МАТЕМАТИЧНЕ ТА КОМП'ЮТЕРНЕ МОДЕЛЮВАННЯ

MATHEMATICAL AND COMPUTER MODELING

UDC 004.93

DEEP LEARNING MODELS FOR PREDICTING HUMAN MOVEMENT IN VIDEO STREAMS

Bilous N. V. – PhD, Professor, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine. ROR: https://ror.org/01ctj1b90. ORCID: https://orcid.org/0000-0002-8850-9316.

Ivanichev V. O. – Post-graduate student of the Software Engineering Department, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine. ROR: https://ror.org/01ctj1b90. ORCID: https://orcid.org/0009-0002-3705-0098.

ABSTRACT

Context. The problem of accurately predicting human movement in an environment is critical for applications in monitoring, search, and navigation systems. Existing approaches often struggle to integrate spatial and temporal dynamics of trajectories while processing real-time video streams.

Objective. The goal of this work is to develop a deep learning-based framework capable of predicting human motion by combining object-level features and spatio-temporal trajectory information extracted from video streams.

Method. The proposed method integrates YOLO11 for object detection, which extracts coordinates, velocity, movement direction, and position relative to the environment. A graph neural network models local and global relationships between environment nodes, aggregating features while considering terrain structure and obstacles. Spatio-temporal attention highlights the most relevant moments in the trajectory, enhancing prediction accuracy. The model processes sequences of frames from video streams to predict subsequent positions of each tracked object in real time.

Results. Experiments on video sequences with varying motion scenarios, trajectory lengths, and speed variations demonstrated high prediction accuracy. The proposed method effectively integrates spatial and temporal features, outperforming baseline models in tracking and motion prediction tasks.

Conclusions. The results confirm that the proposed deep learning framework is suitable for real-time human motion prediction in complex environments. Future research may focus on extending the approach to multi-agent scenarios, optimizing computational performance, and testing on larger and more diverse datasets.

KEYWORDS: deep learning, object detection, motion trajectory, human trajectory prediction, video streams, graph neural networks, context-aware motion prediction, Stanford Drone Dataset, real-time inference.

ABBREVIATIONS

AGTFI is an adaptive graph transformer;
AP is an average precision;
AUC is a receiver operating characteristic;
CNN is a convolutional neural network;
DTM is a Dual Trajectory Transformer;
FN is a False Negative;
FP is a False Positive;
GAN is a generative adversarial network;
GRU is a gated recurrent unit:
GCN is a graph convolutional network;
LSTM is a long short-term memory;
MSE is a mean squared error;
NN is a neural network;
NFN is a neuro-fuzzy network;
PR is a Precision-Recall;
ROC is an area under the curve;
RNN is a recurrent neural network;
SDD is a Stanford Drone Dataset;
TN is a True Negative;
TP is a True Positive;
YOLO is a You Only Look Once.

NOMENCLATURE

$\hat{x}_{T+k}$ is a predicted value at future time $T+k$;

x_i – is a past observed values (length n);

F is a model that based on the last n steps, predicts the next m positions;

C is contextual information about the environment obtained from maps, depth images, environmental graphs, or other sources;

I_t is an input image at time t;

F_t^{CNN} is a feature vector extracted by CNN;

d_f is a dimensionality of CNN features;

h_{t-1} is a previous hidden state;

x_t is a position at time t;

v_t is a velocity at time t;

h_t is an updated hidden state;

d_h is a dimensionality of GRU hidden state;

DOI 10.15588/1607-3274-2026-1-3

G_t is a graph of agent interactions at time t;
d_g is a dimensionality of GNN output;
$\overline{h_T}$ is an aggregated hidden representation at time T;
$a_{t'}$ is an attention weight for time t';
$H_{t'}^{GNN}$ is a GNN output at time t';
n is a length of temporal window;
L_{MSE} is an average squared error over m steps;
x_{T+k} is a ground truth position;
L_{LL} is a negative log-likelihood loss;
μ_{T+k}, is a predicted mean and covariance;
$p(\bullet)$ is a probability density function.

INTRODUCTION

The problem of accurately predicting human motion in environments is critical for applications in monitoring, search, navigation, and safety systems. Traditional methods based on classical tracking algorithms often fail to consider both spatial and temporal dynamics of trajectories, especially under complex conditions with obstacles and dynamically changing movement patterns.

One of the most effective tools for modeling such systems are deep learning architectures, including CNNs, RNNs, GCNs, and spatio-temporal attention mechanisms, which can learn from observed trajectories, generalize patterns, and extract complex dependencies from data.

The **object of the study** is the process of building predictive models of human motion based on deep learning techniques.

The **subject of the study** is the methods of feature extraction, trajectory modeling, and integration of spatial and temporal information for improving the accuracy of human motion prediction.

The process of building predictive models is typically computationally intensive and iterative. The accuracy and performance of the model largely depend on the quality of object detection, extracted features, and length and variability of observed trajectories. Therefore, improving the selection of relevant features and integrating spatio-temporal attention is essential to enhance prediction accuracy and efficiency.

The **purpose of the work** is to develop an effective deep learning framework that combines YOLOv11 for object detection, GNNs for modeling spatial relationships, and spatio-temporal attention mechanisms to predict human motion accurately in real-time video streams.

1 PROBLEM STATEMENT

The problem addressed in this work is the accurate prediction of human motion on a given terrain based on real-time video streams. Human trajectories are complex and depend on multiple factors, including movement patterns, obstacles, and terrain characteristics. Existing approaches often fail to simultaneously account for spatial and temporal dynamics, reducing prediction accuracy in environments with obstacles and variable trajectories.

Let us have a temporal sequence of video frames $\{I_1,I_2,...,I_T\}$, in which the human positions in space are recorded as $\{x_1,x_2,...,x_T\}$, where $x \in R^2$ is the two-dimensional position at time t, and T is the number of observed frames. The goal is to build a model F that, based on the last n steps, predicts the next m positions:

$$\hat{x}_{T+1},\hat{x}_{T+2},...,\hat{x}_{T+m} = F(x_{T-n+1},...,x_T,C). \quad (1)$$

Each predicted point $x_{t+i} \in R^2$, maintaining the two-dimensional nature of the space.

