A Road Accident Detection Method Utilizing Deep Learning and Fast Fourier Transform

Mikhail Gorodnichev
Department of Mathematical Cybernetics and Information Technologies, Moscow Technical University of Communications and Informatics, Russia
m.g.gorodnichev@mtuci.ru (corresponding author)

Kamil Kharrasov
Department of Mathematical Cybernetics and Information Technologies, Moscow Technical University of Communications and Informatics, Russia
k.r.harrasov@mtuci.ru

Marina Moseva
Department of Mathematical Cybernetics and Information Technologies, Moscow Technical University of Communications and Informatics, Russia
m.s.moseva@mtuci.ru

Received: 23 January 2025 | Revised: 17 February 2025 | Accepted: 24 February 2025

 DOI: https://doi.org/10.48084/etasr.10328

ABSTRACT

This paper presents the development of a real-time crash detection method aimed at improving the efficiency of emergency assistance to the location of the accident. The analysis involved reviewing existing classical and neural network-based crash detection approaches, focusing on architectures such as EfficientNet-B1, EfficientNet-B7, MobileNetV2, and ConvNeXtV2. A dedicated dataset consisting of 12,426 crash-related image frames was specifically compiled for this study, combining previous datasets with self-collected images. The performance of accident detection models was evaluated on this dataset, leading to the development of a new crash detection method. The ConvNeXtV2-Femto model was selected as the core architecture for the proposed system, which was modified employing Fast Fourier Convolution (FFC) to improve its performance. Comparative analysis demonstrated that the proposed model achieved a 94% accuracy, outperforming existing approaches in all metrics, including also precision, recall, and F1-score.

Keywords-accident detection; convolutional neural networks; image processing; fast Fourier transform; transport monitoring; classification

I. INTRODUCTION

In recent decades, the rapid advancement of technology has transformed daily life, with new technological solutions penetrating nearly all areas of human activity. The field of street and road network safety has also benefited significantly from these developments. According to the Scientific Center for Road Traffic Safety of the Ministry of Internal Affairs of Russia [1], the number of photo-video monitoring complexes increased from 5,100 units in 2011 to 11,100 in 2016, and further to 21,400 by 2020. By 2021, 23,100 systems were installed, and by the end of 2022, their number reached 27,000. According to statistics from the Russian State Road Traffic Safety Inspectorate, between 2015 and 2023, the number of injuries and fatalities in road accidents decreased by 37% and 40%, respectively. However, the total number of road accidents also fell by 37% during the same period [1]. Thus, the ratio of injuries and fatalities to the total number of accidents has remained nearly unchanged.

Moreover, authors in [2] reported a dramatic increase in Road Traffic Accidents (RTAs) and related casualties in Bangladesh between 1982 and 2000, by 42% and 400%, respectively, with poor road conditions, unqualified drivers, and unsafe driving behaviors being identified as the main causes of RTAs. The substantial economic losses were also highlighted, estimated at US$518 billion globally and £1.2 billion in Bangladesh, equivalent to approximately 2% of Gross Domestic Product (GDP) and all annual foreign aid. Additionally, the authors emphasized the importance of rapid response in the location of the accident by emergency services, which can result in reducing health complications and mortality of the people involved in the accident. As a result, the introduction of immediate road accident detection systems could have a significant impact on the quicker notification of emergency services.

Existing traffic incident detection systems employ diverse algorithms for addressing detection, classification, and tracking. Neural network architectures such as You Only Look Once (YOLO) [3-5], combinations of DeepLab and YOLO [6], EfficientNet and MobileNet [7], custom-designed neural networks [8, 9], ensemble methods [10], and hybrid approaches combining image segmentation thresholds with shadow removal techniques [11] have been widely applied for detection and classification. For object tracking, techniques such as the centroid principle [3], Wenger's algorithm and the Kalman filter [4], temporal templates [12], and ByTrack [5] have been used. Furthermore, systems integrating trajectory prediction [13], Convolutional Neural Network (CNN)-based models combined with traditional video analysis, visual feature extraction, and temporal pattern identification [14] have been developed for enhanced traffic accident detection.

