DOI: https://doi.org/10.15276/aait.09.2026.37
UDC 004:83

Motion-aware adaptive fragmentation and transformer-based temporal modeling for lane-change recognition in Unmanned Aerial Vehicle traffic videos

Tetiana V. Normatova[1)]
ORCID: https://orcid.org/0009-0004-3503-6350; tetiana.normatova@nure.ua Scopus Author ID: 60278045600
Sergii V. Mashtalir[1)]
ORCID: httpi://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

Lane-change recognition in traffic videos captured by unmanned aerial vehicles is challenging due to small vehicle size, global camera motion, and the computational cost of dense video processing. **We propose** a motion-aware adaptive fragmentation framework that combines homography-based camera-motion compensation, optical-flow-guided motion-saliency estimation, and sparse spatiotemporal Transformer modeling. Motion-salient regions are represented using fine-grained patches, whereas regions with lower motion activity are processed using larger patches to reduce redundant spatial representation**. The framework was evaluated** on a lane-change dataset derived from tracking annotations of the Unmanned Aerial Vehicle Benchmark using an automatic trajectory-based label-generation procedure. The resulting dataset contains four hundred and one thousand three hundred forty-five eight-frame temporal windows extracted from thirty aerial video sequences and one thousand three hundred sixteen vehicle tracks, including fifty-four thousand seven hundred and seventy-five lane-change and three hundred forty-six thousand, five hundred seventy normal-driving samples**. Experimental results demonstrate** an accuracy of zero point seven nine, a binary F1-score of zero point seven seven, an area under the receiver operating characteristic curve of zero point nine three, and an area under the precision-recall curve of zero point nine one. The proposed adaptive representation limits the number of spatial tokens to a maximum of sixty-four per vehicle-centered frame, corresponding to a seventy-five percent reduction in the maximum spatial token budget relative to the two hundred fifty-six-token dense uniform-patch tubelet baseline. **The results indicate** that motion-aware adaptive fragmentation can preserve competitive lane-change recognition performance while substantially reducing the spatial token sequence processed by the Transformer. This provides a promising basis for computationally efficient aerial traffic monitoring, while complete assessment of edge-deployment efficiency requires future end-to-end measurements that also account for homography estimation and optical-flow computation.

Keywords: Video classification; lane change detection; optical flow; adaptive fragmentation; transformer; spatio-temporal modeling; Unmanned Aerial Vehicles video analysis

For citation: **Normatova T. V., Mashtalir S. V. "Motion-aware adaptive fragmentation and transformer-based temporal modeling for lane-change recognition in UAV traffic videos". *Applied Aspects of Information Technology.* 2026; Vol.9 No.4: 574–584. DOI: https://doi.org/10.15276/aait.09.2026.37**

INTRODUCTION

Lane-change recognition is important for traffic safety analysis, behavior understanding, and autonomous-driving support. UAVs provide wide-area traffic observations, but aerial video introduces small targets, camera motion, and substantial spatiotemporal redundancy that complicate efficient maneuver recognition.

Reliable recognition of lane-changing maneuvers can support the identification of hazardous situations and improve traffic-dynamics analysis.
UAV-based monitoring offers flexible, wide-field traffic observation; yet global camera motion, viewpoint variation, and small vehicle size make motion estimation difficult. Dense video models also process large amounts of static background information, increasing computational cost.

To address these limitations, we propose a motion-aware adaptive fragmentation framework that combines vehicle-centered tubelets, homography-based UAV motion compensation, weighted optical-flow saliency, hierarchical patching, and transformer-based temporal modeling.

The framework allocates fine patches to motion-salient regions and coarse patches elsewhere, reducing the spatial token sequence while preserving maneuver-related information.

The main contributions of this work are as follows.

1. A motion-aware adaptive fragmentation strategy for efficient lane-change recognition in UAV traffic videos.

2. Integration of homography-based UAV motion compensation with weighted optical-flow-based motion saliency estimation.

© Normatova T., Mashtalir S., 2026

This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/deed.uk)

3. A sparse spatiotemporal transformer framework that reduces the number of processed spatial tokens while preserving recognition performance.

4. Experimental validation on a lane-change dataset derived from UAVDT, demonstrating improved recognition accuracy and reduced spatial token processing compared with a dense uniform-patch tubelet baseline.

1. RELATED WORKS

Lane-change recognition has traditionally relied on vehicle trajectories, lane geometry, and handcrafted motion descriptors. Such approaches can be effective when reliable tracking and lane information are available, but their dependence on structured traffic variables limits direct use in visually complex UAV scenes [1], [2].

Deep video models, including CNN-, 3D-CNN-, and transformer-based architectures, learn spatiotemporal representations directly from visual data [3]-[9]. However, dense processing introduces redundant computation, especially when target vehicles occupy only a small part of an aerial frame.

Optical flow provides explicit pixel-level motion cues and is widely used in action and vehicle-behavior analysis [11], [12], [13], [14]. In UAV video, global camera motion can dominate local vehicle motion, motivating motion compensation before residual motion analysis.