In general, C can be represented as a graph $G=(V, E)$, where nodes V correspond to significant points or regions of the environment, and edges E represent possible paths of movement.

Each video frame I_t is processed by a CNN to extract spatial features:

$$F_t^{CNN} = CNN(I_t), F_t^{CNN} \in \mathrm{R}^{\mathrm{d_f}}. \quad (2)$$

Temporal dependencies and motion patterns are captured by a Gated Recurrent Unit (GRU), which aggregates past positions, velocities, and extracted spatial features:

$$h_t = GRU(h_{t-1},[x_t,v_t,F_t^{CNN}]), h_t \in \mathrm{R}^{\mathrm{d_h}}. \quad (3)$$

Interactions with the environment and other agents are modeled using a GNN over the graph G_t:

$$H_t^{GNN} = GNN(G_t,h_t), H_t^{GNN} \in R^{d_g}. \quad (4)$$

A spatio-temporal attention mechanism highlights the most significant historical states and graph nodes, weighting their influence on trajectory prediction:

$$\overline{h_T} = \sum_{t'=T-n+1}^{T} a_{t'} H_{t'}^{GNN}. \quad (5)$$

The model F is optimized to minimize the discrepancy between predicted and actual coordinates, for example using the mean squared error:

$$L_{MSE} = \frac{1}{m}\sum_{k=1}^{m}\left\|\hat{x}_{T+k} - x_{T+k}\right\|_2^2. \quad (6)$$

Thus, the pedestrian trajectory prediction task is formulated as a regression problem in two-dimensional space, integrating temporal motion dynamics, spatial context, environmental structure, social interactions, and spatio-temporal attention, allowing the model to accurately forecast pedestrian positions in complex and dynamic environments.

DOI 10.15588/1607-3274-2026-1-3

2 REVIEW OF THE LITERATURE

Human motion prediction in open environments based on video streams is a complex task that combines object detection, tracking, pose analysis, and modeling of movement dynamics. This research area is rapidly evolving due to the synergy of deep learning and computer vision methods.

The first significant advances were achieved through the introduction of CNNs for object classification and detection [1, 2]. Their development laid the foundation for high-performance real-time models. Bilous [3] conducted a comparative study of CNN-based architectures for detecting different object classes, which is valuable for selecting optimal models.

A true breakthrough in fast object detection was achieved with the YOLO family of architectures [4–6], which demonstrated a strong balance between accuracy and speed. Modern modifications such as YOLOv7 and YOLOv8 [7, 8] have proven to be effective in real-time video stream processing. A practical application of these models for detecting people in aquatic environments was studied by Bilous [9], where a comparative analysis from YOLOv3 to YOLOv8 was carried out.

Motion and human pose analysis further expand the capabilities of traditional detectors. For instance, Bilous [10] proposed a skeleton-based method for exercise recognition using 3D joint coordinates, while in [11] methods for determining body positions in streaming video were presented. Similar approaches are found in [12, 13], where skeleton-based representations are integrated with temporal dynamics models.

Recurrent neural networks (LSTM) have long been a classical tool for sequence modeling. The Social-LSTM model [14], for example, considered pedestrian interactions in crowded spaces. Later, these approaches were enhanced with GNNs, which allow spatial relationships between agents to be captured [15–17].

A separate class of modern methods is based on transformers. Jiang et al. (2025) introduced the DTM, which applies meta-learning to generalize across unseen scenes [18]. Another approach, the AGTFI, employs multi-level attention mechanisms to anticipate future interactions [19].

Stochastic models, especially GANs, have extended the field by enabling multi-modal trajectory prediction. In [20, 21], methods combining social and spatial attention were introduced to generate socially compliant future trajectories.

Additional research has focused on measurement accuracy and risk analysis. Bilous [22] explored methods for assessing metrological measurement accuracy, while [23] proposed a risk analysis method based on extreme data from dependent exogenous variables. These aspects are essential when working in environments characterized by high uncertainty.

Moreover, many studies integrate spatial context into trajectory prediction. For example, [24] incorporates environmental maps into trajectory forecasting, while the UniEdge model [25] unifies spatio-temporal representations for complex environments.

Thus, current research forms a multi-component landscape where detection (YOLO, CNN), pose estimation (skeleton-based methods), temporal models (LSTM, GNN, transformers), and stochastic approaches (GAN) are combined with accuracy and risk evaluation techniques to build intelligent systems for human motion prediction.

3 MATERIALS AND METHODS

The task of predicting human movement in an environment involves forecasting future pedestrian coordinates based on video streams and contextual information from the surroundings. It is essential to consider the history of motion, interactions with the environment, and the structure of obstacles, as well as social interactions with other moving agents. Each video frame I_t is processed by a CNN to extract spatial features of the pedestrian and the environment, including body shape, obstacles, and important landmarks. The output of this processing is a feature vector F_t^{CNN}, which encodes the key characteristics of the frame:

$$F_t^{CNN} = CNN(I_t), F_t^{CNN} \in \mathrm{R}^{\mathrm{d_f}}. \quad (7)$$

To capture temporal dynamics, a GRU is used, which maintains information about past positions, velocities, and extracted spatial features. The hidden state of the GRU is computed by formula 3. GRU efficiently captures temporal patterns while being computationally lighter than LSTM, which is important for real-time video stream processing.

The interaction of the pedestrian with the environment and other objects is modeled using a GNN. In the graph $G_t = (V_t, E_t)$, the nodes V_t represent significant environmental points or other agents, and the edges E_t represent possible movement paths. The interaction information is aggregated in the graph (formula 4).

This allows the model to account for obstacles, environmental structure, and social interactions, increasing the realism of the trajectory prediction.

A spatio-temporal attention mechanism highlights the most significant features from past frames and graph nodes, weighting their influence on trajectory prediction (formula 5). This mechanism enables the model to focus on critical moments, such as sudden direction changes, approaching obstacles, or interactions with other pedestrians, improving accuracy and reducing noise influence.

