^2 - Nn \right) \quad (1)$$

where N is the number of images, n is the number of observers, and $K = 2$ to represent if the observer fixated on the abnormal incident or not. That is, n_{ij} for $j = 1$ is the number of observers who do not fixate on the ith object, and n_{ij} for $j = 2$ is the number of observers who fixate on the ith object. The observer fixated on the abnormal incident if 80% of its fixations are in the bbox. $\bar{P}$ is, thus, computed for 718 abnormal traffic images only and is equal to 0.85. A $\bar{P}$ value approaching 1.00 indicates high agreement.

The ground truth fixation maps for the image samples are generated by passing a low pass Gaussian filter on the fixation matrix and then performing normalization (refer to Figure 2a). Unlike previous findings (Judd 2011; Zhao and Koch 2013), TaskFix average fixation map shows slight preference towards the top portion of the fixation maps, which corresponds to the farther part of the road (please refer to Figure 2b), presumably because observers were under a task to look for abnormal traffic scenes rather than free-viewing.

Experiments on the Dataset

The results of the experiments performed on TaskFix are summarized in Table 2. Note that, on average, 87% of the fixations in the abnormal traffic scenes are in the bbox.

Entropy is a statistical measure of the randomness of the fixation map of each image. Thus, in free-viewing fixation maps that reflect the salient image region, the entropy depends on the presence of distinct objects that consistently attract human attention (Xu *et al.* 2014). Interestingly, for task-driven fixation maps, the clusters of fixations in the location of the abnormal scenes result in lower entropy, as opposed to the scattered fixations in the normal traffic scenes. The histogram of entropy levels is shown in Figure 2c. The abnormal traffic scenes have a mean entropy level of 2.07 with a standard deviation of 0.58. The normal road traffic images have higher entropy with a mean equal to 4.50 and standard deviation equal to

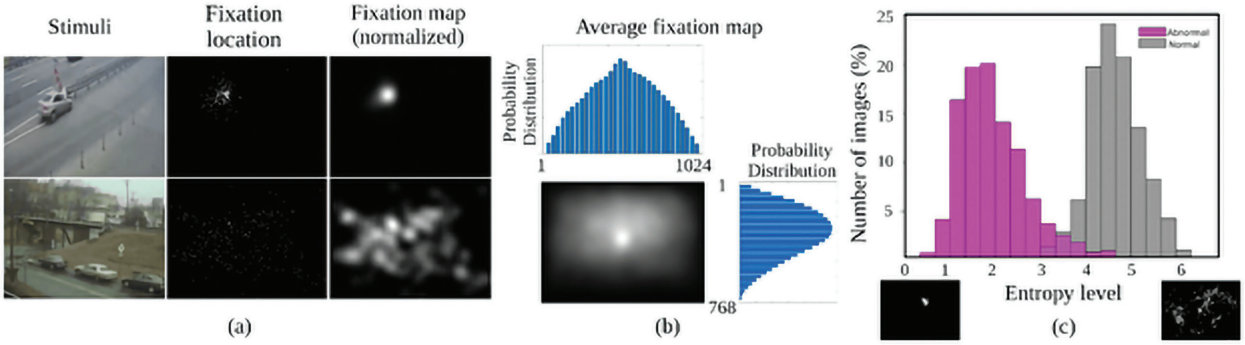

Figure 2. TaskFix is composed of abnormal [*e.g.* first row (a)] and normal [*e.g.* second row (a)] scenes from various traffic cameras. The fixation data collected are downsampled and filtered to generate the ground truth fixation maps, shown in last column (a). The average fixations of TaskFix show center bias in (b), with slight preference at the upper portion of the image. The fixation maps' entropy distribution implies that the difference between normal and abnormal traffic scenes (c).

Table 2. t-test results on the fixation densities of TaskFix.

