Deep learning method for UAV detection and flight path forecasting in real-time

Iurii V. Krak[1,2], Maksim O. Shurypa[3], Maryna O. Molchanova[3], Olexander V. Mazurets[3,*], Olena V. Sobko[3] and Olexander V. Barmak[3]

[1]Taras Shevchenko National University of Kyiv, Kyiv, 64/13 Volodymyrska Str., Kyiv, 01601, Ukraine
[2]Glushkov Institute of Cybernetics of NAS of Ukraine, 40 Glushkov Ave., Kyiv, 03187, Ukraine
[3]Khmelnytskyi National University, 11 Instytutska Str., Khmelnytskyi, 29016, Ukraine

Abstract

The paper is devoted to the creation and approbation of the neural network method for detecting and forecasting the trajectories of small UAVs in a video stream that integrates detection, tracking, and transformer prediction into a single high-performance pipeline. The architecture combines the YOLOv12 model for object localization, the BYTETrack algorithm for identity maintenance, and the modified position-coded Transformer for short-term motion forecasting based on coordinate history. The system generates continuous trajectories in conditions of a changing background, maneuvering, and partial overlaps and provides an average RMSE prediction error of 12.66 pixels, which is interpreted as approximately 0.25–0.32 m in the physical space of the scene. The obtained values of MOTA = 0.79 and IDF1 = 0.64 confirm the stability of tracking and preservation of object identity. On the Anti-UAV benchmark of 593,802 frames of real surveillance video, the YOLOv12 detector within the pipeline achieves mAP50 = 91.04% and mAP50–95 = 62.76% on the test split, running at 55 FPS on GPU and 13 FPS on CPU, while the BYTETrack-based tracker reaches 332 FPS on CPU. The experiments are carried out on the visible-spectrum subset of the Anti-UAV dataset and cover both offline processing of recorded video and online processing of live camera streams. Comparative analysis shows that the proposed approach significantly outperforms statistical models (IMM/Kalman), reducing the forecast error by an order of magnitude. The identified limitations relate to work in two-dimensional space with a fixed observation angle and a short forecasting horizon; extending the model to mobile platforms, multidimensional coordinates, and behavioral forecasting are identified as promising areas for further work. The results of the study confirm the effectiveness of combining transform forecasting with modern detection and tracking methods for building situational awareness systems, airspace monitoring, and countering unauthorized use of UAVs.

Keywords

UAV, trajectory forecasting, UAV detection, UAV tracking

1. Introduction

Modern UAVs are increasingly used for both civilian and military purposes, which necessitates the need for reliable systems for their automated detection and tracking [1, 2]. Traditional methods of airspace monitoring, in particular radar and acoustic systems, are not effective enough in cases of using small-sized, low-noise or low-altitude UAVs [3], especially in environments with a high level of interference [4]. At the same time, video analytics in combination with deep neural networks opens up new opportunities for building intelligent surveillance systems [5], capable not only of recognizing objects in difficult conditions, but also of predicting their trajectory in real time [6]. The problem of predicting UAV trajectories is critical for timely decision-making in systems of counteraction, support or prevention of potential threats [7]. However, most of the existing research focuses either on the problem

CS&SE@SW 2025: 8th Workshop for Young Scientists in Computer Science & Software Engineering, November 21, 2025, Kryvyi Rih, Ukraine
https://doi.org/10.55056/cssesw2025/paper24
*Corresponding author.
yuri.krak@gmail.com (I. V. Krak); shurypamo@khmnu.edu.ua (M. O. Shurypa); m.o.molchanova@gmail.com (M. O. Molchanova); exe.chong@gmail.com (O. V. Mazurets); olenasobko.ua@gmail.com (O. V. Sobko); alexander.barmak@gmail.com (O. V. Barmak)
0000-0002-8043-0785 (I. V. Krak); 0009-0003-7025-4647 (M. O. Shurypa); 0000-0001-9810-936X (M. O. Molchanova); 0000-0002-8900-0650 (O. V. Mazurets); 0000-0001-5371-5788 (O. V. Sobko); 0000-0003-0739-9678 (O. V. Barmak)

of detecting objects in static frames [8], or on tracking their position without taking into account the dynamic characteristics of the movement [9]. Methods that combine spatiotemporal processing of a video stream with trajectory forecasting based on the history of observations remain underdeveloped, particularly in cases of maneuvering or chaotic trajectories. Therefore, the development of a method for neural network detection and forecasting of UAV trajectories in a video stream is a scientifically and practically relevant task that will contribute to increasing the efficiency of situational awareness systems, critical infrastructure protection, tactical support, and autonomous navigation.