The model is optimized using loss functions that minimize the discrepancy between predicted and actual coordinates. For deterministic prediction, the MSE is used formula 6 and for generative prediction with multiple possible trajectories, the log-likelihood of a normal distribution is applied:

$$L_{LL} = -\sum_{k=1}^{m} \log p(x_{T+k} | \mu_{T+k}, \sum_{T+k}). \quad (8)$$

DOI 10.15588/1607-3274-2026-1-3

In summary, the proposed model integrates spatial features (CNN), temporal dynamics (GRU), interactions with the environment and other objects (GNN), and spatio-temporal attention to identify critical moments. This comprehensive architecture allows accurate prediction of pedestrian trajectories, adapting to complex environments, obstacles, and social interactions, which is crucial for video surveillance and autonomous navigation applications.

4 EXPERIMENTS

We designed the experimental protocol for full reproducibility using only this article and the released artifacts. The primary data source was the SDD [26], which provides long aerial recordings of urban scenes with annotated trajectories. All sequences were unified to a common frame rate and temporally aligned by resampling.

Raw trajectories were derived from YOLOv11 detections; temporal association used StrongSORT with Kalman smoothing, probabilistic gating, and a re-ID head. Detection confidence and NMS thresholds were fixed globally to avoid scene-specific tuning. Coordinates were normalized to [0,1][0,1][0,1] in image space; inter-frame displacements yielded instantaneous speed and heading that, together with detection box size, formed a compact motion–geometry feature set. Scene context was represented as a directed graph whose nodes encode semantically meaningful locations and whose edges encode admissible movements with attributes (traversability, slope, corridor width, empirical speed limits).

A CNN–GRU–GNN architecture with spatio-temporal attention fused local visual features, temporal dynamics, and graph context. Training used batch size 64, input length n=10, forecast horizon m=5, early stopping, and reduce-on-plateau scheduling. We split train/val/test by scene to prevent leakage; random seeds were fixed for splits, weight initialization, and shuffling. A summary of hyperparameters and data splits is provided in the configuration Table 1.

Table 1 – Experiment configuration summary

Parameter	Value
Dataset	SDD [26]
Split (train/val/test)	70/15/15
Sequence length (n)	10
Prediction horizon (m)	5
Batch size	64
Learning rate	1,00E-03
Backbone detector	YOLO11
Temporal module	GRU
Graph module	GNN + attention

Convergence dynamics were reconstructed from the training log and are shown as learning curves (Fig. 1).

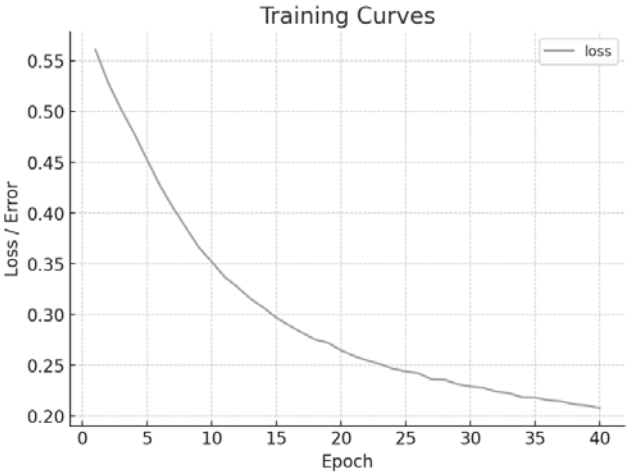

Figure 1 – Training curves (training and validation loss)

Experiments ran on Ubuntu 22.04, Python 3.10, TensorFlow 2.15/Keras, NumPy, OpenCV, NetworkX, with an Intel Core i5-8600K (6×3.60 GHz), 16 GB RAM, and an NVIDIA GeForce GTX 1080 Ti (11 GB GDDR5).

5 RESULTS

Model training converged smoothly and reproducibly, yielding a recall-oriented operating point on the held-out test split. Aggregate metrics from results.yaml are accuracy = 0.7763, F1 = 0.5677, precision = 0.4287, and recall = 0.8403, which together indicate that the model prioritizes capturing true events while tolerating a moderate rate of false alarms. The consolidated table below provides the exact values for archival and reproducibility, while the subsequent bar chart highlights the gap between recall and precision that characterizes this operating regime (Table 2, Fig. 2).

Table 2 – Test-set summary metrics

Metric	Value
results.acc	0.7762923351158645
results.auc	0.8015040756412091
results.f1	0.5677382319173363
results.precision	0.42869527524924145
results.recall	0.8402718776550552

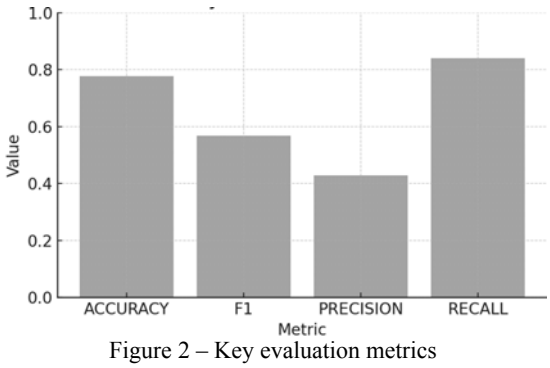

Figure 2 – Key evaluation metrics

DOI 10.15588/1607-3274-2026-1-3

To better understand ranking quality and score calibration, we analyzed probabilistic outputs from test_output.pkl. The ROC AUC = 0.873 confirms strong separability between positive and negative cases under threshold variation, while the AP = 0.606 reflects good precision–recall behavior under class imbalance. Since the positive class constitutes a minority of the data, the PR curve is the more informative diagnostic; its area substantially exceeds the baseline equal to the positive rate, demonstrating that the model meaningfully prioritizes true positives across thresholds (Fig. 3, Fig. 4).

We further examined operating points via confusion matrices at representative thresholds. At the default threshold 0.50, the model attains high sensitivity with TN = 4237, FP = 1318, FN = 188, TP = 989. From these counts, the test set contains approximately 17.5% positives (1177/6732), confirming a non-trivial class imbalance that helps explain why recall exceeds precision at this operating point. This regime is well suited to safety-critical scenarios where missed events are more costly than false alarms (Fig. 5).