Metrics	Average	*t*(*df*)	*p*
Fixations inside bbox and total fixations ratio	0.87	717	< 0.001
Entropy abnormal scenes	2.05	717	< 0.001
Entropy normal scenes	4.70	717	< 0.001
AttI of the most salient part in the abnormal scenes	0.72	717	< 0.005
AttI of the most salient part in the normal scenes	0.34	717	< 0.005

0.50. model visual attention in different resolutions. The VGG-16 (Simonyan and Zisserman 2014), GoogLeNet (Szegedy *et al.* 2015), and SSD (Liu *et al.* 2016) feature networks are considered in the evaluation. For each branch of the DNN, the input image is resized into corresponding image resolutions – one is 300 × 400 and the other is 600 × 800. The fully connected layers are replaced with an interpolation layer. The neural responses of the coarse and fine images are then concatenated. Finally, a 1 × 1 convolutional layer is used to linearly combine the concatenated feature maps generating the saliency map. A classifier, composed of a 100-node fully connected layer and an output activation node, is finally cascaded to the 1 × 1 convolution layer.

Architecture Training and Testing

In fine-tuning the visual attention model, cross-entropy is used as the objective function. The TaskNet model is first trained with SALICON dataset (Jiang *et al.* 2015), which has 10000 sample images and fixation maps. Afterward, the TaskFix dataset is used for fine-tuning. We performed fine-grain analysis by evaluating which set of TaskFix dataset is best for fine-tuning, *i.e.* the complete TaskFix ($TaskFix_{ALL}$) that consists of all the normal and abnormal scene fixation maps, the modified TaskFix ($TaskFix_{MOD}$) that consists of the abnormal scene fixation maps and black images as the fixation maps for normal scenes, and only the abnormal scene fixation maps of TaskFix ($TaskFix_{ABN}$). We also fine-tuned the SALICON-trained architectures using OSIE (Xu *et al.* 2014) to compare our proposed network with free-viewing visual attention models. Table 3 summarizes the experiments we performed.

Table 3. Summary of experiments conducted to determine the best configuration and fine-tuning samples to use.

Name	Feature detector	Training set	Fine-tuning
TaskNet	VGG-16	SALICON dataset	$TaskFix_{ALL}$
TNv1	GoogLeNet		$TaskFix_{ALL}$
TNv2	VGG-16		$TaskFix_{MOD}$
TNv3	GoogLeNet		$TaskFix_{MOD}$
TNv4	VGG-16		$TaskFix_{ABN}$
TNv5	GoogLeNet		$TaskFix_{ABN}$
TNv6	SSD		$TaskFix_{ALL}$
TNv7	SSD		$TaskFix_{MOD}$
TNv8	SSD		$TaskFix_{ABN}$
FNv1	VGG-16		OSIE
FNv2	GoogLeNet		OSIE
FNV3	SSD		OSIE

The $TaskFix_{ALL}$ training set has 718 samples, which consists of 359 abnormal and 359 normal road traffic scenarios. The other 359 abnormal samples in the $TaskFix_{ALL}$ dataset are used for the evaluation of the generated saliency maps. The $TaskFix_{MOD}$ dataset contains the same 718 stimuli images of the training set. However, the fixation maps of the normal images are replaced with black maps only. The fixation maps for the abnormal images are retained. The $TaskFix_{ABN}$ dataset only consists of abnormal road traffic images from the $TaskFix_{ALL}$ training set and test set.

The training and testing are implemented using the Caffe framework. The feature detectors are first trained and fine-tuned. All the configurations in Table 3 are trained using the SALICON dataset with a momentum of 0.9 and an initial learning rate of 1e–5. The learning rate decreases by a factor of 0.1 every 8000 iterations. Due to a large amount of training data and limited memory resources, the loading of input images and training were performed one image per iteration. Validation data shows that after three epochs, the performance started to stabilize. The trained model is then fine-tuned using the datasets indicated in Table 3. All networks use the same test set from the TaskFix dataset for the qualitative and quantitative evaluation. After training the feature detectors, its weights are frozen, and the classifier is trained (last layers after the saliency map in Figure 3) using $TaskFix_{ALL}$ for 10 epochs. Classifier weights are randomly initialized with mean equal to 0 and standard deviation equal to 1e–4.