Efficient transformers reduce redundant processing through sparse attention, adaptive token selection, and dynamic spatial allocation [15], [16], [17], [18]. These ideas motivate the proposed motion-guided fragmentation strategy.

Existing work rarely combines UAV motion compensation, adaptive motion-aware spatial representation, and sparse transformer temporal modeling for lane-change recognition. This work addresses that gap in a unified framework.

2. PROBLEM STATEMENT

Lane-change recognition in UAV traffic video requires distinguishing subtle vehicle motion from global camera motion while avoiding redundant processing of static image regions.

Dense full-frame video models can capture maneuver dynamics but are inefficient for aerial scenes in which vehicles occupy only a small fraction of the frame. UAV translation, rotation, and vibration further distort optical-flow estimates.

The objective of this work is to develop an efficient framework for lane-change recognition in UAV traffic videos that reduces spatial-temporal redundancy while preserving motion-sensitive vehicle dynamics.

Specifically, the proposed approach aims to:

1) compensate UAV-induced camera motion to improve motion estimation accuracy;

2) use optical-flow analysis to detect motion-salient regions;

3) adaptive fragmentation for dynamic spatial resolution distribution;

4) sparse spatio-temporal transformer for vehicle behavior modeling;

5) to achieve a good trade-off between recognition accuracy and computational efficiency.

The resulting framework is designed to focus computational resources on motion-relevant traffic regions while minimizing redundant processing of static background areas, making it suitable for deployment in resource-constrained UAV-based traffic monitoring systems.

3. PROPOSED METHOD

This work proposes a motion-aware sparse spatiotemporal framework for lane-change recognition in UAV traffic videos. The method combines global motion compensation, vehicle-centered tubelets, motion-guided adaptive fragmentation, and factorized spatial-temporal transformer modeling.

Overview. Because UAVDT does not provide lane-change annotations, labels are generated automatically from vehicle trajectories. The recognition pipeline then performs UAV motion compensation, vehicle-centered cropping, residual optical-flow analysis, adaptive fragmentation, and transformer-based window classification.

Given an input UAV video sequence, global UAV-induced camera motion is first compensated between consecutive full-resolution frames using homography estimation. The frame-specific vehicle tracking annotations are then used to extract 128×128 object-centric regions around individual vehicles. Eight consecutive regions form a tubelet, and residual optical flow is computed between consecutive compensated regions using the Farnebäck algorithm [19].

Motion-Aware Adaptive Fragmentation

The residual optical-flow field obtained after UAV motion compensation is representedii through horizontal and vertical motion components. Since lane-change maneuvers often involve significant lateral displacement, the motion saliency M at pixel

coordinates (x,y) is computed according to formula (1):

$$M(x,y) = \sqrt{\{^{(w_u \cdot u)^2 + (w_v \cdot v)^2}}, \qquad (1)$$

where u and v denote the horizontal and vertical optical flow components, respectively, and $w_u > w_v$ are the weighting coefficients that prioritize lateral motion. In all experiments, the weighting coefficients were fixed to $w_u = 1.5$ and $w_v = 1$, emphasizing lateral motion associated with lane-changing behavior. The weighting wu =1.5 and wv =1.0 was empirically selected to increase sensitivity to horizontal motion patterns observed in the analyzed UAVDT sequences. No explicit road-orientation normalization was applied during preprocessing. Consequently, the current weighting is not rotation-invariant and its effectiveness may depend on the orientation of the road relative to the image coordinate system. This weighted motion representation increases sensitivity to motion patterns potentially associated with lane-changing behavior while preserving information about overall vehicle dynamics.

The resulting motion map serves as a motion-saliency representation that guides adaptive fragmentation. Each frame is partitioned using a hierarchical patching strategy.

Depending on local motion activity, one of three fragment resolutions is assigned:

- 8×8 patches for regions with high motion activity;
- 16×16 patches for regions with moderate motion activity;
- 32×32 patches for regions with weak motion activity.

Fragments are extracted without overlap: smaller patches preserve fine motion details, while larger patches reduce redundancy in low-motion regions.

All fragments are resized to a common resolution and projected into fixed-dimensional embeddings.

Spatiotemporal Representation Learning. Adaptive fragment embeddings are augmented with positional information and processed by a spatial transformer encoder.

A learnable classification token [20] aggregates the spatial patch information into a compact frame-level representation.

Frame-level representations are passed to a temporal transformer that models motion evolution across the eight-frame tubelet. This factorized design separates within-frame spatial modeling from temporal dynamics and avoids dense spatiotemporal attention over all patches and frames simultaneously.

During inference, overlapping eight-frame windows are processed independently; their predictions are not aggregated in the reported window-level evaluation.

Window-Level Lane-Change Classification. A lightweight classification head converts the temporal representation into a lane-change probability for each eight-frame window.

The framework therefore performs visual maneuver recognition without requiring lane geometry or handcrafted behavioral rules during inference.