In this study, we propose a neural network-based method for real-time video-based accident detection aimed at improving the efficiency of road monitoring systems. The developed method utilizes the ConvNeXtV2 model with modified convolution operations based on Fast Fourier Transforms (FFT). This approach significantly improves file processing speed compared to existing solutions in this field.

II. DATASET

The dataset used in this study comprises three individual datasets: i) the TrafficNet dataset available in [15] and presented in [16], ii) the dataset in [17] presented in [6], and a self-collected dataset. The resulting dataset is available in [18].

Since the goal of this work is traffic accident detection, the TrafficNet dataset was slightly modified to better suit this purpose. TrafficNet originally includes four classes: Accident, Dense Traffic, Fire, and Sparse Traffic, each containing 1,100 images (4,400 total). The Accident class was retained, while Sparse Traffic was used to represent normal traffic conditions. In contrast, the Fire class was removed because it includes many non-traffic-related fire events, such as industrial or structural fires, while the Dense Traffic class was also omitted to maintain class balance and avoid asymmetry in the dataset.

The dataset described in [6] includes four subsets: weather, road_data, daytime, and crash-likelihood. Only the crash-likelihood subset was relevant for this study, consisting of frame-by-frame video images with and without crashes. It provides 6,400 training images (3,200 per class) and 2,800 testing images (1,400 per class).

In addition, we collected 1,026 images from the Internet and public street surveillance cameras, 426 depicting accidents and 600 without accidents. The resulting dataset consists of 12,426 images, of which 5,946 were classified as Accident and 6,300 as Non-Accident. The dataset was partitioned into 60/15/25 ratios for training, validation, and testing, respectively.

III. MODEL RESEARCH

In this work, three representative CNN architectures were selected for comparative evaluation. MobileNetV2 [19] was chosen as a lightweight baseline architecture optimized for real-time inference on edge and embedded devices, providing fast processing with limited computational resources. EfficientNet models (B1 and B7) [20] were included as state-of-the-art CNNs that employ compound scaling to achieve higher accuracy with moderate complexity. The model proposed in [9] was chosen because of its task-oriented design for Closed-Circuit Television (CCTV) traffic monitoring under real-time constraints, representing application-tailored CNN pipelines commonly deployed in transportation analytics.

Model performance was evaluated using accuracy, precision, recall, F1-score, and Frame Processing Time (FPT) on both the Central Processing Unit (CPU) and Graphics Processing Unit (GPU). All experiments were conducted on a server equipped with an AMD EPYC 7742 processor (64 cores, 128 threads) and an NVIDIA Tesla A100-SXM4-40GB GPU. The performance results are presented in Table I.

TABLE I. PERFORMANCE COMPARISON OF THE MODELS

	Accuracy	**Precision**	**Recall**	**F1-score**	**CPU FPT (s)**	**GPU FPT (s)**
MobileNetV2	**0.91**	0.89	**0.96**	**0.92**	0.01	0.01
EfficientNet-B1	**0.9**	**0.94**	0.92	0.9	0.02	0.01
EfficientNet-B7	0.87	0.91	0.84	0.87	0.1	0.03
[9]	0.86	0.85	0.88	0.86	0.01	0.01

MobileNetV2 achieved the highest accuracy of 91%, indicating strong predictive reliability. EfficientNet-B1 also performed very well (90%), whereas EfficientNet-B7 (87%) and the model from [9] (86%) showed slightly lower accuracy, suggesting reduced classification consistency. In terms of precision, EfficientNet-B1 achieved the highest score of 94%, demonstrating superior reliability in positive predictions. EfficientNet-B7 followed (91%), while MobileNetV2 (89%) showed a slightly higher rate of false positives. The model in [9] recorded the lowest precision (85%).