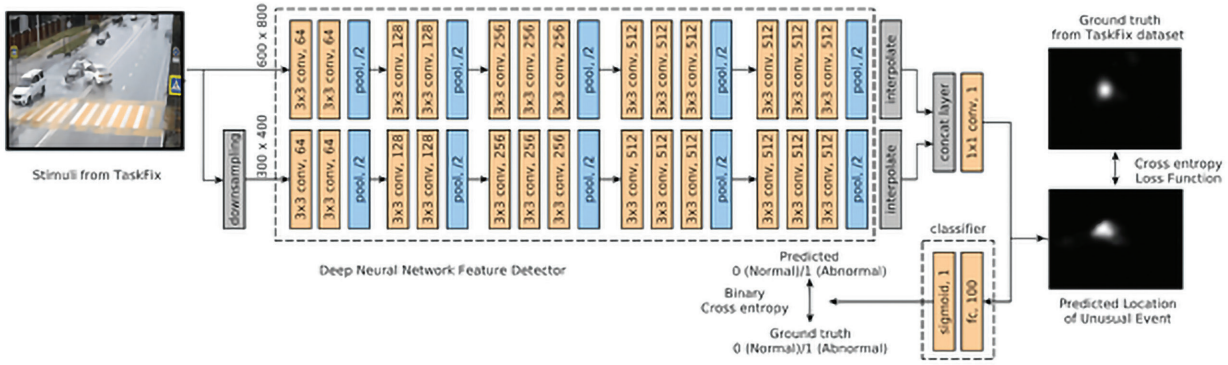

Figure 3. The TaskNet model is composed of parallel feature extractors that capture human attention responses at different resolutions. These are combined linearly *via* interpolation layers at the end of each branch, *via* concatenation layers, and a 1 × 1 convolution layer. The abnormal images from the TaskFix dataset are used for fine-tuning the proposed model. Fully connected layers are used as classifiers to predict if a scene contains an abnormal incident.

RESULTS AND DISCUSSION

We performed a comparison of quantitative and qualitative performance of TaskNet using different combinations of fine-tuning datasets (TNv's in Table 3) and DNN architectures, as well as with other free-viewing models (FNv's in Table 3) and saliency models.

What Do Saliency Metrics Mean?

The saliency metric scores reported in this work are AUC-Judd, sAUC, NSS, CC, KL, SIM, and EMD. AUC-based scores are the most used metric for saliency evaluation. It is invariant to contrast and monotonic transformation such that it is particularly good in detection applications. NSS and CC are highly related saliency metrics because of their analogous computation. NSS measures the correspondence between the prediction and the ground truth fixation. It is sensitive to false positives, relative differences in saliency across the image, and monotonic transformation. Similarly, CC measures the correlation between the prediction and the ground truth fixation. As opposed to NSS, CC equally penalizes false positives, and false negatives such that the increase in CC cannot distinguish whether the gain is due to false positives or false negatives.

KL and SIM rank differently the predicted fixation maps, as opposed to NSS and CC because these metrics (KL and SIM) are extremely sensitive to false positives. KL is a dissimilarity metric that evaluates the loss of information when the saliency map is used to approximate the ground truth fixation map. SIM measures the similarity between the saliency map and the ground truth fixation map. Both SIM and KL highly penalize misdetections. The EMD score depicts the amount needed to move the density of the saliency map to match the ground truth fixation map. The best score for this metric is 0 since the density of the saliency map does not need to be moved. These saliency metrics are exhaustively characterized by Bylinskii *et al.* (2012, 2019).

Results

The quantitative results are presented in Table 4. TaskNet shows promising performance in predicting the location of the observers' object of gaze – as indicated by AUC-Judd metric – although not the best in terms of sAUC, which penalizes models incorporating center bias. TaskNet used the TaskFix dataset, which shows slight center bias (see Figure 2b). In terms of CC and SIM, TaskNet also shows the best performance when the fixation map is considered as a probability distribution. When compared with free-viewing models, SALICON and SalNet (Pan *et al.* 2019), the evaluation shows that TaskNet is the most optimized for task-based attention prediction. Finally, with respect to NSS and KL, TaskNet shows the best results in predicting the location of the eye-fixation of observers looking at traffic accidents. NSS and KL metrics consider the range of values during the evaluation, thus capturing the relative values assigned to image regions. It is important to determine the most important image region because it could contain the location of a traffic accident.

We present in Figure 4 the qualitative performance of the different attention models. The bboxes in the stimuli column indicate the location of the traffic accident. From the figure, we make the following key observations. First,

Table 4. Quantitative comparison of TaskNet performance with other attention models, in predicting observers' attention while looking for traffic accidents. The values in bold are the best in each metric. ↑ means higher is better, ↓ means value is better.