Overall, the method combines object-centric representation, camera-motion compensation, adaptive spatial processing, and temporal attention to reduce redundant token processing while preserving lane-change cues.

Inference and Decision Strategy. Each tubelet is processed independently by the spatial and temporal transformer modules, and the classification head outputs the lane-change probability p.

The final binary prediction is obtained using the threshold-based decision rule in Eq. (2).

$$\hat{y} = \begin{cases} 1, p \geq \tau \\ 0, p < \tau \end{cases} \qquad (2)$$

where p is the predicted probability of the lane-change class; $\tau = 0.6$ is the decision threshold selected on the validation set; $\hat{y}$ is the final predicted label; $\hat{y} = 0$ corresponds to normal driving.

The decision threshold is selected on the validation set; no explicit trajectory or lane-geometry analysis is required at inference time.

4. TRAINING

The framework is trained on UAVDT-derived traffic sequences using track-separated temporal windows, weighted loss for class imbalance, and cached motion preprocessing.

Dataset Preparation and Annotation. UAVDT provides diverse aerial traffic scenes and vehicle tracking annotations. These annotations were adapted to lane-change recognition as follows.

Vehicle-Centered Tubelet Generation. Frame-specific UAVDT bounding boxes locate each tracked vehicle.

After global motion compensation, a 128×128 vehicle-centered crop is extracted from each frame using the corresponding bounding box. Eight consecutive crops form the spatiotemporal tubelet,

with updated bounding boxes allowing the crop to follow the vehicle.

Automatic Trajectory-Based Labeling. Vehicle trajectories were derived from the UAVDT tracking annotations using normalized bounding-box center coordinates. Candidate lane-change events were identified automatically using nine-frame trajectory segments. For each segment, lateral motion was quantified as the standard deviation of the normalized horizontal vehicle position, while directional variation was measured as the mean absolute change in trajectory angle. A segment was identified as a lane-change candidate when the lateral-motion measure exceeded 0.004 and the mean angular change exceeded 0.05 rad; otherwise, it was classified as normal driving. The resulting trajectory-level labels were then mapped to temporal windows of L=8 consecutive frames used as model inputs. An eight-frame window was assigned the positive Lane Change label if it overlapped with at least one trajectory segment identified as a lane-change candidate; otherwise, it was labeled as Normal Driving.

Temporal Window Generation. Fixed-length temporal windows of L=8 consecutive frames were extracted using a sliding-window procedure with a stride of one frame. At 30 fps, the temporal interval between the first and last frames of each window is approximately 0.233 s, while consecutive window starting positions are separated by approximately 0.033 s. The resulting windows overlap substantially and were treated as independent samples during the reported window-level evaluation; predictions from overlapping windows were not aggregated.

Optical-flow maps and adaptive fragmentation layouts are precomputed and cached on disk to avoid repeated preprocessing during training.

Training Strategy and Implementation. The framework is implemented in PyTorch and optimized with AdamW using the configuration summarized in Table 1.

Class Imbalance Handling. Lane-change events occur considerably less frequently than normal-driving events. To address class imbalance [21], weighted Binary Cross-Entropy loss is used during training (3), assigning a higher penalty to misclassification of the minority Lane Change class:

$$L_W = -\left[w_p y log(p) + (1-y) log(1-p)\right], \quad (3)$$

where y is the ground-truth label; p is the predicted probability; w_p is the positive-class weight computed from the class distribution.

In the experiments, the positive-class weight was set to wp≈6.33, reflecting the imbalance between the Normal Driving and Lane Change classes.

The positive-class weighting reduces bias toward the majority Normal Driving class.

Architecture Details. Adaptive fragments are linearly projected to embeddings with positional encodings, processed by the spatial transformer, and then passed to the temporal transformer.

Evaluation Protocol. Training, validation, and test data are separated at the vehicle-track level to prevent temporal leakage.

Performance is evaluated using Accuracy, Precision, Recall, Binary F1-score, ROC-AUC, and PR-AUC.

Because the classes are imbalanced, ROC-AUC and PR-AUC are reported alongside threshold-dependent metrics, with PR-AUC providing particular insight into minority-class performance [22].

Computational efficiency is evaluated by the average number of adaptive patches and spatial tokens processed by the transformer.

Model and Training Configuration. Table 1 summarizes the main architectural, fragmentation, and training parameters for reproducibility.

Table 1. **Model and Training Configuration**

Parameter	Value
Input region size	128 × 128 pixels
Temporal window length	L = 8 frames
Sliding-window stride	1 frame
Video frame rate	30 fps
High-motion patch size	8 × 8
Medium-motion patch size	16 × 16
Low-motion patch size	32 × 32
High-motion threshold	P85 (85th percentile)
Medium-motion threshold	P60 (60th percentile)
Patch embedding dimension	128
Transformer encoder layers	4
Attention heads	4
Batch size	32
Optimizer	AdamW
Learning rate	2×10^{-4}
Number of epochs	50
Classification threshold	$\tau = 0.6$
Positive-class BCE weight	wp ≈ 6.33
Hardware platform	NVIDIA T4 GPU

Source: **compiled by the authors**

Motion thresholds are defined by the saliency distribution: P85 for high-motion 8×8 patches, P60-P85 for medium-motion 16×16 patches, and values below P60 for low-motion 32×32 patches.