Moreover, MobileNetV2 achieved the best recall (96%), followed by EfficientNet-B1 (92%), while EfficientNet-B7 (84%) and [9] (88%) achieved lower scores, likely due to insufficient data generalization during training. In terms of F1-score, MobileNetV2 again achieved the highest F1-score (92%), indicating a strong overall balance between precision and recall. EfficientNet-B1 followed with 90%, while EfficientNet-B8 and [9] reached 87% and 86%, respectively. The inference time on the CPU of MobileNetV2, EfficientNet-B1, and [9] demonstrated efficient processing (0.01-0.02 s per frame), while EfficientNet-B7 was notably slower (0.10 s), which may be critical for real-time applications. All models exhibit similar inference speeds on the GPU, confirming their suitability for accelerated parallel processing.

Overall, MobileNetV2 performed best across most evaluation metrics, combining high accuracy with fast inference, making it the most suitable choice for real-time accident detection applications. EfficientNet-B1 remained competitive and could serve as an alternative with minor optimization. In contrast, EfficientNet-B7 and the model in [9] require substantial refinement, particularly regarding recall and computational efficiency, to achieve satisfactory real-time performance.

IV. DEVELOPMENT OF A METHOD FOR ROAD TRAFFIC ACCIDENT DETECTION

Besides the previous CNN models' performance, recent transformer-based vision models have demonstrated superior accuracy by using global feature interaction, yet their high computational cost limits real-time applicability. However, the ConvNeXt family of architectures [21] bridges this gap by adopting key design principles from visual transformers, such as large kernel sizes, layer normalization, and enhanced feature aggregation, while maintaining a purely convolutional structure. This balance of efficiency and representational power makes ConvNeXtV2 a strong candidate for advancing accident detection systems.

A key innovation in ConvNeXtV2 is the Fully Convolutional Masked Autoencoder (FCMAE) for self-supervised learning. Unlike the original transformer-based masked autoencoder, which used a transformer-based asymmetric encoder-decoder, FCMAE is entirely based on convolutional operations. The masked input is processed as sparse patches using sparse convolution, allowing only visible parts of the image to be computed, thereby improving learning efficiency. The transformer decoder is replaced by a single ConvNeXt block, making the architecture fully convolutional.

A. Proposed Method

Our proposed method of road traffic accident detection was based on ConvNeXtV2, due to its established potential. Specifically, ConvNeXtV2 was used as the underlying network for feature extraction, the output of which generates a vector of size 512. The said vector is passed to a fully connected neural network consisting of 512, 256, 128, and 2 neurons. The output layer provides a binary classification, where 0 denotes "no accident detected" and 1 denotes "accident detected". The test results of the five ConvNeXtV2 versions, Atto, Femto, Pico, Nano, and Tiny, are presented in Table II.

TABLE II. PERFORMANCE COMPARISON OF CONVNEXTV2 VERSIONS

Version	Accuracy	Precision	Recall	F1-score	CPU FPT (s)	GPU FPT (s)
Atto	0.86	0.87	0.84	0.83	**0.01**	**0.004**
Femto	**0.87**	**0.92**	**0.88**	**0.87**	**0.01**	**0.004**
Pico	0.86	0.86	**0.89**	0.83	0.04	**0.004**
Nano	**0.87**	0.89	0.84	0.83	0.05	0.01
Tiny	0.84	0.82	**0.90**	0.85	0.10	0.01

Analysis of the results in Table II indicates that the Femto and Nano models achieved the highest accuracy (87%), followed closely by Atto and Pico, achieving 86%, while Tiny reached 84%. In terms of precision, Femto achieved the highest value (92%), indicating superior reliability in positive predictions, followed by Nano (89%), Atto (87%), and Pico (86%), while Tiny achieved lower precision (82%).

Regarding recall, Tiny achieved the highest score (90%), demonstrating strong sensitivity to positive classes, while Pico followed closely (89%). Femto and Atto scored 88% and 84%, respectively, while Nano achieved the lowest recall of 83%. Additionally, Femto achieved the highest F1-score of 87%, Tiny achieved the second best of 85%, while the other three versions tied with an 83% F1-score. In terms of inference time, Atto and Femto demonstrated the shortest CPU processing time of 0.01 s, indicating high computational efficiency. In contrast, Pico required 0.04 s, Nano required 0.05 s, and Tiny required 0.10 s. In addition, all models exhibited competitive GPU inference times, with Femto, Pico, and Atto showing the best performance.