	Metrics	AUC-Judd↑	sAUC↑	NSS↑	KL↓	EMD↓	CC↑	SIM↑
Task-driven	TaskNet	**0.91**	0.63	**3.29**	**0.99**	4.78	**0.67**	**0.52**
	TNv1	0.89	**0.65**	2.72	1.31	3.11	0.58	0.47
	TNv2	**0.91**	0.62	3.21	1.13	5.37	0.65	**0.52**
	TNv3	0.89	0.64	2.65	1.31	3.04	0.56	0.45
	TNv4	**0.91**	**0.65**	2.83	1.18	2.72	0.57	0.50
	TNv5	0.87	0.64	1.80	1.60	**0.94**	0.38	0.31
	TNv6	0.90	0.63	3.26	1.02	5.13	0.66	0.50
	TNv7	0.89	0.63	2.91	1.10	3.98	0.64	0.49
	TNv8	**0.91**	0.64	3.25	1.15	4.21	0.65	0.48
Free-viewing	FNv1	0.87	0.57	2.06	1.50	2.87	0.42	0.34
	FNv2	0.87	0.64	1.91	1.55	1.12	0.40	0.32
	FNV3	0.88	0.62	2.61	1.26	1.65	0.44	0.40
	SALICON	0.85	0.59	1.60	1.72	1.35	0.34	0.29
	SalNet	0.87	**0.67**	2.06	1.61	**0.91**	0.43	0.30