5. EXPERIMENTS

The experiments evaluate classification quality and spatial-token efficiency of the proposed framework under UAVDT-derived aerial traffic conditions.

Experimental Setup and Evaluation Metrics. The dataset contains 401,345 eight-frame windows from 30 UAV sequences and 1,316 vehicle tracks: 54,775 Lane Change (13.6 %) and 346,570 Normal Driving (86.4 %). A 70%/15%/15% train/validation/test split is applied at vehicle-track level so that all windows from a track remain in one subset, preventing temporal leakage.

Windows use L=8 frames with stride 1. Weighted binary cross-entropy addresses class imbalance, and the validation-selected decision threshold τ=0.6 is fixed for all reported experiments.

Comparison with Baseline Approaches. Table 2, Table 3 and Table 4 position the method relative to dense action-recognition, transformer, sparse/adaptive, and trajectory-based approaches. Direct full-frame quantitative comparison on UAVDT is limited by small vehicle size and the memory cost of dense high-resolution processing; therefore, controlled ablation against the dense uniform-patch tubelet baseline provides the primary quantitative comparison. UAV-derived trajectories have also been used for lane-change identification [25].

Ablation Study. Table 4 shows consistent gains: Binary F1-score increases from 0.66 to 0.77, ROC-AUC from 0.81 to 0.93, and PR-AUC from 0.72 to 0.91, while spatial tokens decrease from 256 to 64 per vehicle-centered frame (75%).

Dense Uniform-Patch Baseline. Uniform 8×8 patching of each 128×128 vehicle-centered frame produces 256 spatial tokens and yields Binary F1=0.66 and ROC-AUC=0.81.

- Optical Flow Guidance [26]. Motion guidance raises Binary F1 to 0.69 while reducing the representation to 192 tokens.
- Adaptive Fragmentation. Three-level 8×8/16×16/32×32 patching increases Binary F1 to 0.72 and reduces the representation to 96 tokens.
- UAV Motion Compensation. Homography-based alignment [27] suppresses camera-induced displacement, improving Binary F1 to 0.74 and ROC-AUC to 0.91 without directly changing the 96-token count.
- Full Proposed Framework. With horizontal-motion weighting, the full model reaches Accuracy 0.79, Binary F1=0.77, and ROC-AUC=0.93 while processing 64 tokens per vehicle-centered frame.

Table 2. **Comparison of lane-change recognition approaches**

Method	Input Type	UAV Support	Motion-Aware
CNN+LSTM	RGB video	–	partially
I3D / SlowFast	RGB video	–	–
ViViT	RGB video	–	–
TokenLearner	RGB video	–	partially
Trajectory-based RNN	RGB video	–	–
Proposed Method	UAV video	+	+

Source: **compiled by the authors**

The dense uniform-patch baseline uses 256 uniform 8×8 tokens. Optical-flow guidance permits coarser representation in lower-motion regions (192 tokens), and three-level adaptive fragmentation reduces this to 96 tokens. UAV motion compensation refines the saliency map without directly changing that count. In the full framework, candidate fragments are ranked according to their motion-saliency scores, and the number of retained spatial fragments is limited to a maximum of 64 per vehicle-centered frame.

6. RESULTS AND DISCUSSIONS

The results show that motion-aware adaptive fragmentation improves lane-change recognition while reducing the spatial token sequence processed by the transformer.

Classification Performance. Table 4 and Fig. 1 and Fig. 2 summarize the quantitative results. Receiver Operating Characteristic (ROC) curves comparing the classification performance of the proposed motion-aware sparse framework against the dense uniform-patch tubelet baseline. Precision-Recall curves (Fig. 2) comparing the proposed motion-aware sparse framework with the dense uniform-patch tubelet baseline.

Table 3. **Comparison of lane-change recognition approaches**

Method	Sparse Processing	Transformer	Requires Trajectories	Adaptive Patches
CNN+LSTM	–	–	–	–
I3D / SlowFast	–	–	–	–
ViViT	–	+	–	–
TokenLearner	+	+	–	–
Trajectory-based RNN	–	–	+	–
Proposed Method	+	+	–	+

Source: **compiled by the authors**

Table 4. **Ablation study of proposed components**

Configuration	Accuracy	Binary F1	Precision	Recall	ROC-AUC	PR-AUC	Tokens per frame
Dense uniform-patch tubelet baseline	0.68	0.66	0.68	0.64	0.81	0.72	256
+ Optical-flow-guided fragmentation	0.71	0.69	0.70	0.68	0.85	0.78	192
+ Three-level adaptive fragmentation	0.74	0.72	0.73	0.71	0.88	0.84	96
+ UAV motion compensation	0.76	0.74	0.75	0.73	0.91	0.88	96
Full proposed framework	0.79	0.77	0.76	0.78	0.93	0.91	64

