DOI: https://doi.org/10.15276/hait.08.2025.29
UDC 004:83

Road traffic accident classification using a sparse video transformer and adaptive fragmentation

Tetiana V. Normatova[1)]
ORCID: https://orcid.org/0009-0004-3503-6350; tetiana.normatova@nure.ua
Sergii V. Mashtali[1)]
ORCID: https://orcid.org/0000-0002-0917-6622; sergii.mashtalir@nure.ua. Scopus Author ID: 36183980100
[1)] Kharkiv National University of Radio Electronic, 14 Nauky Ave. Kharkiv, 61166, Ukraine

ABSTRACT

In this work, we propose a simple yet effective approach for classifying short road traffic video clips into car accident and normal scenes. From each clip 8 frames are uniformly sampled across the sequence to ensure that key events are preserved even in longer videos. Based on the Farnebäck optical flow map, an adaptive fragment selection is performed, where the patch size (eight/sixteen/thirty-two pixels) is determined for each region of the base grid. Smaller patches are used in areas with intensive motion (to capture finer details), while larger patches are used in static regions (to reduce computations). The selected fragments are non-overlapping, resized to a uniform scale, and converted into feature vectors. The architecture operates in two stages. First, a spatial transformer processes each frame independently; attending only to the selected fragments this drastically reduces the number of feature tokens. Second, a temporal transformer processes the sequence of classify tokens (compact per-frame representations), aggregating temporal dynamics across frames. This space-to-time factorization significantly lowers computational cost and memory consumption while maintaining high informativeness in motion-intensive regions. To address class imbalance, we employ a weighted cross-entropy loss (or focal loss emphasizing hard examples) and weighted random sampling during training. Optical flow maps and fragment lists are precomputed and cached on disk, which accelerates training epochs even on CPUs without specialized hardware. Evaluation was conducted on the Car Crash Dataset (one thousand and five hundred accident and three thousand normal videos) using an eighty to twenty percent train-test split with preserved class proportions. The proposed method achieved Accuracy = 0.864 and Macro-F1 = 0.851. Preliminary comparisons show that our approach outperforms both the baseline uniform-patch Vision Transformers and traditional temporal aggregation schemes. The key advantage of the method lies in combining motion-guided feature reduction with a two-stage spatial-temporal processing pipeline, making the model suitable for realistic computational constraints (CPU-level inference) while maintaining high sensitivity to short and localized accident events. The approach is easily scalable and can be integrated with self-supervised pertaining techniques (e.g., masked video reconstruction). All experimental conditions, hyperparameters, and configurations are documented to ensure full reproducibility.

Keywords: Video classification; neural networks; convolutional neural networks; object classification; video stream analysis; data classification; image fragment processing

For citation: Normatova T. V., Mashtalir S. V. "Road traffic accident classification using a sparse video transformer and adaptive fragmentation ". ***Herald of Advanced Information Technology*****. 2025; Vol. 8 No.4: 464–475. DOI: https://doi.org/10.15276/hait.08.2025.29**

INTRODUCTION

Artificial intelligence (AI) and computer vision are rapidly transforming the way traffic environments are monitored, analyzed, and understood. In recent years, intelligent traffic systems have become a cornerstone of modern smart city infrastructure, enabling automated accident detection, vehicle tracking, congestion analysis, and overall road safety management. The ability to automatically abnormal situations in real time, such as traffic collisions or near-miss incidents is particularly critical because it allows for immediate response and can significantly reduce the severity of consequences.

Traditionally, traffic video analysis has relied on handcrafted feature extraction and conventional machine learning techniques such as background subtraction, motion history analysis, or Support Vector Machines. However, these approaches are highly sensitive to lighting conditions, occlusions, and camera viewpoints, which limits their robustness in real-world deployments. With the advent of deep learning, Convolutional Neural Networks (CNNs) and 3D Convolutional Neural Networks (3D CNNs) have become the dominant tools for video understanding tasks, including accident detection. These models are capable of automatically learning spatio-temporal features directly from data, eliminating the need for manual feature engineering.

Despite their success, most existing deep learning-based solutions process the entire video frame in a uniform manner, frame by frame or as a 3D volume. Such exhaustive processing leads to high computational and memory demands, especially for high-resolution video streams. The high dimensionality of video data results in models with tens or hundreds of millions of parameters,

which require powerful GPUs and large-scale infrastructure to operate efficiently. As a result, these systems are often impractical for CPU-based environments or real-time applications, such as traffic monitoring from roadside cameras or embedded systems in vehicles. Moreover, redundant processing of static background regions wastes resources without contributing valuable information to the classification process.

In this work, we aim to address these limitations by introducing an adaptive and motion-aware approach to video representation. Instead of processing all regions of a frame equally, we propose to reduce unnecessary computation by selectively extracting fragments (patches) based on the intensity of motion detected between consecutive frames. Regions with high motion speed are represented by smaller fragments, which preserve detailed spatial information, while low-motion regions are represented by larger fragments, reducing the total number of tokens passed through the model. This strategy allows the model to focus its attention on dynamic and informative areas such as moving vehicles or potential collisions while ignoring large portions of static background like the road surface or sky, which are less relevant for accident detection.

In this work, we propose reducing unnecessary computations by selectively extracting fragments from each frame: regions with high motion speed are represented by smaller fragments, while regions with low motion speed use larger ones.

This approach allows the network to focus its attention on dynamic areas, which are the most critical for accident detection.

The architecture of the proposed method includes a two-stage Sparse-ViT. The spatial attention block operates on image fragments within each frame, while the temporal block processes the sequence of classification tokens (CLS) that summarize frame-level information. Before vectorization, a non-overlapping tiled adaptive fragmentation is performed, guided by the optical flow: for each base grid cell, the magnitude of the flow vector is compared with quantile thresholds (q_1/q_2), and the fragment size is selected (e.g., 8/16/32 pixels). The fragments are then cropped without overlap, resized to a base size, and converted into feature vectors. This selection reduces redundant features and focuses computations on regions with motion, improving both spatial frame analysis and temporal information aggregation.