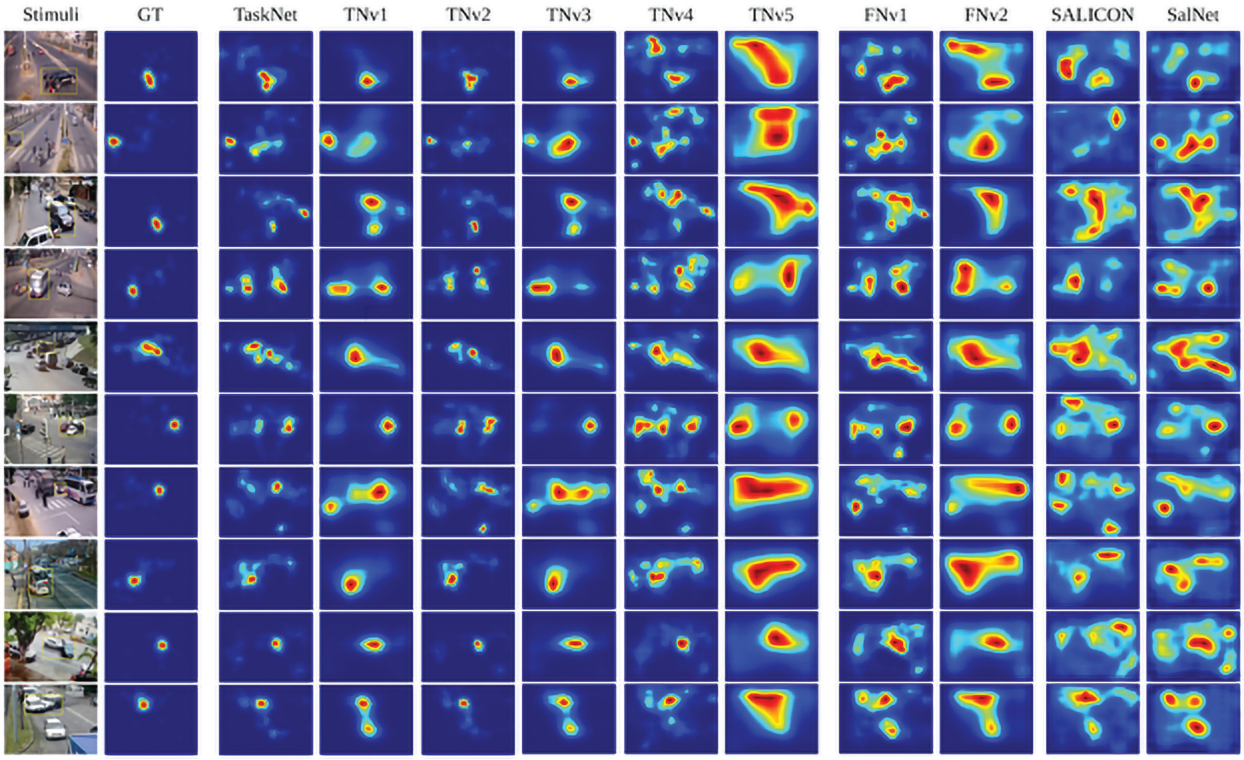

Figure 4. Qualitative comparison of predicted eye fixation from TaskNet, different versions of task-driven models, free-viewing-based models, and saliency models. The yellow boxes indicate the location of the traffic accident. Comparing TaskNet with other models shows improvement in predicting the location of traffic accidents.

as opposed to free-viewing models, visual attention models can be trained to perform an observer's task, attenuating image semantic. Comparing the TaskNet predicted location of traffic accidents to those predicted by the free-viewing based networks (FNv1, FNv2, and FNv3) and the outputs of the saliency models SALICON and SalNet, TaskNet consistently shows better performance. Note that free-viewing-based models mimic human attention spotting the salient part of an image based on semantic features, *e.g.* color contrast, texture, and object size. Second, repurposing visual attention models only need to be trained in scenes where the goal of the observer is present, *i.e.* using $\text{TaskNet}_{\text{ABN}}$. If trained with normal scenes, the model will perform worse. Compared with other TaskNet versions, TNv2 and TNv4 show false detection, which implies that it is best to train the network on how to look for abnormal events only. While TNv1, TNv3, and TNv5 have minimal false detection, TaskNet's quantitative performance is better.

The saliency maps are generated by resizing the output feature map of the 1×1 convolution filter to match input image size. A visualization of the top-five activation nodes from this output feature map of TaskNet, SALICON, and SalNet is shown in Figure 5a. This is performed by getting the top five node values from the output feature map and mapping their respective regions in the 600×800 input image. These nodes correspond to the predicted top-five most salient 32×32 regions in the scene. TaskNet top-five nodes all coincide with the abnormal incident location, as opposed to the top-five nodes of SALICON and SalNet, which do not cluster in one location.

Detecting Traffic Accidents *via* a Fully-connected Network

As discussed in the Experiments on the Dataset, the fixation maps of abnormal traffic scenes have low entropy due to clustered fixations, while fixation maps of normal traffic scenes have high entropy due to scattered fixations. The entropy of the TaskNet saliency maps – as predicted by the saliency models, SALICON, and SalNet – are compared with the entropy of the fixation maps, as shown in Figure 5b. The entropy for the abnormal scenes fixation maps is significantly lower than the entropy of normal scenes fixation maps ($p < 0.001$). For the saliency maps of TaskNet, the mean entropy of abnormal scenes also shows a significant difference from the mean entropy of normal scenes ($p < 0.005$). This indicates that the resulting salient region for the abnormal images is clustered in one area while the normal images have scattered salient regions.

Inspired by these important observations, a classifier is appended and trained using TaskFix. The performance high false-positive rate (= 0.18) and low false-negative