Source: **compiled by the authors**

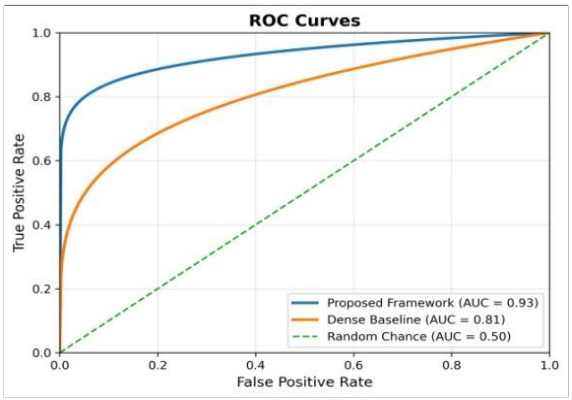


Fig. 1. **ROC curves**
Source: **compiled by the authors**

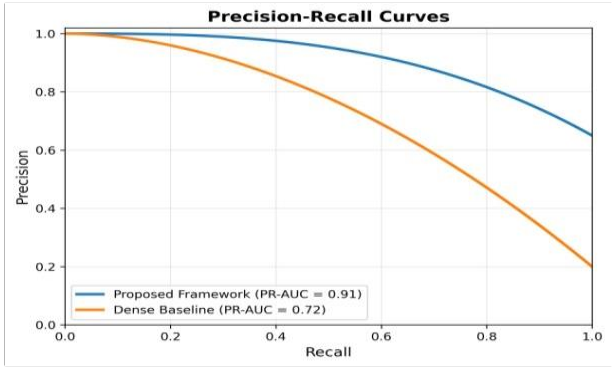


Fig. 2. **Precision-Recall curves**
Source: **compiled by the authors**

Across the ablation, Binary F1 rises from 0.66 to 0.77, ROC-AUC from 0.81 to 0.93, and PR-AUC from 0.72 to 0.91, while spatial tokens decrease from 256 to 64 per vehicle-centered frame.

Computational Efficiency. Fig. 3 shows Ablation study evaluating the impact of individual framework components on classification performance (Binary F1-score) and computational efficiency (number of processed spatial tokens).

The dense baseline processes 256 spatial tokens per vehicle-centered frame and achieves Binary F1=0.66; the proposed framework uses 64 tokens and reaches 0.77. This 75 % token reduction applies only to the Transformer input sequence, not to total end-to-end computation, which also includes homography, alignment, optical flow, and saliency estimation.

Qualitative Analysis of Adaptive Fragmentation. Fig.4, Fig.5, Fig.6 and Fig.7

illustrate tubelet extraction, motion compensation, residual optical flow, and adaptive fragmentation; fine patches concentrate on motion-salient vehicle regions while coarse patches represent low-motion areas. On Fig. 4 the tubelet serves as the input spatiotemporal region for subsequent motion analysis and lane-change recognition.

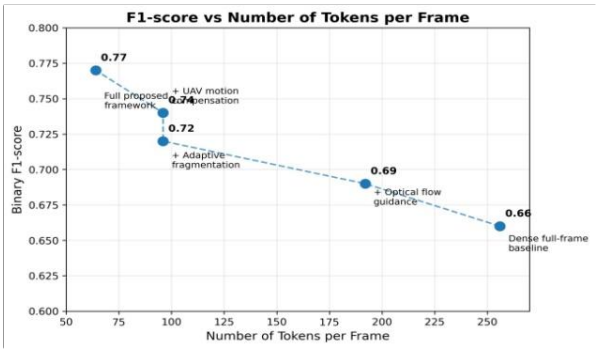


Fig. 3. **The relationship between Binary F1-score and processed spatial tokens**
Source: **compiled by the authors**



Fig. 4. **Object-centric tubelet extracted around the target vehicle**
Source: **compiled by the authors**

On Fig. 5 global UAV motion introduces substantial background displacement, resulting in motion responses distributed across the entire scene.

On Fig. 6 background motion is significantly reduced, allowing vehicle-related motion patterns to become more distinguishable.

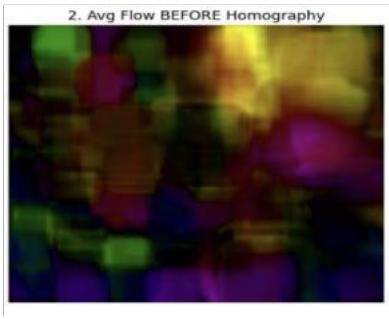


Fig. 5. **Average optical-flow field computed before homography-based motion compensation**
Source: **compiled by the authors**

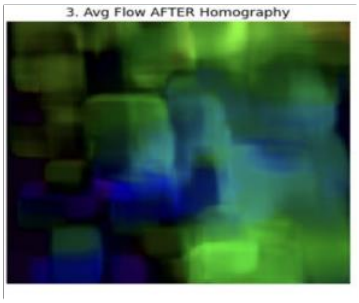


Fig. 6. **Average optical-flow field after homography-based UAV motion compensation**
Source: **compiled by the authors**

On Fig. 7 High-motion regions are represented using 8×8 patches (green), medium-motion regions use 16×16 patches (yellow) and low-motion regions are represented using 32×32 patches (red). The white bounding box indicates the target vehicle region.