The aim of the work is to develop a method for forecasting the trajectories of small-sized UAVs in a video stream based on a modified position-coded transformer architecture taking into account the history of movement and spatial constraints of the scene, which provides a reduction in the average and final prediction error compared to statistical models and retains its suitability for integration into real-time monitoring systems.

The main contribution of the paper is the implemented modification of the transformer architecture for forecasting the trajectories of small-sized UAVs, which allows to preserve the model's ability to contextually match observations while reducing computational costs. The exclusion of a full-fledged decoder block and its replacement with a coordinate-oriented module with positional encoding ensured the constancy of time dependencies without losing spatial binding. Such a restructuring of the structure made it possible to reduce the dimensionality of intermediate representations and reduce latent fluctuations that usually arise when working with short and fragmented tracks. An additional result was an increase in the forecasting accuracy by using windowed coordinate aggregation with a fixed alignment horizon. This allowed to compensate for nonlinear changes in the trajectory and reduce the final error without involving complex recurrent components. The use of sinusoidal position vectors instead of embedding layers improved the model's resistance to noise, large-scale fluctuations, and loss of part of the frames, which is typical for real-time video streams.

2. Related works

Over the past three years, the issue of detecting and tracking small aerial objects in a video stream, in particular UAVs, has become one of the leading research areas in the field of computer vision. Scientists note that the main difficulties of this task are associated with small target sizes, significant variability of angles, overlaps and degradation of the video signal in real conditions [10]. In response to this, in the modern literature there is a tendency to adapt convolutional architectures, primarily the YOLO family, to the specifics of drone datasets (VisDrone, UAVDT, etc.), which is confirmed by stable improvements in the detection of small objects in variable scenes [11]. In the work, the YOLOv8 architecture is modified by enhanced feature fusion at different scales, which allowed to improve the accuracy on the VisDrone2019 set. The authors emphasize that the largest increase in mAP is observed precisely for small-area objects, due to the optimization of FPN blocks.

In [12], an alternative solution is proposed – SOD-YOLO, where the focus is on adaptive weighted scale compensation. Unlike [11], this model is focused not only on improving the transmission of small features, but also on suppressing background noise, which reduced the number of false positives on moving textures.

Another approach is presented by Yao et al. [13], where local contour enhancement is applied in the early layers of the neural network. The authors prove that the preliminary selection of high-frequency components allows to avoid "blurring" of objects when convolutions of a large kernel, which is a typical problem in videos shot from long distances.

Despite significant progress in detection, for practical application it is critical to preserve the identity and continuity of trajectories in time. Studies show that associative multi-object trackers, such as ByteTrack and DeepSORT, combined with state-of-the-art detectors, provide a significant reduction in ID-switch even under conditions of partial overlaps and changes in perspective [14, 15], as well as in UAV domain work, where the combination of detection and Re-ID improves resistance to uneven motion and video signal degradation [16].

A separate layer of publications focuses on UAV trajectory forecasting as the next step after tracking. Along with classical statistical approaches (Kalman/IMM), sequential neural network models – LSTM and Transformer [17] are increasingly used. It has been shown that taking into account the trajectory history, contextual features of the scene and behavioral intentions of the target reduces prediction errors and increases the stability of estimating future positions in a short prediction horizon [18], and for related transport scenarios – and transformer trajectory predictors taking into account complex urban dynamics [19]. At the same time, many works emphasize insufficient attention to estimating forecast uncertainty and robustness to observation losses – these aspects remain open and identify promising areas for further research.

In [20], the authors investigate the problem of recognizing drone types to improve airspace safety. To solve this problem, they used a lightweight Yolov7 model, which integrated the CNeB, C3C2 modules and the Re-Parameterization Decoupled (RePD) framework. This approach allowed to reduce the size of the model, increase the accuracy and speed of real-time image processing. The results showed that the proposed Yolov7-CNeB-C3C2-RePD model achieves 91.9% Precision, 90.5% Recall and 95% mAP@0.5, outperforming other algorithms in test comparisons, while providing processing of 27.88 frames per second, which meets real-time requirements. Xie et al. [21] investigate the detection of UAVs in real time. To do this, they used the YOLOv5s model trained on five object categories, with mAP50 = 93.2%, mAP50-95 = 71.7%, size 13.76MB and detection speed 11.26ms per image. The model is integrated into a modified DeepStream for simultaneous use by multiple users and fast recovery of the video stream after interruptions, providing efficient automatic real-time UAV inspection.