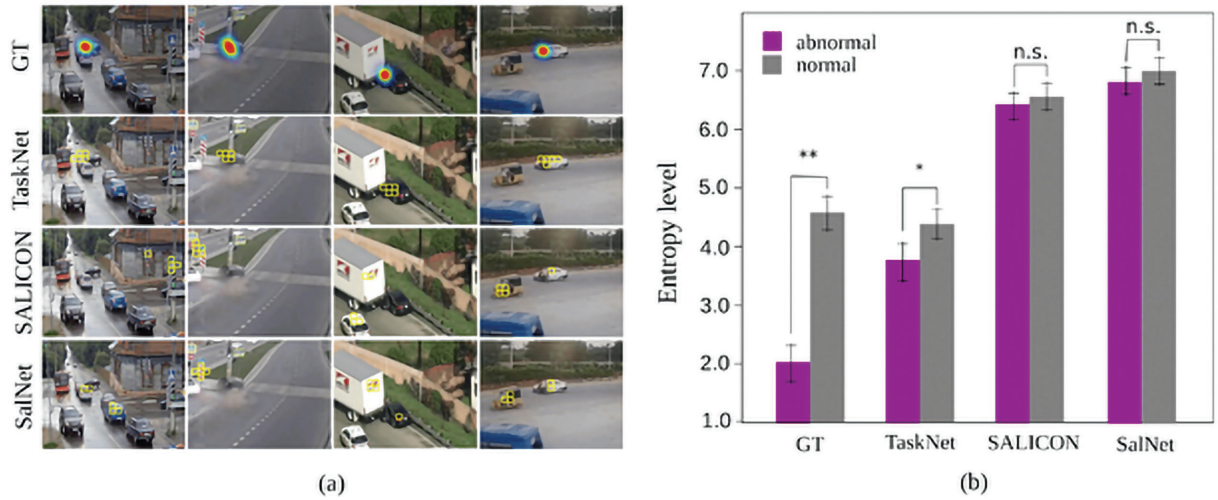

Figure 5. (a) Output feature map visualization shows that the TaskNet's top-5 nodes cluster around the location of traffic accidents (see second row) while the free viewing-based models' top-5 predictions are in other image portions. Also, (b) Wilcoxon's signed-rank test shows that the abnormal and normal scenes of the ground truth fixation maps and TaskNet predicted maps differ significantly in terms of entropy levels. ** means $p < 0.001$, * means $p < 0.005$, and n.s. means no significant difference.

rate (FNR; = 0.09) is observed (refer to Table 5 for the summary). Low FNR is aimed at systems that detect possible cases of abnormal events. The overall classification accuracy is 0.86, with precision equal to 0.83 and recall equal to 0.91.

Table 5. Classification performance of the proposed system. The false omission rate measures the proportion of misdetected accidents among those which are rejected.

Metrics	Values
True positive instances	326
False-negative instances	33
True negative instances	293
False-positive instances	66
Accuracy	0.86
Precision	0.83
Recall	0.91
False omission rate	0.33
False discovery rate	0.17

CONCLUSION

Most of the road cameras are monitored simultaneously by human observers. It is, thus, practicable to develop a traffic accident detection system that mimics a human observer performing abnormal event spotting tasks. In this work, we propose the first, human visual attention-based, traffic accident detection system.

We present a novel, task-driven fixation dataset of normal scenes and traffic accident scenes called TaskFix. TaskFix is composed of 718 image samples for each scene type containing normal scenes and traffic accident scenes. Statistics show that there is a significant difference in the ground truth fixation maps' entropy level of normal scenes and traffic accident scenes. Using TaskFix, we then encoded in a visual attention model called TaskNet, the human observers' task of catching a traffic accident. TaskNet performance *vis-à-vis* other free-viewing-based visual attention models, in mimicking task-driven observers, is demonstrated using qualitative, quantitative, statistical, and visualization experiments. TaskNet outperforms other visual attention models.

To our knowledge, this is the first attempt to use a fixation data-driven, visual attention model for abnormal incident detection. TaskNet is unique from other existing anomaly detection systems in that it is visual attention-based. It only needs fixation maps collected from observers under a task. Thus, TasKNet avoids the problems of imbalanced distribution and the sparse occurrence of abnormal events. Automatic detection of anomalous events, especially traffic accidents, is a critical task; thus, TaskNet use case is for alert systems to assist the authorities.

One major limitation of the current TaskNet is fixation dataset specific, *i.e.* it mainly works for traffic accident scenes that were used in the fixation training. We will make a more generic anomaly fixation dataset for our future work. Meanwhile, we foresee that as we make our model more generic to different anomalous events, both in scenes and in videos, anomaly detection using only saliency will be a challenge.

ACKNOWLEDGMENTS

This study was supported by the New Ph.D. Grant under the De La Salle University Research Coordination Office Project No. 05 N 1TAY19-2TAY20.

REFERENCES

AHMED SAA, DOGRA DPD, KAR S, ROY PP. 2018. Surveillance Scene Representation and Trajectory Anomaly Detection Using Aggregation of Multiple Concepts. Expert Systems with Applications 101: 43–55.