1. RELATED WORKS

There are several approaches for segmentation in video streams, which can be grouped as follows:

– transformer-based approaches for video: Vision Transformers (ViT) [1], [2] laid the foundation for representing images as a sequence of small patches, each transformed into a feature vector. TimeSformer [3] separates attention into spatial and temporal components [4], [5]. Video Vision Transformers (ViViT) [6] factorizes spatio-temporal attention by applying separate blocks for per-frame patch processing and temporal aggregation. Video Swin Transformer [7] introduces hierarchical shifted windows, while Multiscale Vision Transformers (MViT) [8] builds multi-level representations. These lines of work demonstrate that separating space and time, as well as using multi-scale representations, improves efficiency. However, the quadratic computational cost of applying full attention over the entire sequence of features remains a bottleneck;

– sparse and efficient attention for long sequences: models such as Sparse Transformer [9], Longformer [10], and BigBird [11] reduce computational complexity by employing local and random attention patterns. Selective attention with learned biases has shown that attending to a small number of the most relevant key elements can effectively replace full global attention. For video tasks, this is particularly important when processing long clips. In the approach described in this work, “sparsity” is achieved not by modifying the internal attention mechanism, but by reducing the number of sequence elements beforehand: only frame patches with high motion saliency are retained, so the sequence length is shortened even before being processed by the transformer;

– optical flow and two-stream signals: FlowNet2 [12], PWC-Net [13], and Recurrent All-Pairs Field Transforms (RAFT) [14] have made optical flow a reliable source of motion features. Motion is useful both as a separate input branch (two-stream) and as a cue for spatial selection and feature masking. In the implementation described here, classical Farnebäck flow is used as an “importance map” that guides adaptive patching, providing a significant speedup on CPU while maintaining acceptable quality;

– token selection and representation reduction: to avoid processing all frame patches, the TokenLearner [15] method dynamically aggregates a small subset of the most informative patches, while DynamicViT learns to skip less important patches during inference. For video, motion-guided

approaches have been proposed (e.g., ViViT variants, MotionFormer), as well as methods focusing on salient regions (AdaFocus), which enhance resolution in critical areas. The approach in this work follows this line: adaptive patch selection is applied, where the window size (8/16/32 pixels) is determined by quantile thresholds of optical flow magnitude. Patches are extracted without overlap and resized to a base size, which significantly reduces the number of processed representations while preserving regions with pronounced motion;

– class imbalance: common approaches include balanced cross-entropy loss or reweighting; Focal Loss is a specialized loss function [16]; oversampling can be performed using WeightedRandomSampler. For video tasks, temporal segmentation methods such as Temporal Segment Networks (TSN) [17] are also applied to avoid overrepresentation of long normal segments. In the implementation described here, loss reweighting and weighted random sampling in the data loader are already employed; Focal Loss and TSN-style segmentation can be enabled if needed;

– pretraining and masked modeling (perspective): MAE [18] and VideoMAE [19] have shown that masking patches followed by reconstruction significantly improves feature robustness, especially when labeled data are limited. In this work, the MAE approach has not yet been applied in the presented experiments; however, it is fully compatible with the frame patch generation and selection module used here and represents a logical next step;

– positioning summary: unlike approaches that focus on improving the attention mechanism itself, the present work reduces computational cost through motion-based patch selection and a two-stage processing pipeline: the spatial stage operates on the selected patches of each frame, while the temporal stage aggregates only the per-frame CLS tokens [20]. This design yields a compact, CPU-friendly configuration without overly complex components and with moderate requirements for computational resources and data volume.

2. PROBLEM STATEMENT

Traffic accident detection from surveillance streams is a challenging problem due to the high dimensionality of video, scene variability (weather, lighting, camera angle), and class imbalance accidents occur far less frequently than normal traffic. In addition, real-world deployments are often constrained by latency and limited hardware resources (edge devices, CPU-only servers). Therefore, a practical approach must achieve a balance between accuracy, computational efficiency, and robustness while maintaining sensitivity to short, safety-critical events.

The aim of this work is to develop a lightweight yet accurate method for binary video classification determining whether a given video clip depicts an accident or a normal situation that can operate effectively under real-time or CPU-based conditions.

The main tasks of the study are as follows.

1. To design an adaptive fragment selection mechanism that dynamically adjusts spatial resolution according to motion intensity, preserving details in dynamic regions while reducing redundancy in static areas.

2. To implement a two-stage Sparse Video Transformer architecture that first extracts spatial features within individual frames and then aggregates temporal dependencies across frames using CLS tokens.

3. To optimize the model for limited-memory environments by reducing the number of tokens processed per frame by approximately 30–60%, depending on scene complexity.

4. To evaluate the proposed approach against conventional baselines such as uniform-patch ViT and 3D CNN-based models in terms of accuracy, macro-F1, precision, recall, and inference speed on both CPU and GPU.

5. To analyze the impact of class imbalance and introduce techniques such as weighted loss functions and sampling strategies to improve the model's sensitivity to rare accident events.

The proposed method relies on optical flow estimation (Farnebäck algorithm) to detect motion-salient regions and guide patch size selection (8×8, 16×16, or 32×32 pixels). Smaller patches are assigned to high-motion areas, while larger patches cover static backgrounds. The selected non-overlapping patches are embedded and processed by the spatial transformer block to produce frame-level representations. These CLS tokens are then passed to a temporal transformer that captures inter-frame dynamics and outputs the final classification result.

The approach assumes the presence of visible motion cues associated with accident dynamics (e.g., rapid deceleration, collision, sudden trajectory change). In scenes with very subtle or absent motion, the quality of optical flow estimation becomes the limiting factor. Additionally, precomputing optical flow maps introduces moderate preprocessing overhead. However, due to

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

the quantile-based thresholding mechanism, patch selection remains adaptive across scenes with varying motion magnitudes.

In summary, the study aims to achieve efficient clip-level classification of road traffic accidents under realistic computational constraints by combining motion-guided adaptive fragmentation with a sparse transformer architecture. The resulting framework selectively allocates computational capacity to the most informative regions, maintaining accuracy while substantially reducing computational cost and inference time.

3. PROPOSED METHOD

The proposed method introduces an adaptive motion-aware video transformer architecture designed to efficiently detect traffic accidents in surveillance streams. Like classical Vision Transformer (ViT) architectures, our model is composed of two main stages: a spatial transformer block and a temporal transformer block preceded by a motion-guided adaptive patch selection module. The goal of the framework is to allocate computational resources dynamically, focusing the model's attention on regions exhibiting high motion activity, which are more likely to contain critical events such as collisions, lane departures, or abrupt stops.