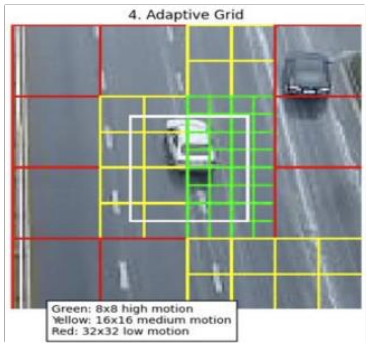


Fig. 7. **Adaptive spatial fragmentation guided by motion saliency**
Source: **compiled by the authors**

Error Analysis. Fig. 8 illustrate confusion matrix of the proposed framework that evaluated on a representative subset of 1,000 test windows (450 Lane Change and 550 Normal Driving samples). The model achieves an Accuracy of 0.79 and Binary F1-score of 0.77, indicating effective discrimination between lane-change and normal-driving events.

The confusion matrix is consistent with the reported Accuracy of 0.79 and Binary F1-score of 0.77.

Fig. 8 uses 1,000 representative test windows (450 Lane Change, 550 Normal Driving) only for error visualization; aggregate metrics are computed on the complete test set.

False negatives mainly occur with subtle or gradual lateral motion and partial occlusion, where short-window motion cues resemble normal driving.

False positives mainly arise from temporary within-lane drift, neighboring-vehicle interactions, or complex motion patterns that resemble lane changes.

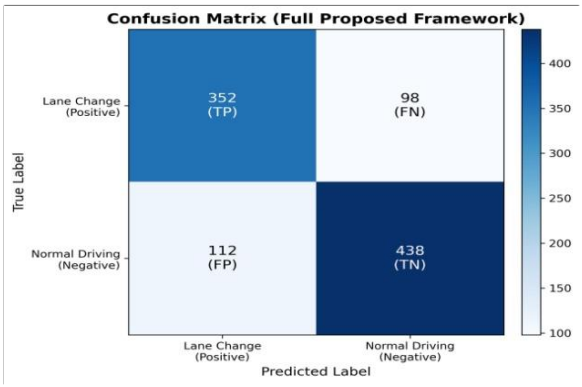


Fig. 8. **Model errors summarization on a representative test subset**
Source: **compiled by the authors**

Overall, the results support the proposed combination of motion compensation, adaptive fragmentation, and sparse spatiotemporal modeling, while indicating the need for full end-to-end efficiency evaluation before edge deployment.

CONCLUSIONS AND PROSPECTS OF FURTHER RESEARCH

This paper introduced a motion-aware sparse framework combining homography-based UAV motion compensation, weighted optical-flow saliency, adaptive fragmentation, and transformer temporal modeling for lane-change recognition.

The framework achieves Accuracy 0.79, Binary F1=0.77, ROC-AUC=0.93, and PR-AUC=0.91. The proposed framework reduces the maximum spatial token budget from 256 to 64 tokens per vehicle-centered frame, corresponding to a 75 % reduction relative to the dense uniform-patch tubelet baseline.

The results indicate a favorable recognition-efficiency trade-off and potential for resource-constrained UAV traffic monitoring, subject to future end-to-end deployment measurements.

Future work will focus on end-to-end detector/tracker integration, more robust camera-motion compensation for non-planar scenes, longer temporal contexts and overlapping-window aggregation, validation and refinement of automatic trajectory-based labels, and hardware-aware evaluation of latency, memory, and energy costs including homography and optical-flow preprocessing.

- **End-to-End Integration with Vehicle Detection and Tracking.** The current framework relies on pre-computed tracking annotations during both training and inference. Future work should investigate end-to-end integration with real-time vehicle detectors and trackers to enable fully autonomous deployment without reliance on pre-computed tracking annotations.
- **Robust Camera Motion Compensation**. Homography-based motion compensation assumes an approximately planar scene, which may be violated at low UAV altitudes or in the presence of tall objects such as buses and trucks. More robust camera motion models, e.g., depth-aware or feature-based alternatives, are a promising direction to improve compensation accuracy in complex aerial environments.
- **Long-Term Temporal Modeling**. The current time window of 8 frames might not be sufficient to capture slow or gradual lane change maneuvers. Future work should examine longer time contexts and whether aggregation strategies across overlapping windows can improve the recognition of extended maneuvers compared to the current separate window-by-window classification.
- **Improved Label Generation**. The current lane-change labels are generated automatically using trajectory-based motion criteria and fixed thresholds for lateral displacement and directional variation. Although this procedure enables scalable label generation from UAVDT tracking annotations, borderline cases may remain sensitive to the selected thresholds. In future work, data-driven threshold optimization and validation against manually annotated lane-change events to further evaluate will be explored and improve label reliability.