AHMED SA, DOGRA DP, KAR S, ROY PP. 2019. Trajectory-based Surveillance Analysis: A Survey. IEEE Transactions on Circuits and Systems for Video Technology 29(70): 1985–1997.

BASTANI V, MARCENARO L, REGAZZONI CS. 2016. Online nonparametric Bayesian activity mining and analysis from surveillance video. IEEE Transactions on Image Processing 25(5): 2089–2102.

BYLINSKII Z, JUDD T, BORJI A, ITTI L, DURAND F, OLIVA A, TORRALBA A. 2019. MIT Saliency benchmark. Available at http://saliency.mit.edu/

BYLINSKII Z, JUDD T, OLIVA A, TORRALBA A, DURAND F. 2019. What Do Different Evaluation Metrics Tell Us About Saliency Models? IEEE Transactions on Pattern Analysis and Machine Intelligence 41(3): 740–757.

CHANDOLA V, BANERJEE A, KUMAR V. 2009. Anomaly detection: a survey. ACM Computing Survey 41(3).

CORDEL MO, FAN S, SHEN Z, KANKANHALLI MS. 2019. Emotion-aware Human Attention Prediction. In: Proc. of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 15–20 Jun 2019; Long Beach, CA, USA.

CORNIA M, BARALDI L, SERRA G, CUCCHIARA R. 2016. A deep multi-level network for saliency prediction. In: 23rd IEEE International Conference on Pattern Recognition; 04–08 Dec 2016; Cancun, Mexico. p. 3488–3493.

DHOLE H, SUTAONE M, VYAS V. 2019. Anomaly Detection using Convolutional Spatio-temporal Autoencoder. In: 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT); 06–08 Jul 2019; Kanpur, India.

HE S, TAVAKOLI HR, BORJI A, MI Y, PUGEAULT N. 2019. Understanding and Visualizing Deep Visual Saliency Models. In: Proc. of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 15–20 Jun 2019; Long Beach, CA, USA. p. 10198–10207.

HUANG X, SHEN C, BOIX X, ZHAO Q. 2015. SALICON: Reducing the Semantic Gap in Saliency Prediction by Adapting Deep Neural Networks. In: Proc. of the IEEE International Conference on Computer Vision (ICCV); 07–13 Dec 2015; Santiago, Chile. p. 262–270.

IONESCU RT, KHAN FS, GEORGESCU M, SHAO L. 2019 Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In: Proc. of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 15–20 Jun 2019; Long Beach, CA, USA. p. 7834–7843.

ITTI L, KOCH C. 2000. A Saliency-based Search Mechanism for Overt and Covert Shifts of Visual Attention. Vision Research 40(10–12): 1489–1506.

JIANG M, HUANG S, DUAN J, ZHAO Q. 2015. SALICON: Saliency in Context. In: Proc. of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 07–12 Jun 2015; Boston, MA, USA. p. 1072–1080.

JUDD T. 2011. Understanding and predicting where people look in images [Dissertation]. Massachusetts Institute of Technology, Cambridge, MA. Available at https://dspace.mit.edu/

KRUTHIVENTI S, AYSUH K, BABU RV. 2015. DeepFix: A Fully Convolutional Neural Network for Predicting Human Eye Fixations. IEEE Transactions on Image Processing 26(9): 4446–4456.

KUMARAN SK, DOGRA DP, ROY PP. 2020. Anomaly Detection in Road Traffic Using Visual Surveillance: A Survey. ACM Computing Surveys.

LIU W, ANGUELOV D, ERHAN D, SZEGEDY C, REED S, FU CY, BERG AC. 2016. SSD: Single Shot MultiBox Detector. Lecture Notes in Computer Science. Proceedings of the 14th European Conference on Computer Vision; 11–14 Oct 2016; Amsterdam, The Netherlands. Springer. p. 22–36.

LU C, SHI J, JIA J. 2013. Abnormal Event Detection at 150 FPS in MATLAB. In: IEEE ICCV; 01–08 Dec 2013; Sydney, Australia

LUO W, LIU W, GAO S. 2017. A Revisit of Sparse Coding Based Anomaly Detection in Stacked RNN Framework. In: Proc. of the IEEE International Conference on Computer Vision (ICCV); 22–29 Oct 2017; Venice, Italy. p. 341–349.