Each video clip is uniformly sampled to 8 frames from the original video stream to provide a reasonable trade-off between temporal context and processing speed. Uniform sampling ensures that the model receives frames that represent the temporal dynamics of the event without introducing redundancy. All frames are resized to a fixed spatial resolution (128×128 px) and converted to grayscale copies for optical flow computation, while the RGB versions are retained for subsequent patch embedding in the transformer.

To identify motion-relevant regions, we use the Farnebäck dense optical flow algorithm, which estimates per-pixel motion vectors between consecutive frames. Let $X_t \in \mathrm{R}^{\mathrm{H}\times\mathrm{W}\times 3}$ be the RGB frame at time *t*. We compute a dense Farnebäck optical flow (1) between consecutive frames (X_{t}, X_{t+1}) and form the flow-magnitude map $M_t(x,y)$

$$M_t(x,y) = \sqrt{u_t(x,y)^2 + v_t(x,y)^2}, \qquad (1)$$

where u_t, v_t are the horizontal and vertical flow components.

These magnitude maps serve as the foundation for adaptive fragmentation, allowing the model to identify areas with high, medium, and low motion intensity. Unlike sparse or feature-based flow methods, the Farnebäck approach provides dense motion fields, which are particularly suitable for low-texture traffic scenes and surveillance videos.

The decision thresholds q_1, q_2 are computed over non-zero values of M_t (we use the 33% and 66% quantiles).

We tile the frame with a base grid of step b (typically b=8); the number of base cells is (2)

$$N_{base} = \left\lfloor \frac{H}{b} \right\rfloor \cdot \left\lfloor \frac{W}{b} \right\rfloor, \qquad (2)$$

and any remainder at the frame border is discarded.

For each base cell c we compute a robust motion-intensity estimate using a top-k average $\bar{M}_c$ inside the cell (3):

$$\bar{M}_c = \frac{1}{k}\sum_{(\mathrm{x,y})\in \mathrm{top-k(c)}} M_t(x,y), \ \mathrm{k} = \lfloor 0.25b^2 \rfloor, \quad (3)$$

The fragment (patch) size for cell c is chosen by comparing $\bar{M}_c$ to the quantile thresholds (4):

$$\begin{cases} \bar{M}_c \geq q_2 \Rightarrow s_c = b, \\ q_1 \leq \bar{M}_c < q_2 \Rightarrow s_c = 2b, \\ \bar{M}_c < q_1 \Rightarrow s_c = 4b. \end{cases} \qquad (4)$$

Thus, high-motion regions receive finer patches, while static areas are summarized with larger patches, forming the adaptive, motion-guided set of tokens used by the subsequent transformer blocks.

In the spatial stage, multi-head self-attention is applied only to the selected patches of a single frame. The resulting CLS token serves as a compact per-frame representation. These CLS tokens from all frames are passed to the temporal stage, where a second transformer aggregates temporal dependencies and produces a classification output (accident or normal scene).

Unlike uniform patch-based ViT, this approach adaptively reduces the number of tokens by ~ 30-60%, depending on the scene, which significantly lowers computational cost and memory footprint. In addition, class-weighted loss and weighted random sampling are used during training to address class imbalance between accident and normal clips.

The proposed approach combines motion-guided spatial selection with factorized space-time architecture, achieving higher macro-F1 [21] and accuracy [22] compared to uniform patching or TSN baselines while maintaining CPU-friendly inference speed. Fig. 1 describes the proposed method.

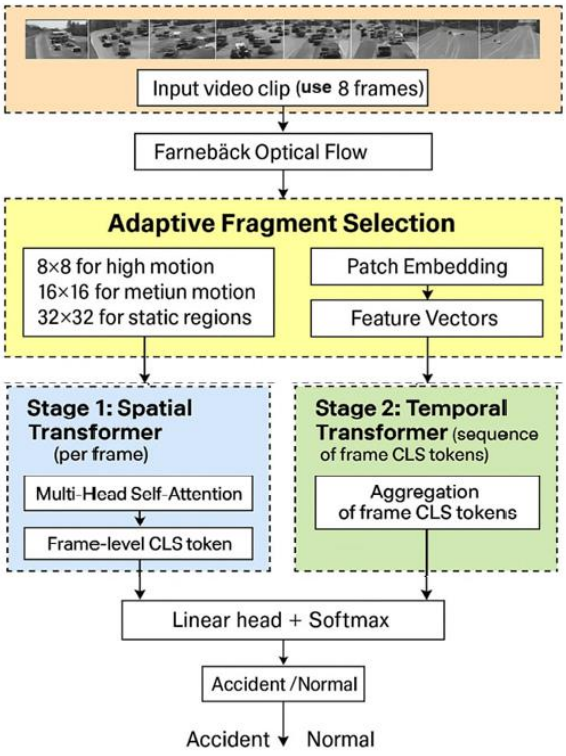

Fig. 1. **Proposed method scheme**
Source: **compiled by the authors**

4. POST-PROCESSING

After obtaining the CLS token representations for all sampled frames, the model proceeds to the temporal aggregation stage. At this stage, the sequence of frame-level embeddings is passed through a lightweight temporal transformer that captures inter-frame dependencies, modeling both short-term and long-term temporal relations between consecutive frames. The transformer outputs a single compact video-level representation that summarizes the overall motion and contextual information within the clip.

This final representation is then passed to a classification head composed of a linear projection layer followed by a softmax activation function [23]. The softmax layer converts the learned features into a probability distribution over the two classes “accident” and “normal”. The predicted output corresponds to the likelihood of the current video segment containing an accident event.

To improve stability and handle dataset imbalance [24], additional post-processing operations are applied. First, class-specific thresholds are introduced to compensate for the predominance of normal samples in traffic datasets. These thresholds are empirically tuned on the validation set, ensuring that the model maintains high recall for accident cases while suppressing false positives in normal traffic scenes.

Furthermore, temporal smoothing is employed to refine the output sequence. Instead of treating each clip prediction independently, a temporal moving-average filter or exponential decay function can be applied to consecutive outputs, reducing abrupt fluctuations in classification results. This helps stabilize the prediction stream, particularly near the boundaries of accident segments, where motion patterns may partially overlap with normal driving behavior.