ACKNOWLEDGEMENTS

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

Conflict of interests: Authors declare that they do not have conflict of interests, in particular financial, personal, copyright or any of a different nature, which could have influenced the research, as well as the results published in this articles

Financing: Research was carried out without financial support

Accessibility data: All data available in digital or graphic form in the main text manuscript

Using artificial intelligence tools: During preparation manuscript used tool ChatGPT, GPT-5.6 Sol, OpenAI of artificial intelligence (AI) for grammatical and stylistic text checks. All fragments processed by him were checked and edited by the author. AI was not used to generate text, formulas, and tables, formation scientific provisions, receipt results or formulation conclusions. Scientific the provisions, results and conclusions are my own, intellectual by the contribution of the authors

REFERENCES

1. Liang K., Wang J. & Bhalerao A. “Lane change classification and prediction with action recognition networks”. *Lecture Notes in Computer Science*. *Cham*. 2023. p. 617–632. DOI: https://doi.org/10.1007/978-3-031-25056-9_39.

2. Biparva, M., et al. “Video action recognition for lane-change classification and prediction of surrounding vehicles”. *IEEE Transactions on Intelligent Vehicles*. 2022. DOI: https://doi.org/10.1109/tiv.2022.3164507 .

3. Tran, D., et al. “A closer look at spatiotemporal convolutions for action recognition”. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City, UT. 2018. DOI: https://doi.org/10.1109/cvpr.2018.00675.

4. Wang, L., et al. “Temporal segment networks for action recognition in videos” *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2019; 41 (11): 2740–2755, https://www.scopus.com/pages/publications/85052804139. DOI: https://doi.org/10.1109/tpami.2018.2868668.

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

6. Dosovitskiy, A. “An image is worth 16x16 words: Transformers for image recognition at scale”. *ICLR 2021*. 2021. DOI: https://doi.org/10.48550/arXiv.2010.11929.

7. Arnab, A., et al. “ViViT: A video vision transformer”. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal, QC, Canada. 2021. DOI: https://doi.org/10.1109/iccv48922.2021.00676.

8. Hu, M., et al. “Store and fetch immediately: Everything is all you need for space-time video super-resolution”. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2023; 37 (1): 863–871. DOI: https://doi.org/10.1609/aaai.v37i1.25165.

9. Liu, Z., et al. “Video swin transformer”. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA. 2022. DOI: https://doi.org/10.1109/cvpr52688.2022.00320.

10. Fan, H., et al. “Multiscale vision transformers”. *IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal, QC, Canada. 2021. DOI: https://doi.org/10.1109/iccv48922.2021.00675.

11. Ilg, E., et al. “FlowNet 2.0: Evolution of optical flow estimation with deep networks”. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI. 2017. DOI: https://doi.org/10.1109/cvpr.2017.179.

12. Teed, Z. & Deng, J. “RAFT: Recurrent all-pairs field transforms for optical flow”. *Computer Vision – ECCV*. 2020. p. 402–419, https://www.scopus.com/pages/publications/85097229895. DOI: https://doi.org/10.1007/978-3-030-58536-5_24.

13. Hui, T.-W., Tang, X. & Loy, C. C. “LiteFlowNet: A lightweight convolutional neural network for optical flow estimation”. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City, *UT*, *USA*. 2018. DOI: https://doi.org/10.1109/cvpr.2018.00936.

14. Argaw, D. M., et al. “Optical flow estimation from a single motion-blurred image”. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2021; 35 (2): 891–900. DOI: https://doi.org/10.48550/arXiv.2103.02996.

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

16. Zaheer, M., et al. “BigBird: Transformers for longer sequences”. *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*. Vancouver, Canada. 2020. DOI: https://doi.org/10.48550/arXiv.2007.14062.

17. Kitaev, N. & Kaiser, L., Levskaya A. “Reformer: The efficient transformer”. *ICLR*. 2020. DOI: https://doi.org/10.48550/arXiv.2001.04451.

18. 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.

19. Farnebäck, G. “Two-frame motion estimation based on polynomial expansion”. *Image Analysis: 13th Scandinavian Conference, SCIA, Lecture Notes in Computer Science*. 2003; 2749: 363–370, https://www.scopus.com/authid/detail.uri?authorId=56628974700. DOI: https://doi.org/10.1007/3-540-45103-X_50.

20.Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G. & Jégou, H. “Going deeper with image transformers”. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021. p. 32–42. DOI: https://doi.org/10.1109/ICCV48922.2021.00010.

21.He, H. & Garcia, E. A. “Learning from imbalanced data”. *IEEE Transactions on Knowledge and Data Engineering*. 2009; 21 (9): 1263–1284. DOI: https://doi.org/10.1109/TKDE.2008.239.