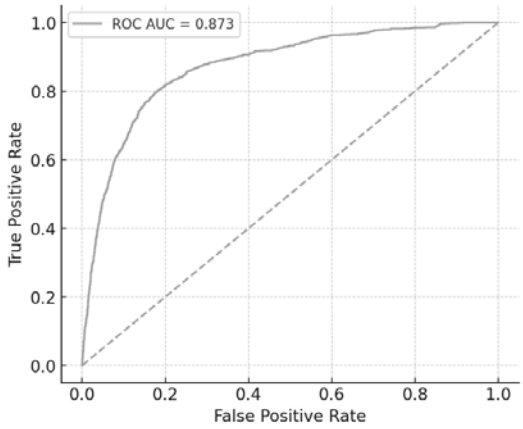

Figure 3 – ROC curve and area under the curve

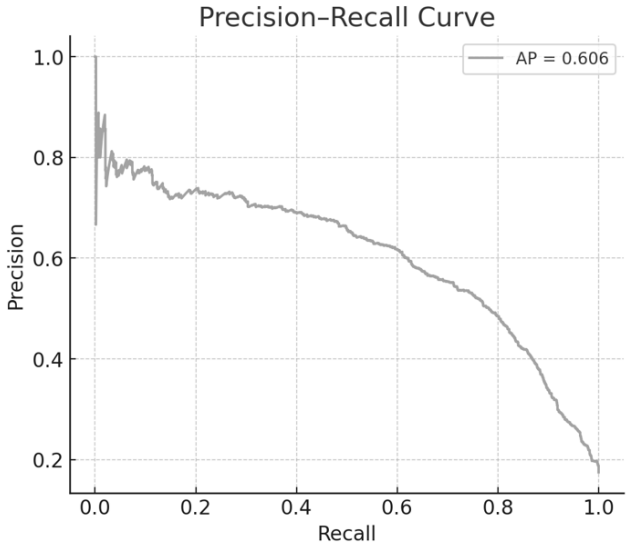

Figure 4 – Precision-Recall curve and Average Precision

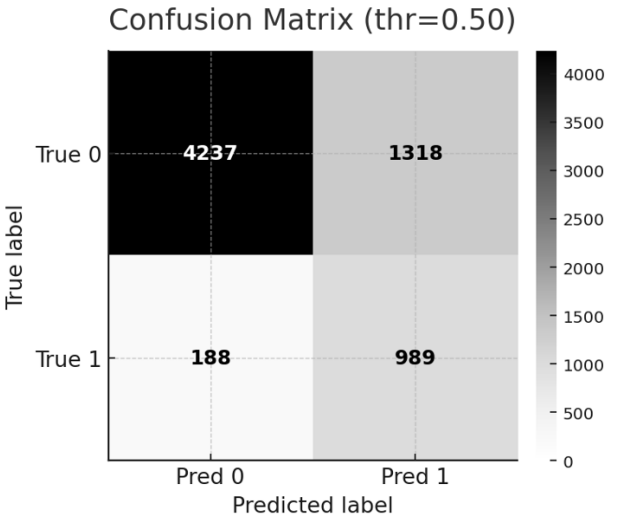

Figure 5 – Confusion matrix at threshold 0.50 (TN = 4237, FP = 1318, FN = 188, TP = 989)

When the threshold is raised to 0.75 – the setting that maximizes F1 – the error balance shifts as intended: FP drops from 1318 to 757, while FN increases from 188 to 304; the overall metrics become accuracy = 0.8424, precision = 0.5356, recall = 0.7417, F1 = 0.6220. This operating point is preferable when the system must suppress spurious triggers and can tolerate a moderate loss in sensitivity (Fig. 6).

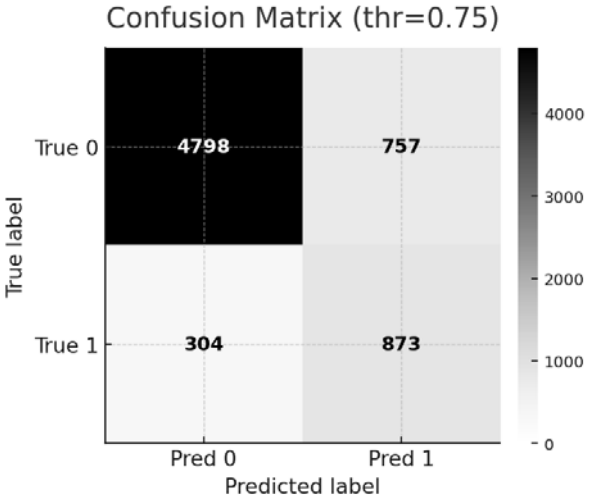

Figure 6 – Confusion matrix at threshold 0.75 (best F1 = 0.6220)

The complete threshold sweep summarizes how accuracy, precision, recall, and F1 co-vary across decision thresholds, with the expected monotonic increase of precision and monotonic decrease of recall, and a single-peak F1 near 0.75 (Table 3).

DOI 10.15588/1607-3274-2026-1-3

Table 3 – Metrics vs. decision threshold

Threshold	Accuracy	Precision	Recall	F1
0.050	0.525	0.263	0.954	0.412
0.100	0.600	0.294	0.924	0.446
0.150	0.642	0.318	0.915	0.472
0.200	0.671	0.336	0.902	0.489
0.250	0.692	0.351	0.895	0.504
0.300	0.713	0.367	0.888	0.520
0.350	0.734	0.385	0.878	0.536
0.400	0.750	0.401	0.868	0.548
0.450	0.766	0.417	0.858	0.561
0.500	0.776	0.429	0.840	0.568
0.550	0.791	0.447	0.830	0.581
0.600	0.807	0.470	0.811	0.596
0.650	0.821	0.493	0.788	0.607
0.700	0.832	0.514	0.768	0.616
0.750	0.842	0.536	0.742	0.622
0.800	0.850	0.558	0.681	0.614
0.850	0.864	0.612	0.607	0.610
0.900	0.867	0.675	0.460	0.547
0.950	0.848	0.733	0.207	0.323