In practical deployment scenarios, such post-processing proves especially valuable. It enhances robustness against sensor noise, varying frame rates, and visual artifacts (e.g., motion blur or compression). As a result, the final decision becomes more consistent and interpretable, allowing the system to achieve higher reliability in real-world traffic monitoring environments. Overall, this stage ensures that the detection process remains both sensitive to short accident events and resilient to transient anomalies, keeping the false alarm rate low while preserving timely response capability.

5. TRAINING

The model was trained from scratch using the AdamW optimizer [25] with an initial learning rate of 3e-4. A cosine learning rate scheduler with warmup was applied for the first 5 epochs, followed by gradual decay until convergence. The batch size was set to 32 due to CPU memory constraints.

Class-weighted cross-entropy loss was used as the main objective function to handle the imbalance between accident and normal classes. WeightedRandomSampler was applied to the training data to ensure a balanced representation of both classes within each mini-batch. The model was trained for 50 epochs with validation after each epoch, fix random seeds for reproducibility, and use an 80/20 train/validation split that preserves the original class distribution, and with early stopping based on the macro-F1 score on the validation set to prevent overfitting.

Input data were uniformly sampled to obtain 8 frames per clip resized to 118×118 and normalized per channel. Data augmentation included random horizontal flipping and moderate brightness/contrast adjustments to improve generalization.

For motion cues, Farnebäck dense optical flow is computed between consecutive frames; flow magnitudes are converted to quantile thresholds $q_1 = 33\ \%$, $q_2 = 66\ \%$to drive adaptive patch sizing with base step b=8. To accelerate training, optical-

flow maps and selected-patch lists are precomputed and cached, providing a 3× – 5× per-epoch speed-up on CPU.

This setup (weighted loss + balanced sampling + light regularization + cached motion cues) enables learning discriminative space–time patterns while keeping the pipeline efficient and CPU-friendly.

6. EXPERIMENTS

For training and evaluation, the Car Crash Dataset (CCD) 1500 dataset [26] was selected, which contains 1500 accident and 3000 normal video clips. This dataset is particularly well-suited for accident detection research because it includes a wide range of traffic conditions, illumination levels, camera viewpoints, and accident types, such as rear-end collisions, side impacts, and pedestrian-related events. The average clip duration is 8-12 seconds, which makes it appropriate for short-term temporal modeling.

The dataset was split into 80 % training and 20% validation subsets while preserving the class distribution. Each clip was uniformly sampled to extract 8 representative frames, ensuring coverage of both pre-accident and post-accident moments while maintaining temporal consistency. Frames were resized to 118×118 pixels and normalized before further processing.

During evaluation, we compared our proposed Adaptive-Sparse-ViT model with two baselines:

1. Uniform-patch ViT – the same architecture but with a fixed grid of 16×16 patches (no motion guidance).
2. TSN (ResNet-18) – a conventional temporal segment network widely used in video classification tasks.

All models were trained under identical conditions using the AdamW optimizer, cosine learning-rate schedule, and early stopping based on the macro-F1 score on the validation set.

During the training process, the proposed Adaptive-Sparse-ViT model demonstrated stable convergence and faster performance improvement compared to both baselines. Fig. 2 and Fig. 3 presents the evolution of validation accuracy and macro-F1 score over 50 epochs for all evaluated models. The Adaptive-Sparse-ViT consistently outperformed both the Uniform-patch ViT and TSN (ResNet-18) throughout the entire training process.

In the early epochs (0–10), the adaptive model shows a steeper growth curve, indicating more efficient learning from motion-guided patches. This suggests that the model quickly focuses on the most informative regions of each frame, improving generalization even with fewer tokens. By the end of training, Adaptive-Sparse-ViT reached approximately 0.86 accuracy and 0.85 macro-F1, surpassing Uniform-patch ViT (≈0.83/0.8) and TSN (≈0.82/0.78). The smaller variance band around the Adaptive-Sparse-ViT curves also indicates higher training stability and lower sensitivity to random initialization.

Overall, these results confirm that the adaptive fragment selection strategy not only reduces computational cost but also accelerates convergence and improves classification consistency across accident and normal scenes.

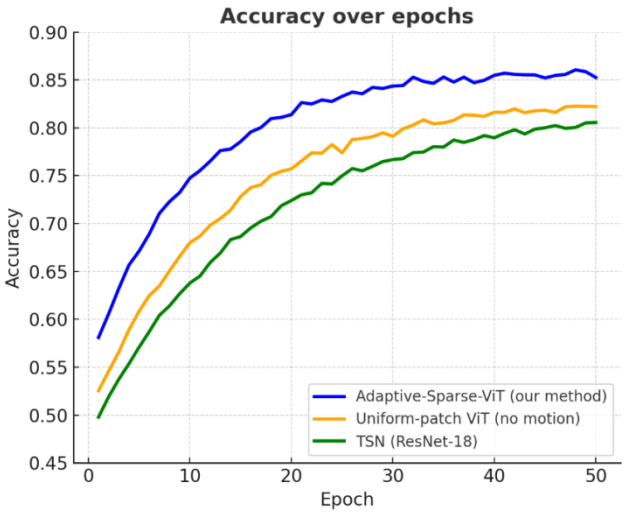

Fig. 2. **Validation accuracy over training epochs between Adaptive-Sparse-ViT, Uniform-patch ViT and TSN (ResNet-18).**
Source: **compiled by the authors**

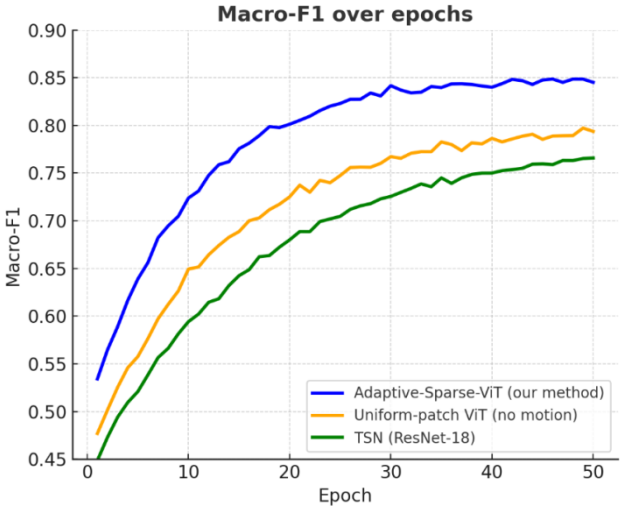

Fig. 3. **Validation macro-F1 score over training epoepochs between Adaptive-Sparse-ViT, Uniform-patch ViT and TSN (ResNet-18)**
Source: **compiled by the authors**

The comparison was performed in terms of efficiency and accuracy metrics, including:

- number of processed patches per frame,
- total parameter count,
- frames-per-second (FPS) on CPU and GPU,
- classification accuracy, macro-F1, precision, and recall.