22.Saito, T. & Rehmsmeier, M. “The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets”. *PLOS ONE*. 2015; 10 (3): e0118432. DOI: https://doi.org/10.1371/journal.pone.0118432.

23.Bertasius, G., Wang, H. & Torresani, L. “Is space-time attention all you need for video understanding?” *Proceedings of the 38th International Conference on Machine Learning (ICML)*.2021; 139: 813–824. DOI: https://doi.org/10.48550/arXiv.2102.05095.

24.Du, D., et al. “The unmanned aerial vehicle benchmark: Object detection and tracking”. *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018. p. 375–391, https://www.scopus.com/pages/publications/85055113352. DOI: https://doi.org/10.1007/978-3-030-01249-6_23.

25.Barmpounakis, E., Sauvin, G. M. & Geroliminis, N. “Lane detection and lane-changing identification with high-resolution data from a swarm of drones”. *Transportation Research Record: Journal of the Transportation Research Board*. 2020; 2674 (7): 63–75. DOI: https://doi.org/10.1177/0361198120920627.

26.Sun, S., et al. “Optical flow guided feature: A fast and robust motion representation for video action recognition”. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City, UT. 2018. DOI: https://doi.org/10.1109/cvpr.2018.00151.

27. Wang, L. & Zhang, F. “Decoupling ego-motion from target dynamics via Dual-interval motion cues for UAV detection”. 2026. DOI: https://doi.org/10.48550/arXiv.2605.22605.

Received 03.09.2026
Received after revision 22.10.2026
Accepted 30.10.2026

DOI: https://doi.org/10.15276/aait.09.2026.37
УДК 004:83

Метод адаптивної фрагментації з урахуванням руху та трансформерного темпорального моделювання для розпізнавання маневрів зміни смуги руху у відеопослідовностях дорожнього руху, отриманих з безпілотних літальних апаратів

Норматова Тетяна Віталіївна[1)]
ORCID: https://orcid.org/0009-0004-3503-6350; tetiana.normatova@nure.ua. Scopus Author ID: 60278045600
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

Розпізнавання маневрів зміни смуги руху у відео дорожнього руху, отриманих за допомогою безпілотних літальних апаратів, є складним завданням через невеликий розмір транспортних засобів, глобальний рух камери та обчислювальну вартість щільної обробки відеоданих. **Запропоновано** фреймворк адаптивної фрагментації з урахуванням руху, який поєднує компенсацію руху камери на основі гомографії, оцінювання значущості руху за допомогою оптичного потоку та розріджене просторово-часове моделювання на основі Transformer. Області зі значущим рухом подаються за допомогою дрібних фрагментів, тоді як області з нижчою руховою активністю обробляються більшими фрагментами для зменшення надлишкового просторового представлення. **Фреймворк оцінено** на наборі даних для розпізнавання зміни смуги руху, сформованому на основі трекінгових анотацій набору Unmanned Aerial Vehicle Benchmark із застосуванням автоматичної процедури генерації міток за траєкторіями. Сформований набір даних містить чотириста одну тисячу триста сорок п'ять восьмикадрових часових вікон, отриманих із тридцяти відеопослідовностей та тисячу триста шістнадцять треків транспортних засобів, з яких п’ятдесят чотири тисячі сімсот сімдесят п’ять належать до класу зміни смуги руху, а триста

сорок шість тисяч п’ятсот сімдесят - до класу нормального руху**. Експериментальні результати демонструють** точність у розмірі нуль крапка сімдесят дев’ять, бінарну F1-міру у розмірі нуль крапка сімдесят сім, площу під кривою робочих характеристик приймача у розмірі нуль крапка дев’яноста три та площу під кривою «точність–повнота» у розмірі нуль крапка дев’яноста один. Запропоноване адаптивне представлення обмежує кількість просторових токенів до максимум шістдесят чотири на один об’єктно-центрований кадр, що відповідає сімдесят п’яти відсоткам скороченню максимального бюджету просторових токенів порівняно з базовим методом щільних рівномірних фрагментів у tubelet, який використовує двісті п’ятдесят шість токенів. **Результати свідчать**, що адаптивна фрагментація з урахуванням руху дає змогу зберігати конкурентоспроможну якість розпізнавання зміни смуги руху, водночас істотно скорочуючи просторову послідовність токенів, яку обробляє Transformer. Це створює перспективну основу для обчислювально ефективного моніторингу дорожнього руху за допомогою безпілотних літальних апаратів, однак повна оцінка ефективності розгортання на крайових пристроях потребує подальших наскрізних вимірювань, що також ураховують витрати на оцінювання гомографії та обчислення оптичного потоку.

Ключові слова: класифікація відео; виявлення зміни смуги руху; оптичний потік; адаптивна фрагментація; трансформер; просторово-часове моделювання; аналіз відео з безпілотних літальних апаратів

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
Галузь досліджень: Обробка зображень та відео, аналіз даних



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 Харківського національного університету радіоелектроніки, пр. Науки, 14. Харків, 61166, Україна
Галузь досліджень: Обробка зображень та відео, аналіз даних