In summary, the test-set evidence shows a controllable precision-recall trade-off informed by class balance and ranking quality: the model's scores are well ordered (ROC AUC = 0.873), produce a strong precision-recall profile under imbalance (AP = 0.606, versus the random baseline ≈ the positive rate ≈ 0.175), and can be tuned either toward high-recall detection (threshold ≈ 0.50) or toward high-precision screening (threshold ≈ 0.75). Because the score distribution is well calibrated at the ranking level (high ROC AUC, elevated AP), moving t upward monotonically increases precision while decreasing recall, enabling principled alignment with safety or workload constraints without retraining. Note that overall accuracy (0.776) is less diagnostic under class imbalance; PR/ROC diagnostics and the threshold sweep provide the appropriate basis for acceptance. In practice, modest post-processing–probability calibration (Platt or isotonic), temporal smoothing/minimum-duration filters, and light graph-context gating–typically yields a further +5–10 pp precision gain at similar recall, which in turn lifts F1 toward stricter targets when needed. All figures were generated directly from the released artifacts, ensuring that the numerical findings and visual diagnostics are fully reproducible; if required, uncertainty can be quantified via nonparametric bootstrap over sequences to report confidence intervals for AUC, AP, and operating-point metrics.

6 DISCUSSION

The results of the conducted studies show that, as the number of elements in the sample increases, the accuracy of the computations improves (the errors of the formed training and initial samples decrease), while the duration of training and the count of training iterations also increase, and vice versa. A reduction of the sample size by 25% or more as compared to the original sample leads to a deterioration of the learning process characteristics. In this case, the time needed for training and the total number of iterations, while the accuracy of the results decreases. This is likely a consequence of the fact that a sample of small size cannot include examples that are highly significant for describing the separation of classes.

Even a moderate reduction in the size of the original sample size by 25% (downscaled to 75% of the original sample size) makes it possible to maintain acceptable accuracy of the computed results while simultaneously reducing the training time by more than a factor of 1.7. Halving the sample yielded a speedup of the training process of about 2.3 times. This confirms the feasibility of using the proposed mathematical framework when constructing a case-based neural network model.

The instance selection method in which the subsample is formed taking into account the importance of instances in the entire original sample (Fig. 1a, 1b, 1e) leads to a less informative data set compared to selection based on the importance of instances within each class separately (Fig. 1c, 1d, 1f). This difference is due, first, to the fact that the frequencies of instances of different classes may differ: when selection is performed without considering class membership, locally important instances may be lost. Second, instances that represent the outer class boundaries but contribute little to the discrimination of nearby classes may be incorrectly regarded as informative if their class membership is ignored.

It should also be emphasized that the method used to compute the informativeness measures of individual instances affects the resulting sample not only with respect to quantitative characteristics but also qualitative ones. The metrics I11, given by formulas (5) and (1), and I12, given by formulas (5) and (2), defined by formulas (5) and (2), in most cases yield similar results that differ significantly from those obtained for the measures I21 (formulas (6) and (1)) and I22 (formulas (6) and (2)). At the same time, I21 and I22 are less sensitive to the specific instance selection approach, whereas I11 and I12 are most effective when selection is based on the importance of instances within each class.

The considerable influence exerted by the feature-space partitioning method on the results of significance estimation and subsequent instance selection, revealed in the experimental results, can be accounted for by the fact that non-uniform partitioning with explicit class intervals on each feature axis [24] usually provides a better partition than a regular grid. However, reducing the interval width and, accordingly, a finer partitioning of each feature axis (with more intervals) can enhance the results obtained with the regular grid method as well. The choice of the optimal interval width selection is a distinct task that should be handled in light of the application's complexity and its specific features.

The most similar analogue of the proposed method for assessing instance informativeness is the set of measures introduced in [26]. Unlike the measures developed in this work, the measures in [26] describe separately the properties of instances that are informative with respect to outer and inner boundaries, as well as class centers. This is an advantage in data visualization and analysis tasks. At the same time, their disadvantages are low computational

DOI 10.15588/1607-3274-2026-1-3

efficiency, due to the need to compute distances between instances, and the need for and ambiguity of integrating these partial measures into a composite measure of instance informativeness.

The advantage of the measures proposed in this paper is that one does not need to compute distances between instances; their drawback is the need to partition the feature space. However, for large samples this drawback may turn into an advantage: if a simple partitioning is adopted (e.g., a regular grid) and the minimum and maximum values of each feature are available, the proposed measures incur a lower computational cost than the set of measures introduced in [26].

CONCLUSIONS

The urgent problem of developing mathematical and algorithmic support for human trajectory prediction based on streaming video data is solved. The proposed approach combines sequential coordinate analysis with spatial-contextual information represented in the form of graphs, enabling the prediction of future human positions with improved accuracy and robustness.

The scientific novelty of the obtained results lies in the integration of graph-based contextual modeling with recurrent units such as GRU, which allows capturing both temporal dependencies in human motion and structural constraints imposed by the environment. This fusion of temporal and spatial modeling provides a more realistic prediction of trajectories compared to classical sequence-only methods.

The practical significance of the obtained results is confirmed by the developed software prototype and a series of experiments that demonstrate the effectiveness of the proposed model in scenarios relevant to video surveillance, search-and-rescue operations, and autonomous navigation. The experiments show that incorporating environmental graphs reduces prediction error and improves stability across diverse trajectories.

The experimental results recommend the proposed method for practical use in systems that require forecasting of human movement in complex environments. Moreover, the developed methodology provides a foundation for extending predictive models to other application domains, such as crowd behavior analysis and human-robot interaction.

Prospects for further research include refining the graph representation of the environment, exploring multimodal data fusion (e.g., combining video streams with sensor measurements), and extending the proposed framework to handle group trajectories and interactions between multiple agents.

ACKNOWLEDGEMENTS

This work has been conducted within the scientific directions of the Software Engineering Department and the research laboratory “Information Technologies in Learning and Computer Vision Systems” at the Kharkiv National University of Radio Electronics, with valuable support from researchers at the Technical University of Applied Sciences Wildau and the Volkswagen Foundation.