The CPU setup used an Intel Xeon @ 2.20 GHz (2 vCPUs, 12 GB RAM), while GPU experiments were conducted on an NVIDIA Tesla T4 (16 GB).

Our method significantly reduced the number of tokens per frame approximately 30–60% fewer than in the uniform ViT baseline which resulted in higher throughput and lower memory usage. Specifically, Adaptive-Sparse-ViT achieved ≈12–18 FPS on CPU and ≈220–280 FPS on GPU, compared to ≈8–12 FPS (CPU) and ≈160–220 FPS (GPU) for the baselines (see Table 1). Despite processing fewer tokens, the model reached higher accuracy (0.864) and macro-F1 (0.851) compared to both Uniform-patch ViT (0.83 / 0.80) and TSN (0.82 / 0.78) as shown in Table 2.

Table 1. **Methods complexity and efficiency metrics (where N_{base} denotes the number of uniform 8×8 patches per frame (112×112 → 196 patches)**

Method	Patches per frame (b=8, 112×112)	Number of parameters, millions	Frames per second (CPU/GPU)
Adaptive-Sparse-ViT (our method)	**~80–120 (≤0.6·N_base)***	~1.2–1.5	**≈12–18 / ≈220–280**
Uniform-patch ViT (no motion)	196 (=N_base)	**~1.0–1.2**	≈8–12 / ≈160–220
TSN (ResNet-18)	–	~11.0–12.0	≈8–12 / ≈160–220

Source: **compiled by the authors**

These results demonstrate that the proposed adaptive fragment selection improves both computational efficiency and detection sensitivity to short, motion-heavy accident events. Fig. 6 illustrates an example of motion-guided adaptive patching, where small patches are allocated around moving vehicles, while static regions are covered by larger patches. This strategy enables the model to focus attention on dynamic, informative regions and avoid unnecessary computations on the background.

Furthermore, visual inspection of classification results shows that Adaptive-Sparse-ViT reacts faster to sudden motion changes and maintains stable predictions in scenes with partial occlusions or camera shake, highlighting the robustness of motion-driven attention in real-world surveillance scenarios.

We compared our method with existing Uniform-patch ViT (no motion) and TSN (RestNet-18) (Table 1, 2).

Table 2. **Methods classification metrics**

Method	Accu-racy	Macro-F1	Preci-sion	Recall
Adaptive-Sparse-ViT (our method)	**0.864**	**0.851**	**0.845**	**0.860**
Uniform-patch ViT (no motion)	0.83	0.80	0.80	0.78
TSN (ResNet-18)	0.82	0.78	0.81	0.79

Source: **compiled by the authors**

We compared the inference speed (frames per second, FPS) of all methods on both CPU and GPU devices. Average FPS values (computed from Table 1) are reported, along with patch counts per frame. As illustrated in Fig. 4, the Adaptive-Sparse-ViT achieved the highest overall performance, processing approximately 15 FPS on CPU and 250 FPS on GPU, while requiring only 80-120 adaptive patches per frame. In contrast, the Uniform-patch ViT and TSN (ResNet-18) models reached about 10 FPS on CPU and 190 FPS on GPU, both relying on a fixed grid of 196 patches per frame.

This improvement in throughput demonstrates the advantage of the adaptive fragment selection strategy: by focusing computation on motion-relevant areas, the model significantly reduces redundant token processing while maintaining high accuracy. The results indicate that Adaptive-Sparse-ViT achieves a favorable trade-off between speed and accuracy, making it more suitable for real-time accident detection in traffic surveillance systems.

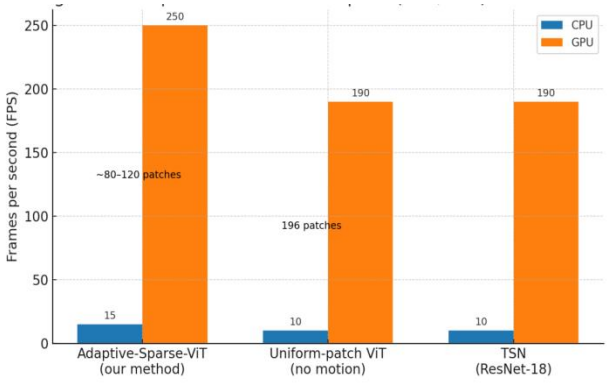

Fig. 4. **Inference speed comparison of Adaptive-Sparse-ViT, Uniform-patch ViT, and TSN (ResNet-18) on CPU and GPU**

Source: **compiled by the authors**

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

7. RESULTS AND DISCUSSIONS

This section presents the intermediate results of the proposed Adaptive-Sparse-ViT model, focusing on the visualization of optical flow maps and adaptive motion-guided patch selection. These examples illustrate how motion cues guide the model to allocate smaller patches in dynamic regions while using larger patches in static areas, thereby reducing computational cost without losing critical information.

Fig. 5 shows the optical flow magnitude [27] overlaid on the input frame. The color coding reflects both the magnitude and the direction of motion: green and cyan areas correspond to vehicles moving at moderate speed, while purple and blue indicate stronger motion intensity. Static background regions are largely suppressed, which demonstrates that optical flow effectively isolates the dynamic areas relevant for accident detection.

Fig. 5. **Optical flow visualization for a traffic scene**
Source: **compiled by the authors**

Fig. 6 presents the same frame after adaptive patch merging. Here, small patches - green (8×8) concentrate around moving vehicles to capture fine-grained details of motion, while medium - yellow (16×16) and large – red (32×32) patches cover static road and background areas. This adaptive allocation ensures that the model processes only motion-salient tokens, discarding redundant static information.