Overall, Femto and Pico showed the most balanced performance across all metrics, making them preferable for applications requiring both accuracy and completeness of predictions. On the other hand, Atto and Nano displayed some limitations in balancing precision and recall, while Tiny displayed the lowest results overall, limiting its application in complex tasks

B. Modified Femto Model with Fast Fourier Convolution

Based on the previous analysis, the Femto model was selected as the baseline for modification. The most effective modification that was implemented was replacing CNN layers with Fast Fourier Convolution (FFC) layers.

The FFC [22] method extends the receptive field and combines information from different scales using the FFT. Its key component, the spectral transformer, aims to efficiently increase the receptive field of the convolution operation to fully resolve the input feature tensor using the discrete Fourier transform. The input tensor, represented as $X \in R^{H \times W \times C}$, where $H \times W$ denotes the spatial resolution and C the number of channels, is divided into spatial and spectral components: $X^l \in R^{H \times W \times (1-\alpha_{in})C}$ and $X^g \in R^{H \times W \times \alpha_{in}C}$. The parameter $\alpha_{in} \in [0, 1]$ represents the proportion of feature channels, controlling the balance between local and global information at different network layers. The output vector $Y \in R^{H \times W \times C}$ is similarly decomposed into a local and global components, with the ratio of global output channels determined by the hyperparameter $\alpha_{out} \in [0, 1]$. The operation can be expressed as:

$$Y^l = Y^{(l \to l)} + Y^{(g \to l)} = f_l(X^l) + f_{g \to l}(X^g) \quad (1)$$

$$Y^g = Y^{(g \to g)} + Y^{(l \to g)} = f_g(X^g) + f_{l \to g}(X^l) \quad (2)$$

where $Y^{(l \to l)}$, $Y^{(g \to l)}$ and $Y^{(l \to g)}$ apply standard convolution, while $Y^{(g \to g)}$ applies the discrete Fourier transform.

FFC layers are fully compatible with standard convolutional operations and can be integrated into existing CNNs without architectural modifications. In terms of computational complexity, FFC is comparable to conventional convolution in both parameter count and operation cost, with its main advantage emerging with large kernels, as the spectral transformer can be trained with a global receptive field using only a 1x1 kernel.

Table III presents the performance results of the proposed modified Femto model alongside the previously evaluated architectures. The proposed model achieved the highest accuracy of 94%, tied with EfficientNet-B1 in terms of precision (94%), and achieved the highest recall and F1-score (both 94%). Additionally, the inference times further confirm

the model's suitability for real-time use, with 0.01 seconds on CPU and 0.003 seconds on GPU. These results show that the proposed approach combines both accuracy and speed, meeting the stringent requirements of real-time video surveillance and emergency response systems.

To further assess the performance of the proposed modified Femto model, the confusion matrix is illustrated in Figure 1. Among 3,050 test cases, the model correctly classified 1,667 out of 1,738 accident cases (96%) and 1,200 out of 1,312 non-accident cases (91%). These results further demonstrate the potential of the proposed model, providing a promising foundation for efficient and accurate real-time traffic accident detection.

TABLE III. PERFORMANCE EVALUATION OF THE PROPOSED MODEL

	Accuracy	**Precision**	**Recall**	**F1-score**	**CPU FPT (s)**	**GPU FPT (s)**
Proposed	**0.94**	**0.94**	**0.96**	**0.94**	**0.01**	**0.003**
MobileNetV2	0.91	0.89	**0.94**	**0.92**	**0.01**	0.01
EfficientNet-B1	0.9	**0.94**	0.92	0.9	0.02	0.01
EfficientNet-B7	0.87	0.91	0.84	0.87	0.1	0.03
[9]	0.86	0.85	0.88	0.86	**0.01**	0.01