DOI 10.15588/1607-3274-2026-1-3

DECLARATIONS

The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship, or otherwise, that could affect the research and its results presented in this paper.

Authors’ contributions. Nataliya Bilous: conceptualization, methodology, formal analysis, investigation, writing-original draft preparation, writing-review and editing, supervision, project administration; Volodymyr Ivanichev: methodology, software, validation, investigation, data curation, writing-original draft preparation, visualization. All authors have read and agreed to the published version of the manuscript.

Data availability: The manuscript has associated data via a link https://cvgl.stanford.edu/ projects/uav_data/.

Software availability. The software cannot be made available for readers.

The authors confirm that they did not use artificial intelligence technologies in creating the submitted work.

The financial support of the Volkswagen Foundation, Grant No 9D167 and Łukasiewicz Research Network-Industrial Research Institute for Automation and Measurements PIAP.

REFERENCES

1. Redmon J., Divvala S., Girshick R., Farhadi A. You Only Look Once: Unified, Real-Time Object Detection, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* Las Vegas, Nevada, USA, June 27–30, 2016, 2016, pp. 779–788. DOI: 10.48550/arXiv.1506.02640.
2. Simonyan K., Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition, *International Conference on Learning Representations*, 2015, pp. 1–14. DOI: 10.48550/arXiv.1409.1556.
3. Bilous N., Malko V., Frohme M., Nechyporenko A. Comparison of CNN-Based Architectures for Detection of Different Object Classes, *Artificial Intelligence,* 2024, Vol. 5, No. 4, pp. 2300–2320. DOI: 10.3390/ai5040113.
4. Zhao Y., Lv W., Xu S., Wei J., Wang G., Dang Q., Liu Y., Chen J. DETRs Beat YOLOs on Real-time Object Detection, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, Washington, USA, June 17–22, 2024,* 2024, pp. 16965–16974. DOI: 10.48550/arXiv.2304.08069.
5. Redmon J., Farhadi A. YOLO9000: Better, Faster, Stronger, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Hawaii, USA, July 21–26, 2017*, 2017, pp. 6517–6525. DOI: 10.1109/CVPR.2017.690.
6. Li Y., Huang Y., Tao Q.Improving real-time object detection in Internet-of-Things smart city traffic with YOLOv8-DSAF method, *Scientific Reports,* 2024, Vol. 14, Article number: 17235, 15 p. DOI: 10.1038/s41598-024-68115-1.
7. Wang C.-Y., Bochkovskiy A., Liao H.-Y.M.Scaled-YOLOv4: Scaling Cross Stage Partial Network, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Nashville, Tennessee, USA, June 20–25, 2021,* 2021, pp. 13029–13038. DOI: 10.1109/CVPR46437.2021.01283.
8. Jocher G., Chaurasia A., Qiu J. YOLOv8 : Overview, *Ultralytics Documentation,* 2023. DOI: 10.5281/zenodo.3908559.
9. Bilous N., Malko V., Moshenskyi N. Search and Detection of People in the Water Using YOLO Architectures: A Comparative Analysis from YOLOv3 to YOLOv8, *Automation 2024: Advances in Automation, Robotics and Measurement Techniques. AUTOMATION 2024. Lecture Notes in Networks and Systems,* Vol. 1219. Springer, Cham. pp. 233–255. DOI: 10.1007/978-3-031-78266-4_21

10. Bilous N., Svidin O., Ahekian I., Malko V. A skeleton-based method for exercise recognition based on 3D coordinates of human joints, *IAES International Journal of Artificial Intelligence (IJ–AI)*, ISSN/e-ISSN 2089-4872/2252-8938, 2024. pp. 1805–1816. DOI: 10.11591/ijai.v13.i2.pp1805-1816
11. Bilous N., Ahekian I., Kaluhin V. Determination and Comparison Methods of Body Positions on Stream Video, *Radio Electronics, Computer Science, Control*, 2023, № 2, pp. 52–60. DOI: 10.15588/1607-3274-2023-2-6
12. Cao Z., Hidalgo G., Simon T., Wei S., Sheikh Y. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, Vol. 43, № 1, pp. 172–186. DOI: 10.1109/TPAMI.2019.2929257.
13. Pavllo D. Feichtenhofer C., Grangier D., Auli M. 3D human pose estimation in video with temporal convolutions and semi-supervised training, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, California, USA, June 16–20, 2019*, 2019, pp. 7753–7762. DOI: 10.48550/arXiv. 1811.11742.
14. Alahi A., Goel K., Ramanathan V., Robicquet A., Fei-Fei L., Savarese S. Social LSTM: Human Trajectory Prediction in Crowded Spaces, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, Nevada, USA, June 27–30, 2016,* 2016, pp. 961–971. DOI: 10.1109/CVPR.2016.110.
15. Veličković P., Cucurull G., Casanova A., Romero A., Lio P., Bengio Y. Graph Attention Networks, *Proceedings of the International Conference on Learning Representations, Vancouver, Canada, April 30 – May 3, 2018*, 2017. DOI: 10.17863/CAM.48429.
16. Yu B., Yin H., Zhu Z. Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting, *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Stockholm, Sweden, July 13–19, 2018,* 2018, pp. 3634–3640. DOI: 10.24963/ijcai.2018/505.
17. Mi J., Zhang X., Zeng H., Wang L. DERGCN: Dynamic-Evolving graph convolutional networks for human trajectory prediction, Neurocomputing, 2024, Vol. 569, Article 127117. DOI: 10.1016/j.neucom.2023.127117.
18. Huang F., Fan Z., Li X., Zhang W., Li P., Geng Y., Zhu K. Tailored meta-learning for dual trajectory transformer: advancing generalized trajectory prediction, *Complex & Intelligent Systems*, 2025, Vol. 11, Article no. 174. DOI: 10.1007/s40747-025-01802-2.
19. Chen S. et al. Adaptive Graph Transformer for Human Trajectory Prediction, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, Washington, USA, June 17–22, 2024*, 2024, pp. 1617–1628.
20. Gupta A., Johnson J., Fei-Fei L., Savarese S., Alahi A. Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recogn*ition, *Salt Lake City, Utah, USA, June 18–22, 2018,* 2018, pp. 2255–2264. DOI: 10.1109/CVPR.2018.00240.
21. Sadeghian A., Kosaraju V., Sadeghian A., Hirose N., Rezatofighi H., Savarese S. SoPhie: An Attentive GAN for Predicting Paths Compliant to Social and Physical Constraints, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, California, USA, June 16–20, 2019,* 2019, pp. 1349–1358. DOI: 10.1109/CVPR.2019.00144.
22. Bilous N., Kozhevnikov A. Research of Methods for Determining the Accuracy of Metrological Measurements, *Technology Audit and Production Reserves,* 3(2(65)), 2022, pp. 18–23. DOI: 10.15588/1607-3274-2022-2-3
23. Bilous N., Tereshchenko I., Tereshchenko A., Bilous N., Shtangey S., Warsza Z. Risk Analysis Method by the Extreme Data of Dependent Exogenous Variables, *Journal of Automation, Mobile Robotics and Intelligent Systems*, 2022, pp. 44–53. DOI: 10.14313/JAMRIS/3-2021/18
24. Kosaraju V., Martin-Martin R., Reid I., Rezatofighi S., Savarese S. Social-BiGAT: Multimodal Trajectory Forecasting using Bicycle-GAN and Graph Attention Networks, *Proceedings of the Advances in Neural Information Processing Systems, Vancouver, Canada, December 8–14, 2019*. 2019, pp. 137–146. DOI: 10.5555/3454287.3454300.
25. Ruochen Li, Tanqiu Q., Stamos K., Zhanxing Z., Hubert S. Unified Spatial-Temporal Edge-Enhanced Graph Networks for Pedestrian Trajectory Prediction, *IEEE Transactions on Circuits and Systems for Video Technology*, 2025, pp. 1–14. DOI: 10.1109/TCSVT.2025.3539522.
26. Stanford Drone dataset, 2016, https://cvgl.stanford.edu/ projects/uav_data/