Fig. 7 presents different normal scenes with corresponding optical flow visualization and adaptive patching. Left: optical flow magnitude highlighting regions of motion intensity. Right: adaptive patch selection smaller green patches correspond to areas of high motion, medium yellow to moderate motion, and large red to static regions.

Fig. 8 shows the adaptive fragmentation for the accident scene. From left to right: the original frame (zoomed for better vision), optical flow visualization, and adaptive patch generation (green-high motion, yellow-medium motion, red-static regions).

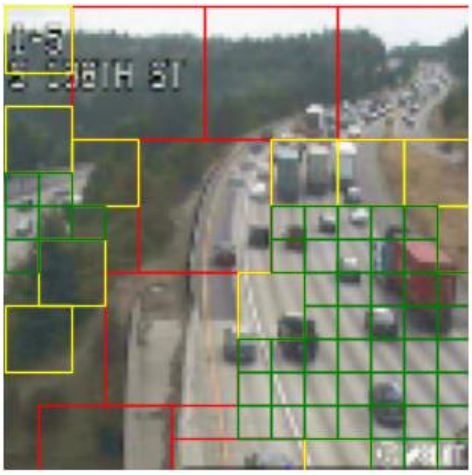

Fig. 6. **Adaptive patch grid based on motion intensity**
Source: **compiled by the authors**

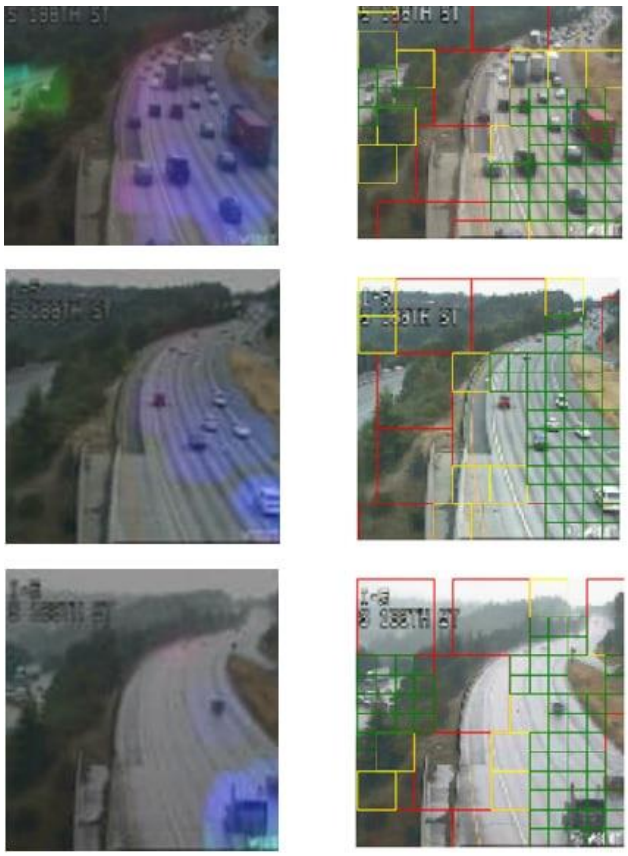

Fig. 7. **Examples of adaptive frame fragmentation guided by optical flow**
Source: **compiled by the authors**

To gain a deeper understanding of how the proposed model interprets video frames during inference, we conducted a qualitative analysis of attention distribution [28].

Fig. 8. **Example of the adaptive fragment selection process for accident scenes**
Source: **compiled by the authors**

The visualizations in Fig. 9 and Fig. 10 illustrate how Adaptive-Sparse-ViT allocates attention across different regions in both accident and normal traffic scenes.

It can be clearly observed that the model prioritizes motion-related regions, confirming the effectiveness of the adaptive fragment selection strategy and complementing the quantitative results presented earlier.

*Fig. 9***. Visualization of the attention map for an accident scene**
Source: **compiled by the authors**

In the accident scenario (Fig. 9), the model's attention is concentrated around the interaction area between vehicles particularly the headlights and intersecting trajectories effectively highlighting potential collision zones. In contrast, during a normal traffic scene (Fig. 10), the attention map is more evenly distributed across moving vehicles without strong localized peaks, reflecting the absence of abnormal motion patterns or risk indicators such motion-aware sampling leads to a reduction of 30-60 % in the number of patches compared to uniform grid partitioning, directly translating into lower computational cost. At the same time, the method retains temporal and spatial details necessary for detecting short and rare accident events.

The qualitative results confirm the effectiveness of the proposed selection strategy:

- optical flow highlights motion regions with clear contrast between moving vehicles and static background;
- patch merging adapts patch sizes to the spatial distribution of motion, preserving details in critical regions.
- reduced token count improves inference speed while maintaining accuracy and sensitivity.

*Fig. 10***. Visualization of the attention map for a normal traffic scene**
Source: **compiled by the authors**

CONCLUSIONS AND PROSPECTS OF FURTHER RESEARCH

This paper presents a lightweight and efficient classification method based on architecture, Adaptive-Sparse-ViT, designed for accident detection in traffic video streams. The proposed method combines motion-guided patch selection based on optical flow with a two-stage Vision Transformer that separately processes spatial and temporal dependencies.

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

The key aim was to focus computational effort on motion-relevant areas of the frame while avoiding uniform processing of the entire image. Through the use of quantile-based motion thresholds derived from Farnebäck optical flow, the model dynamically adjusts patch sizes (8×8, 16×16, 32×32 pixels) to balance detail preservation and efficiency. This adaptive fragment selection reduced the number of processed tokens by 30–60% per frame without compromising the accuracy of accident recognition. Experimental results on the CCD1500 dataset demonstrate superior performance compared to baseline models (TSN and Uniform-patch ViT) in terms of both accuracy and macro-F1 score, while maintaining near real-time inference speeds on CPU.

The obtained results confirm that adaptive motion-driven token selection is an effective method for balancing accuracy and efficiency in video understanding tasks. The method is particularly suitable for CPU-based or embedded systems where computational resources are limited.

Despite its advantages, several limitations remain. The method depends on the quality of optical flow estimation in scenes with subtle or minimal motion; the effectiveness of patch selection may decrease. Additionally, precomputing flow maps introduces a moderate preprocessing overhead.