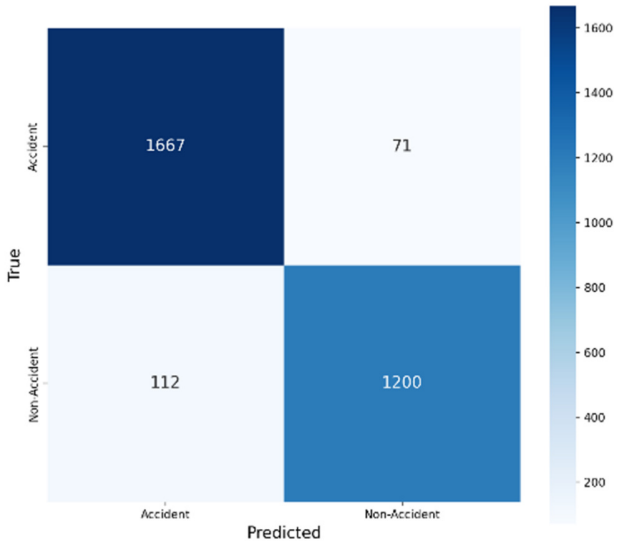

Fig. 1. Confusion matrix of the proposed model.

V. CONCLUSION

The study addressed the detection of traffic accidents on road networks by reviewing and assessing experimentally several accident detection frameworks, including MobileNetV2, EfficientNet-B1, EfficientNet-B7, a model proposed in [9], and several versions of ConvNeXtV2. Additionally, we proposed a modified Femto ConvNeXtV2 model that employs Fast Fourier Convolution (FCC) for convolution operations, which improved frame processing speed compared with existing approaches. The proposed model also achieved the highest F1-score (94%), accuracy (94%), recall (96%), and precision (94%) among all evaluated methods, while also displaying competitive inference time on Central Processing Unit (CPU) (0.01 s) and Graphics Processing Unit (GPU) (0.003 s). However, throughout our analysis, it became apparent that more road accident data encompassing a wider variety of accident types and road conditions are needed to improve the generalizability of the model and its adaptability to real-world conditions.

Lastly, future work will focus on improving the architecture of the accident detection system through the integration of more advanced deep learning models to further increase detection accuracy. In addition, incorporating additional contextual factors such as weather conditions, time of day, and other external parameters will also be explored to develop a more comprehensive and reliable accident detection system.

REFERENCES

[1] The Ministry of Internal Affairs of the Russian Federation, "Information on road safety indicators," stat.gibdd.ru. [Online]. Available: http://stat.gibdd.ru.

[2] K. Fathallah, S. Khamlich, E. Mohammed, and B. Mohamed, "Intelligent System for the Automatic Detection and Control of Accidents on the Road in Real Time," *Journal of Theoretical and Applied Information Technology*, vol. 99, no. 11, pp. 2578–2594, Jun. 2021.

[3] Z. Rahman, A. M. Ami, and M. A. Ullah, "A Real-Time Wrong-Way Vehicle Detection Based on YOLO and Centroid Tracking," in *2020 IEEE Region 10 Symposium (TENSYMP)*, Dhaka, Bangladesh, 2020, pp. 916–920, https://doi.org/10.1109/TENSYMP50017.2020.9230463.

[4] H. Ghahremannezhad, H. Shi, and C. Liu, "Real-Time Accident Detection in Traffic Surveillance Using Deep Learning," in *2022 IEEE International Conference on Imaging Systems and Techniques (IST)*, Kaohsiung, Taiwan, Jun. 2022, pp. 1–6, https://doi.org/10.1109/IST55454.2022.9827736.

[5] U. Jartarghar, D. Sanghvi, M. Nidgundi, K. Kumar, and S. Varur, "Vision Transformer-Based Multi-Phase Accident Detection and False Positive Mitigation." In Review, Jan. 2024, https://doi.org/10.21203/rs.3.rs-3903862/v1.

[6] M. M. Karim, Y. Li, R. Qin, and Z. Yin, "A system of vision sensor based deep neural networks for complex driving scene analysis in support of crash risk assessment and prevention." arXiv, Jun. 2021, https://doi.org/10.48550/arXiv.2106.10319.