MURABITO F, SPAMPINATO C, PALAZZO S, GIORDANO D, POGORELOV K, RIEGLER M. 2018. Top-down saliency detection driven by visual classification. Computer Vision and Image Understanding 172: 67–76.

NICHOLS T, WISNER P, GULABCHAND G. 2010. Putting the Kappa Statistic to Use. Quality Assurance Journal 13: 57–61.

PAN J, SAYROL E, GIRO-I-NIETO X, MCGUINNESS K, O'CONNOR NE. 2019. Shallow and deep convolutional networks for saliency prediction. In: Proc. of the 2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 27–30 Jun 2016; Las Vegas, NV, USA. p. 598–606.

SAN MIGUEL JC, MARTINEZ JM. 2012. A semantic-based probabilistic approach for real-time video event recognition. Computer Vision and Image Understanding 116(9): 937–952.

SANTHOSH KK, DOGRA DP, ROY PP. 2019. Temporal Unknown Incremental Clustering Model for Analysis of Traffic Surveillance Videos. IEEE Transactions on Intelligent Transportation Systems 20(5): 1762–1773.

SHINE L, EDISON A, JIJI CV. 2019. A Comparative Study of Faster R-CNN Models for Anomaly Detection in 2019 AI City Challenge. In: Proc. of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 15–20 Jun 2019; Long Beach, CA, USA. p. 306–314.

SIMONYAN K, ZISSERMAN A. 2014. Very Deep Convolutional Networks for Large-scale Image Recognition. ArXiv, abs/1409.1556.

SUK H, JAIN AK, LEE S. 2011. A Network of Dynamic Probabilistic Models for Human Interaction Analysis. IEEE Transactions on Circuits and Systems for Video Technology 21(7): 932–945.

SZEGEDY C, LIU W, JIA Y, SERMANET P, REED S, ANGUELOV D, ERHAN D, VANHOUCKE V, RABINOVICH A. 2015. Going deeper with convolutions. In: Proc. of the 2015 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 07–12 Jun 2015; Boston, MA, USA. p. 1–9.

TANG H, CHEN C, PEI X. 2016. Visual Saliency Detection *via* Sparse Residual and Outlier Detection. IEEE Signal Processing Letters 23(12): 1736–1740.

[UCSD] University of California in San Diego. 2013. UCSD Anomaly Detection Dataset. Retrieved on Feb 2020 from http://www.svcl.ucsd.edu/projects/anomaly/dataset.html

VISHNU VCM, RAJALAKSHMI M, NEDUNCHEZHIAN. 2018. Intelligent traffic video surveillance and accident detection system with dynamic traffic signal control. Cluster Computing 21: 135–147.

XU D, YAN Y, RICCI E, SEBE N. 2017. Detecting anomalous events in videos by learning deep representations of appearance and motion. Computer Vision and Image Understanding 156: 117–127.

XU J, JIANG M, WANG S, KANKANHALLI MS, ZHAO Q. 2014. Predicting human gaze beyond pixels. Journal of Vision 14(1): 1–20.

YANG M, RAJASEGARAR S, ERFANI SM, LECKIE C. 2019. Deep learning and one-class SVM-based anomalous crowd detection. In: Proc. of the 2019 International Joint Conference on Neural Networks (IJCNN); 14–19 Jul 2019; Budapest, Hungary.

YAO Y. XU M, WANG Y, CRANDALL DJ, ATKINS EM. 2019. Unsupervised Traffic Accident Detection in First-person Videos. In: Proc of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 03–08 Nov 2019; Macau, China. p. 273–280.

ZHAO J, YI Z, PAN S, ZHAO Y, ZHAO Z, SU F, ZHUANG B. 2019. Unsupervised Traffic Anomaly Detection Using Trajectories. In: Proc. of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 15–20 Jun 2019; Long Beach, CA, USA. p. 133–140.

ZHAO Q, KOCH C. 2013. Learning saliency-based visual attention: a review. Signal Processing 93(6): 1401–1407.

ZHOU S, SHEN W, ZENG D, FANG M, WEI Y, ZHANG Z. 2016. Spatial-temporal convolutional neural networks for anomaly detection and localization in crowded scenes. Signal Processing: Image Communication 47: 358–368.