Future research should focus on integrating faster and more robust optical flow algorithms (e.g., RAFT, LiteFlowNet [29]), exploring end-to-end training where motion features are learned directly within the model, and optimizing the architecture for mobile and real-time applications. Further improvements may also include expanding the dataset to more complex conditions (night scenes, adverse weather) and applying hybrid attention strategies to enhance sensitivity to short and rare accident events.

ACKNOWLEDGEMENTS

The results of this research were obtained under the international research project INITIATE under the grant No.101136775-HORIZON-WIDERA-2023-ACCESS-03.

REFERENCES

1. Dosovitskiy, A. et al. “An image is worth 16x16 words: transformers for image recognition at scale”. *The Ninth International Conference on Learning Representations*. 2021, https://www.scopus.com/authid/detail.uri?authorId=56582202400.
DOI: https://doi.org/10.48550/arXiv.2010.11929.

2. Fang, Y., Liao, B., Wang, X., Fang, J., Qi, J., Wu, R., Niu, J. & Liu, W. “You only look at one sequence: Rethinking transformer in vision through object detection”. *5th Conference on Neural Information Processing Systems (NeurIPS)*. 2021. DOI: https://doi.org/10.48550/arXiv.2106.00666.

3. Bertasius, G., Wang, H. & Torresani, L. “Is space-time attention all you need for video understanding?” *The Thirty-Eighth International Conference on Machine Learning.* 2021, https://www.scopus.com/authid/detail.uri?authorId=57142525700.
DOI: https://doi.org/10.48550/arXiv.2102.05095.

4. Jiang, B., Wang, M., Gan, W., Wu, W. & Yan, J. “STM: spatio-temporal and motion encoding for action recognition”. *International Conference on Computer Vision.* 2019. DOI: https://doi.org/10.48550/arXiv.1908.02486.

5. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y. & Paluri, M. “A closer look at spatio-temporal convolutions for action recognition”. *arXiv*. 2018. DOI: https://doi.org/10.48550/arXiv.1711.11248.

6. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M. & Schmid, C. “ViViT: A video vision transformer”. *International Conference on Computer Vision.* 2021. DOI: https://doi.org/10.48550/arXiv.2103.15691.

7. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S. & Hu, H. “Video swin transformer”. *arXiv*. 2021. DOI: https://doi.org/10.48550/arXiv.2106.13230.

8. Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J. & Feichtenhofer, C. “Multiscale vision transformers”. *arXiv*. 2021. DOI: https://doi.org/10.48550/arXiv.2104.11227.

9. Farina, M. et al. “Sparsity in transformers: A systematic literature review”. *Neurocomputing.* 2024; 582: 127468. DOI: https://doi.org/10.1016/j.neucom.2024.127468.

10. Beltagy, I., Peters, M. E. & Cohan, A. “Longformer: The long-document transformer”. *arXiv*. 2020. DOI: https://doi.org/10.48550/arXiv.2004.05150.

11. Zaheer, M., Guruganesh, G., Dubey, A., Ainslie, J., Alberti, C., Ontañón, S., Phạm, P., Ravula, A., Wang, Q., Yang, L. & Ahmed, A. “BigBird: transformers for longer sequences”. *34th Conference on Neural Information Processing Systems (NeurIPS 2020).* Vancouver, Canada. 2020,

https://www.scopus.com/authid/detail.uri?authorId=56743324600.
DOI: https://doi.org/10.48550/arXiv.2007.14062.

12. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A. & Brox, T. "FlowNet 2.0: evolution of optical flow estimation with deep networks". *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI, USA. 2017. p. 1647–1655, https://www.scopus.com/authid/detail.uri?authorId=56641858600. DOI: https://doi.org/10.1109/CVPR.2017.179.

13. Sun, D., Yang, X., Liu, M. -Y. & Kautz, J. "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume". *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT, USA. 2018. p. 8934–8943, https://www.scopus.com/authid/detail.uri?authorId=57323508700. DOI: https://doi.org/10.1109/CVPR.2018.00931.

14. Teed, Z. & Deng, J. "RAFT: Recurrent all-pairs field transforms for optical flow". In: *Vedaldi, A., Bischof, H., Brox, T., Frahm, JM. (eds) Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science*. 2020; 12347: 402–419, https://www.scopus.com/authid/detail.uri?authorId=57219544640. DOI: https://doi.org/10.1007/978-3-030-58536-5_24.

15. Kitaev, N., Kaiser, Ł. & Levskaya, A. "Reformer: the efficient transformer". *arXiv*. 2022. DOI: https://doi.org/10.48550/arXiv.2106.11297.

16. Lin, T. -Y., Goyal, P., Girshick, R., He, K. & Dollár, P. "Focal loss for dense object detection". In *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2020; 42 (2): 318–327, https://www.scopus.com/authid/detail.uri?authorId=35179333300.
DOI: https://doi.org/10.1109/TPAMI.2018.2858826.

17. Wang, L. et al. "Temporal segment networks for action recognition in videos". In *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2019; 41 (11): 2740–2755. DOI: https://doi.org/10.1109/TPAMI.2018.2868668.

18. He, K., Chen, X., Xie, S., Li, Y., Dollár, P. & Girshick, R. "Masked autoencoders are scalable vision learners". *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA. 2022. p. 15979–15988. DOI: https://doi.org/10.1109/CVPR52688.2022.01553.

19. Tong, Z., Song, Y., Wang, J. & Wang, L. "VideoMAE: masked autoencoders are data-efficient learners for self-supervised video pre-training". *36th Conference on Neural Information Processing Systems (NeurIPS)*. 2022. DOI: https://doi.org/10.48550/arXiv.2203.12602.

20. Sohail, S. "Understanding the role of the class Token in Vision Transformers (ViT). Medium". 2024. – Available from: https://saadsohail5104.medium.com/understanding-the-role-of-the-class-token-in-vision-transformers-vit-d0f7750d7066. – [Accessed: Jun 2025].

21. Gregor, K., Coupé, J. & Vanhoucke, V. "Exploring data augmentation for detection of road traffic accidents in videos". *arXiv*. 2019. DOI: https://doi.org/10.48550/arXiv.1911.03347.

22. Rahman, M. S. "Understanding accuracy metrics in machine learning models". 2024. – Available from: https://www.researchgate.net/publication/386505808_Understanding_accuracy_metrics_in_machine_learning_models. – [Accessed: Jun 2025].