[7] T. Tamagusko, M. G. Correia, M. A. Huynh, and A. Ferreira, "Deep Learning applied to Road Accident Detection with Transfer Learning and Synthetic Images," *Transportation Research Procedia*, vol. 64, pp. 90–97, 2022, https://doi.org/10.1016/j.trpro.2022.09.012.

[8] B. Kumeda, Z. Fengli, A. Oluwasanmi, F. Owusu, M. Assefa, and T. Amenu, "Vehicle Accident and Traffic Classification Using Deep Convolutional Neural Networks," in *2019 16th International Computer Conference on Wavelet Active Media Technology and Information Processing*, Chengdu, China, Dec. 2019, pp. 323–328, https://doi.org/10.1109/ICCWAMTIP47768.2019.9067530.

[9] M. Tahir, Y. Qiao, N. Kanwal, B. Lee, and M. N. Asghar, "Real-Time Event-Driven Road Traffic Monitoring System Using CCTV Video Analytics," *IEEE Access*, vol. 11, pp. 139097–139111, 2023, https://doi.org/10.1109/ACCESS.2023.3340144.

[10] M. Machoke, J. Mbelwa, J. Agbinya, and A. E. Sam, "Performance Comparison of Ensemble Learning and Supervised Algorithms in Classifying Multi-label Network Traffic Flow," *Engineering, Technology & Applied Science Research*, vol. 12, no. 3, pp. 8667–8674, Jun. 2022, https://doi.org/10.48084/etasr.4852.

[11] M. U. Farooq, A. Ahmed, S. M. Khan, and M. B. Nawaz, "Estimation of Traffic Occupancy using Image Segmentation," *Engineering, Technology & Applied Science Research*, vol. 11, no. 4, pp. 7291–7295, Aug. 2021, https://doi.org/10.48084/etasr.4218.

[12] S. Bakheet and A. Al-Hamadi, "A deep neural framework for real-time vehicular accident detection based on motion temporal templates," *Heliyon*, vol. 8, no. 11, Nov. 2022, Art. no. e11397, https://doi.org/10.1016/j.heliyon.2022.e11397.

[13] Y. Yao, M. Xu, Y. Wang, D. J. Crandall, and E. M. Atkins, "Unsupervised Traffic Accident Detection in First-Person Videos." arXiv, Jul. 26, 2019, https://doi.org/10.48550/arXiv.1903.00618.

[14] S. Robles-Serrano, G. Sanchez-Torres, and J. Branch-Bedoya, "Automatic Detection of Traffic Accidents from Video Using Deep Learning Techniques," *Computers*, vol. 10, no. 11, Nov. 2021, Art. no. 148, https://doi.org/10.3390/computers10110148.

[15] *Release Traffic-Net Dataset V1 · OlafenwaMoses/Traffic-Net*. (2024), M. Olafenwa . [Online]. Available: https://github.com/OlafenwaMoses/Traffic-Net/releases/tag/1.0.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition." arXiv, Dec. 2015, https://doi.org/10.48550/arXiv.1512.03385.

[17] *Image Dataset for driving scene classification*. (2021), M. M. Karim, Y. Li, R. Qin, and Z. Yin. [Online]. Available: https://github.com/monjurulkarim/Crash_road_function_dataset.

[18] *road_accident_data*. (2025), M. Gorodnichev, K. Kharrasov, and M. Moseva. [Online]. Available: https://github.com/KKharrasov/road_accident_data.

[19] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks." arXiv, Mar. 21, 2019, https://doi.org/10.48550/arXiv.1801.04381.

[20] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks." arXiv, Sep. 2020, https://doi.org/10.48550/arXiv.1905.11946.

[21] S. Woo *et al.*, "ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders." arXiv, Jan. 2023, https://doi.org/10.48550/arXiv.2301.00808.

[22] L. Chi, B. Jiang, and Y. Mu, "Fast Fourier Convolution", in *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, Canada, Dec. 2020, pp. 4479-4488.