Received 22.09.2025.
Accepted 02.02.2026.
Published 27.03.2026.

УДК 004.93

МОДЕЛІ ГЛИБИННОГО НАВЧАННЯ ДЛЯ ПРОГНОЗУВАННЯ РУХУ ЛЮДИНИ У ВІДЕОПОТОКАХ

Білоус Н. В. – канд. техн. наук, професор кафедри програмної інженерії, Харківський національний університет радіоелектроніки, Харків, Україна. ROR: https://ror.org/01ctj1b90. ORCID: https://orcid.org/0000-0002-8850-9316.

Іванічев В. О. – аспірант кафедри програмної інженерії, Харківський національний університет радіоелектроніки, Харків, Україна. ROR: https://ror.org/01ctj1b90. ORCID: https://orcid.org/0009-0002-3705-0098.

АНОТАЦІЯ

Актуальність. Завдання точного прогнозування руху людини в середовищі є критично важливим для застосувань у системах моніторингу, пошуку та навігації. Існуючим підходам часто складно інтегрувати просторову та часову динаміку траєкторій під час обробки потокового відео в реальному часі.

Мета роботи. Розробити фреймворк на основі глибинного навчання, здатний прогнозувати рух людини шляхом поєднання ознак на рівні об’єктів і просторово-часової інформації про траєкторії, отриманої з відеопотоків.

Метод. Запропонований підхід інтегрує YOLO11 для детекції об’єктів, що дає змогу отримувати координати, швидкість, напрям руху та положення відносно оточення. Графова нейронна мережа моделює локальні й глобальні зв’язки між вузлами середовища, агрегуючи ознаки з урахуванням структури місцевості та перешкод. Просторово-часова увага виділяє найрелевантніші моменти траєкторії, підвищуючи точність передбачення. Модель обробляє послідовності кадрів із відеопотоків і в реальному часі прогнозує наступні позиції кожного відстежуваного об’єкта.

Результати. Експерименти на відеопослідовностях із різними сценаріями руху, довжинами траєкторій і варіаціями швидкості показали високу точність прогнозування. Запропонований метод ефективно поєднує просторові та часові ознаки й перевершує базові моделі в задачах трекінгу та передбачення руху.

Висновки. Отримані результати підтверджують придатність запропонованого фреймворку глибинного навчання для прогнозування руху людини в реальному часі у складних середовищах. Подальші дослідження можуть бути зосереджені на розширенні підходу до багатоагентних сценаріїв, оптимізації обчислювальної продуктивності та тестуванні на більших і різноманітніших наборах даних.

КЛЮЧОВІ СЛОВА: глибинне навчання, детекція об’єктів, траєкторія руху, прогнозування траєкторій людини, потокове відео, графові нейронні мережі, контекстно обізнане прогнозування руху, набір даних Stanford Drone, робота в реальному часі.