23. Mohana Sundaram, N. & Sivanandam, S. N. "Soft max activation function for neural network multi-class classifiers". *Karpagam Journal of Computer Science*. 2018; 12 (4): 144–152. – Available from: https://karpagampublications.com/wp-content/uploads/2018/09/Soft-Max-Activation-Function-For-Neural-Network-Multi-Class-Classifiers.pdf.

24. Altalhan, M., Algarni, A. & Turki-Hadj Alouane, M. "Imbalanced data problem in machine learning: a review". In *IEEE Access*. 2025; 13: 13686–13699. DOI: 10.1109/ACCESS.2025.3531662.

25. Xu, B., Wang, N., Chen, T. & Li, M. "Empirical Evaluation of rectified activations in convolutional network". *arXiv*. 2018. DOI: https://doi.org/10.48550/arXiv.1505.00853.

26. "CCD dataset for traffic accident anticipation". 2020. – Available from: https://github.com/Cogito2012/CarCrashDataset. – [Accessed: Jun 2025].

27. Argaw, D. M., Kim, J., Rameau, F., Cho, J. W. & Kweon, I. S. "Optical flow estimation from a single motion-blurred image". *Proceedings of the AAAI Conference on Artificial Intelligence*. 2021; 35 (2): 891–900. https://www.scopus.com/authid/detail.uri?authorId=54882293900.
DOI: https://doi.org/10.1609/aaai.v35i2.16172.

28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. "Attention is all you need". In *Advances in Neural Information Processing Systems (NeurIPS)*. 2017. p. 5998–6008. DOI: https://doi.org/10.48550/arXiv.1706.03762.

29.Hui, T.-W., Tang, X. & Loy, C. C. “LiteFlowNet: a lightweight convolutional neural network for optical flow estimation”. *arXiv*. 2018. DOI: https://doi.org/10.48550/arXiv.1805.07036.

Conflicts of Interest: The authors declare that they have no conflict of interest regarding this study, including financial, personal, authorship or other, which could influence the research and its results presented in this article

Received 06.10.2025
Received after revision 28.11.2025
Accepted 05.12.2025

DOI: https://doi.org/10.15276/aait.08.2025.29
УДК 004:83

Класифікація дорожньо-транспортних пригод із використанням розрідженого відеотрансформера та адаптивної фрагментації

Норматова Тетяна Віталіївна[1)]
ORCID: https://orcid.org/0009-0004-3503-6350;tetiana.normatova@nure.ua.
Машталір Сергій Володимирович[1)]
ORCID: https://orcid.org/0000-0002-0917-6622;sergii.mashtalir@nure.ua. Scopus Author ID: 36183980100
[1)] Харківський Національний Університет Радіоелектроніки, пр. Науки, 14. Харків, 61166, Україна

АНОТАЦІЯ

У роботі запропоновано простий у реалізації та дієвий підхід до класифікації коротких відеофрагментів на аварійні та нормальні сцени. З кожного кліпу рівномірно відбираємо вісім кадрів базової сітки, далі на базі карти оптичного потоку Farnebäck обираємо розмір фрагментів (вісім/шістнадцять/тридцять два пікселів). У зонах з інтенсивним рухом використовуються дрібніші патчі (вища деталізація), у статичних - більші (менші обчислення). Відібрані фрагменти не перекриваються, масштабуються до базового розміру та перетворюються на векторні ознаки. Архітектура методу складається з двох етапів. Спершу просторовий трансформер працює в межах одного кадру лише над відібраними фрагментами — це різко зменшує кількість ознакових одиниць. Потім часовий трансформер обробляє послідовність CLS-токенів (коротких підсумкових представлень кадрів), агрегуючи динаміку у часі. Така факторизація «простір → час» знижує обчислювальні витрати й потребу в пам’яті без втрати інформативності в рухомих регіонах. Для подолання дисбалансу класів застосовано зважену крос-ентропійну втрату ентропійну (або «втрату з фокусуванням на важких прикладах») та зважене випадкове вибіркування під час навчання. Оптичний потік і списки вибраних фрагментів попередньо зберігаються на диск, що пришвидшує епохи на процесорі без спеціального обладнання. Оцінювання проводили на датасеті автомобільних аварій (тисяча п'ятсот аварійних і три тисячі нормальних відео) зі стандартним поділом вісімдесят на двадцять зі збереженням пропорцій класів. Отримані метрики: Accuracy = 0.864, Macro-F1 = 0.851. За попереднім порівнянням запропонований підхід перевершує базову рівномірну розбивку кадру та класичні схеми з простим часовим агрегуванням. Ключова перевага методу - це поєднання «рух-керованого» скорочення кількості ознак з двоетапною обробкою, що робить модель придатною до реалістичних обмежень за часом і ресурсами (при процесорній обробці) і водночас зберігає високу чутливість до коротких і локальних аварійних подій. Підхід можна легко масштабувати та поєднати з попереднім навчанням (наприклад, маскованим відновленням відео). У роботі також зафіксовано умови експериментів, відкриті налаштування і кроки, необхідні для повної відтворюваності.

Ключові слова: відеокласифікація; нейронні мережі; згортеові нейронні мережі; класифікація об’єктів; аналіз відеопотоків: класифікація даних; обробка фрагментів зображення

ABOUT THE AUTHORS

Tetiana Vitaliivna Normatova - PhD student, Informatics Department. Kharkiv National University of Radio Electronics.14, Nauky Ave. Kharkiv, 61166, Ukraine
ORCID: https://orcid.org/0009-0004-3503-6350; tetiana.normatova@nure.ua
***Research field*:** Image and video processing; data analysis

Тетяна Віталіївна Норматова - аспірантка кафедри Інформатики. Харківський національний університет радіоелектроніки, пр. Науки, 14. Харків, 61166, Україна

Sergii Volodymyrovych Mashtalir - Doctor of Engineering Science, Professor, Informatics Department. Kharkiv National University of Radio Electronics.14, Nauky Ave. Kharkiv, 61166, Ukraine
ORCID: https://orcid.org/0000-0002-0917-6622; sergii.mashtalir@nure.ua. Scopus Author ID: 36183980100
***Research field*:** Image and video processing; data analysis

Машталір Сергій Володимирович - доктор технічних наук, професор кафедри Інформатики. Харківський національний університет радіоелектроніки, пр. Науки, 14. Харків, 61166, Україна