# LaTeX Author Guidelines for CVPR Proceedings

Anonymous CVPR 2021 submission

Paper ID ****

## Abstract

## 1. Introduction

In order to make Autonomous Vehicles (AV) reliable in a real-world scenario, safety for every agent involved must be the focus of every self-driving system implementation. This objective will be achieved once the AV has a clear understanding of the driving scene around it, focusing on the main visual cues that are needed for its correct behaviour, differentiating between normal and anomalous situations, so that it can react in real-time to make the safest decision. It is necessary to track down all possible causes for the sake of accident avoidance, this must be done with both precision and promptness to assure the maximum reaction space.

Our work put the focus on video analysis by dash-mounted cameras, willing to improve the tools for driving scene interpretation in the context of Advanced Driver Assistance Systems (ADAS). From this perspective, a driving scenario is quite hard to model since there are many information to take into account that can be used to define the driving scene, there are plenty of possible accident classes that must be taken into account and to make matters worse, most of the times, it is quite hard to distinguish normal driving scenes from accident ones at frame level, which further enhances the problem's complexity.

Though we know that accidents are a consequence of an anomalous driving scenario, it is non-trivial to define what a driving anomaly is. We can define an anomaly as an hazardous situation that can lead to an accident, but since the hazardousness prior to the accident may be determined subjectively by each individual, the boundaries for an anomaly are not really clear and this is reflected in some dataset annotations. Some attempts have been made to propose a deterministic method in the interest of defining an anomaly. Yao et al. [14] defines an anomaly as the window in which the accident happens, but since we want to prevent it, this might not be ideal in a prevention perspective. Fang et al. [3] instead want to predict an accident willing to happen in the next 5 seconds labelling the anomaly start from the moment in which half part of the object involved in the accident appears in the view. Yao et al. [12] proposed a Detection of traffic Anomaly (DoTA) dataset that takes into account When the anomalous event starts and ends, locates spatially Where all the involved agents are in each frame and What type of anomaly is. Their work formulates the anomaly start as the instant after which the accident is unavoidable. As said before choosing that instant is quite subjective depending on the situation and personal biases.

Human's capability of evaluating danger on the fly is strictly related to still a matter of study by neuroscientists

Cornia et al.[1] proposed Multi-Level Network for Saliency Prediction (MLNET), an architecture for saliency estimation which simulates what human see at first glance, a step forward towards driving scene comprehension.

The architecture we propose to solve the problem is composed by a Video Swin Transformer [7] as the backbone network, adapted to work in a real-time scenario, as expected from an ADAS implementation. As a further contribution for our work we implemented MLNET to estimate the saliency map for each frame in order to lead the model to focus on the most pertinent regions of the traffic scene. Finally we propose a relabelling of DoTA dataset adopting a different criterion of evaluation which disentangles itself from the subjectivity and grants a deterministic method for estimating anomaly boundaries, allowing the largest possible reaction space while maintaining normal and anomalous scenes well separated.

## 2. Related

**Vision Transformers** Transformers [9] are born as an architecture to solve sequence-to-sequence problems, handling long-range dependencies in a simple way with the advantage of a strong parallelization compared to state-of-the-art architectures such as RNN and derivatives. Initially developed for text analysis tasks, transformers have also found application in the image field. The seminal work [2] first proposed a Vision Transformer (ViT), paving the way for a new generation of detectors, alternative to CNN. Afterwards, with the aim of improving the performance in terms of accuracy of the results and decreasing the computational need, variants such as the Swin Tranformers [6] were born. In order to reduce the computational cost of the self-attention mechanism, authors proposed a shifted-windowing scheme to compute self-attention on smaller non-overlapping windows, introducing cross-window connection to cope with the lack of connections between different regions of the image. As a direct evolution of Swin Tranformers, to process video instead of images, a new architecture was proposed in [7]. The authors proposed to approximate spatiotemporal self-attention by compute self-attention locally, extending spatial domain to the spatiotemporal domain.

**Traffic Anomaly Detection** To detect anomaly in video, in [4], authors proposed a convolutional AutoEncoder (ConvAE) trained only on normal frames with the objective of frames reconstruction. In [8, 10], authors used Convolutional LSTM Auto-Encoder as framework to encode appearance and motion. Authors of AnoPred [5] proposed a multi-task loss which include image intensity, optical flow, gradient, and adversarial losses for video frame-level anomaly detection which apply a UNet to predict a future RGB frame. In [14], authors proposed an unsupervised method which tracks traffic participants trajectories and detect anomaly monitoring prediction consistency. In TRN model [11], authors coupled the action detection task with the future action anticipation during the training. To predict the action, they use both the historical temporal dependencies modeled by a RNN and the anticipation of the future via a temporal decoder. In 2020, authors of [13] proposed a new dataset of video anomaly detection called Detection of Traffic Anomaly (DoTA).

## 3. Theory

## 4. Experiments

**Dataset**.
**Evaluation Metrics**.
**Implementation details.**

### 4.1. Ablation study

## 5. Conclusions

## References

[1] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. A deep multi-level network for saliency prediction, 2016. 1

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 1

[3] Jianwu Fang, Dingxin Yan, Jiahuan Qiao, and Jianru Xue. Dada: A large-scale benchmark and model for driver attention prediction in accidental scenarios, 2019. 1

[4] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 733–742, 2016. 2

[5] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6536–6545, 2018. 2

[6] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[7] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video Swin Transformer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3192–3201, New Orleans, LA, USA, June 2022. IEEE. 1, 2

[8] Weixin Luo, Wen Liu, and Shenghua Gao. Remembering history with convolutional lstm for anomaly detection. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 439–444. IEEE, 2017. 2

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[10] Lin Wang, Fuqiang Zhou, Zuoxin Li, Wangxia Zuo, and Haishu Tan. Abnormal event detection in videos using hybrid spatio-temporal autoencoder. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2276–2280. IEEE, 2018. 2

[11] Mingze Xu, Mingfei Gao, Yi-Ting Chen, Larry S Davis, and David J Crandall. Temporal recurrent networks for online action detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5532–5541, 2019. 2

[12] Yu Yao, Xizi Wang, Mingze Xu, Zelin Pu, Ella Atkins, and David Crandall. When, where, and what? a new dataset for anomaly detection in driving videos, 2020. 1

[13] Yu Yao, Xizi Wang, Mingze Xu, Zelin Pu, Yuchen Wang, Ella Atkins, and David Crandall. Dota: Unsupervised detection of traffic anomaly in driving videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2022. 2

[14] Yu Yao, Mingze Xu, Yuchen Wang, David J Crandall, and Ella M Atkins. Unsupervised traffic accident detection in first-person videos. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 273–280. IEEE, 2019. 1, 2