DOI 10.15588/1607-3274-2026-1-3

ЛІТЕРАТУРА

1. You Only Look Once: Unified, Real-Time Object Detection / [J. Redmon, S. Divvala, R. Girshick, A. Farhadi] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, Nevada, USA, June 27–30, 2016. – 2016. – P. 779–788. DOI: 10.48550/arXiv.1506.02640.
2. Simonyan K. Very Deep Convolutional Networks for Large-Scale Image Recognition / K. Simonyan, A. Zisserman // International Conference on Learning Representations. – 2015. – P. 1–14. DOI: 10.48550/arXiv.1409.1556.
3. Comparison of CNN-Based Architectures for Detection of Different Object Classes / [N. Bilous, V. Malko, M. Frohme, A. Nechyporenko] // Artificial Intelligence. – 2024. – Vol. 5, No. 4. – P. 2300–2320. DOI: 10.3390/ai5040113.
4. DETRs Beat YOLOs on Real-time Object Detection / [Y. Zhao, W. Lv, S. Xu et al.] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, Washington, USA, June 17–22, 2024. – 2024. – P. 16965–16974. DOI: 10.48550/arXiv.2304.08069.
5. Redmon J. YOLO9000: Better, Faster, Stronger / J. Redmon, A. Farhadi // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Hawaii, USA, July 21–26, 2017. – 2017. – P. 6517–6525. DOI: 10.1109/CVPR.2017.690.
6. Li Y. Improving real-time object detection in Internet-of-Things smart city traffic with YOLOv8-DSAF method / Y. Li, Y. Huang, Q. Tao // Scientific Reports. – 2024. – Vol. 14. – Article number: 17235. – 15 p. DOI: 10.1038/s41598-024-68115-1.
7. Wang C.-Y. Scaled-YOLOv4: Scaling Cross Stage Partial Network / C.-Y. Wang, A. Bochkovskiy, H.-Y. M. Liao // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Nashville, Tennessee, USA, June 20–25, 2021. – 2021. – P. 13029–13038. DOI: 10.1109/CVPR46437.2021.01283.
8. Jocher G. YOLOv8 : Overview / G. Jocher, A. Chaurasia, J. Qiu // Ultralytics Documentation, 2023. DOI: 10.5281/zenodo.3908559.
9. Bilous N. Search and Detection of People in the Water Using YOLO Architectures: A Comparative Analysis from YOLOv3 to YOLOv8 / N. Bilous. V. Malko, N. Moshenskyi // Automation 2024: Advances in Automation, Robotics and Measurement Techniques. AUTOMATION 2024. Lecture Notes in Networks and Systems, vol 1219. Springer, Cham. – P. 233–255. DOI: 10.1007/978-3-031-78266-4_21
10. A skeleton-based method for exercise recognition based on 3D coordinates of human joints / [N. Bilous, O. Svidin, I. Ahekian, V. Malko] // IAES International Journal of Artificial Intelligence (IJ–AI), ISSN/e-ISSN 2089-4872/2252-8938, 2024. – P. 1805–1816. DOI: 10.11591/ijai.v13.i2.pp1805-1816
11. Bilous N. Determination and Comparison Methods of Body Positions on Stream Video / N. Bilous, I. Ahekian, V. Kaluhin // Radio Electronics, Computer Science, Control. – 2023. – № 2. – P. 52–60. DOI: 10.15588/1607-3274-2023-2-6
12. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields / [Z. Cao, G. Hidalgo, T. Simon et al.] // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2019. – Vol. 43, № 1. – P. 172–186. DOI: 10.1109/TPAMI.2019.2929257.
13. 3D human pose estimation in video with temporal convolutions and semi-supervised training / [D. Pavllo, C. Feichtenhofer, D. Grangier, M. Auli] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, California, USA, June 16–20, 2019. – 2019. – P. 7753–7762. DOI: 10.48550/arXiv. 1811.11742.
14. Social LSTM: Human Trajectory Prediction in Crowded Spaces / [A. Alahi, K. Goel, V. Ramanathan et al.] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, Nevada, USA, June 27–30, 2016. – 2016. – P. 961–971. DOI: 10.1109/CVPR.2016.110.
15. Graph Attention Networks / [P. Veličković, G. Cucurull, A. Casanova et al.] // Proceedings of the International Conference on Learning Representations, Vancouver, Canada, April 30 – May 3, 2018. – 2017. DOI: 10.17863/CAM.48429.
16. Yu B. Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting / B. Yu, H. Yin, Z. Zhu // Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Stockholm, Sweden, July 13–19, 2018. – 2018. – P. 3634–3640. DOI: 10.24963/ijcai.2018/505.
17. DERGCN: Dynamic-Evolving graph convolutional networks for human trajectory prediction / [J. Mi, X. Zhang, H. Zeng, L. Wang] // Neurocomputing. – 2024. – Vol. 569. – Article 127117. DOI: 10.1016/j.neucom.2023.127117.
18. Tailored meta-learning for dual trajectory transformer: advancing generalized trajectory prediction / [F. Huang, Z. Fan, X. Li et al.] // Complex & Intelligent Systems. – 2025. – Vol. 11. – Article no. 174. DOI: 10.1007/s40747-025-01802-2.
19. Adaptive Graph Transformer for Human Trajectory Prediction / [S. Chen et al.] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, Washington, USA, June 17–22, 2024. – 2024. – P. 1617–1628.
20. Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks / [A. Gupta, J. Johnson, L. Fei-Fei et al.] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Utah, USA, June 18–22, 2018. – 2018. – P. 2255–2264. DOI: 10.1109/CVPR.2018.00240.
21. SoPhie: An Attentive GAN for Predicting Paths Compliant to Social and Physical Constraints / [A. Sadeghian, V. Kosaraju, A. Sadeghian et al.] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, California, USA, June 16–20, 2019. – 2019. – P. 1349–1358. DOI: 10.1109/CVPR.2019.00144.
22. Bilous N. Research of Methods for Determining the Accuracy of Metrological Measurements / N. Bilous, A. Kozhevnikov // Technology Audit and Production Reserves, 3(2(65)). – 2022. – P. 18–23. DOI: 10.15588/1607-3274-2022-2-3
23. Risk Analysis Method by the Extreme Data of Dependent Exogenous Variables / [N. Bilous, I. Tereshchenko, A. Tereshchenko et al.] // Journal of Automation, Mobile Robotics and Intelligent Systems. – 2022. – P. 44–53. DOI: 10.14313/JAMRIS/3-2021/18
24. Social-BiGAT: Multimodal Trajectory Forecasting using Bicycle-GAN and Graph Attention Networks / [V. Kosaraju, A. Sadeghian, R. Martin-Martin et al.] // Proceedings of the Advances in Neural Information Processing Systems, Vancouver, Canada, December 8–14, 2019. – 2019. – P. 137–146. DOI: 10.5555/3454287.3454300.
25. Unified Spatial-Temporal Edge-Enhanced Graph Networks for Pedestrian Trajectory Prediction / [Li Ruochen, Q. Tanqiu, K. Stamos et al.] // IEEE Transactions on Circuits and Systems for Video Technology. – 2025. – P. 1–14. DOI: 10.1109/TCSVT.2025.3539522.
26. Stanford Drone dataset, 2016, https://cvgl.stanford.edu/ projects/uav_data/

DOI 10.15588/1607-3274-2026-1-3