In [22], drone detection using surveillance cameras is investigated, focusing on small, faint and low-contrast objects that are difficult to detect. To do this, they proposed a method based on YOLOv5s that uses frame sequences and inter-frame optical flow, combining spatial and temporal information. This approach improves the accuracy and reliability of detecting moving small targets. Experiments showed an average Precision value of 86.87% (an improvement of 11.49% compared to the baseline model) at a speed of more than 30 frames per second, and the method can be adapted to other detectors with different backbone models.

All the considered works demonstrate that modern methods based on YOLO models and their modifications effectively solve the problem of drone detection in real time, even under difficult conditions – high or low altitude, small object sizes, weak contrast and motion.

3. Methodology

3.1. Improved approach to UAV detection and flight path forecasting

The generalized scheme of the approach to predicting UAV trajectories based on a position-coded transformer is shown in figure 1. The proposed approach to predicting UAV motion predicts the next 60 positions of the object based on the previous 30 coordinates, which corresponds to a forecasting horizon of 2 seconds at a frequency of 30 frames per second.

This format allows using only one second of motion history to obtain a tactical forward forecasting

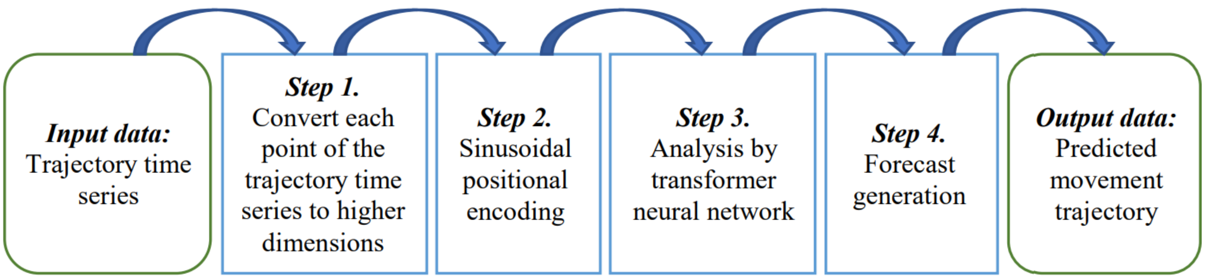

Figure 1: Generalized scheme of the approach to forecasting UAV trajectories.

suitable for real-time monitoring systems. Unlike classical regression or statistical models, the developed architecture does not work with derived quantities such as velocity or acceleration, but directly with the coordinate trajectory, interpreting it as a time sequence of spatial observations. In order to provide the ability to generalize nonlinear motion patterns, each point is transformed from a two-dimensional representation into a vector of hidden features of higher dimension (Step 1). Since the transformer-type model does not operate with the concept of default order [23], a sinusoidal positional encoding is added to the vector representations of coordinates (Step 2), which provides a time reference to each element of the sequence. The central element of the system is the self-attention block, which analyzes the relationships between all trajectory points simultaneously (Step 3), detecting both local changes in direction and longer-term patterns of motion. After processing by the encoder, all generated features are aggregated into a single context vector, which is used as a compact representation of the current state of the object. Based on this vector, a multilayer perceptron (Step 4) generates a forecast of the next coordinates, ensuring a smooth continuation of the trajectory taking into account the previous movement.

Thus, the proposed approach allows for short-term forecasting of UAV trajectories in real time, ensuring smooth prediction even in the presence of nonlinear changes in motion. The use of positional encoding and the self-attention mechanism increases the model's generalization ability and makes it suitable for integration into tactical monitoring and response systems.

3.2. Pipeline for UAV detection and flight path forecasting in real-time

The pipeline based on the neural network method of UAV detection and trajectory forecasting is implemented as a sequential pipeline that combines detection, tracking, forecasting, and result visualization modules (figure 2). The input to the system can be a video file, a link to an online stream, or a live broadcast from a camera. Each frame coming from the source is passed to the YOLOv12 module [24], which performs drone detection, returning the coordinates of the bounding boxes and the probabilities of belonging to the target class.

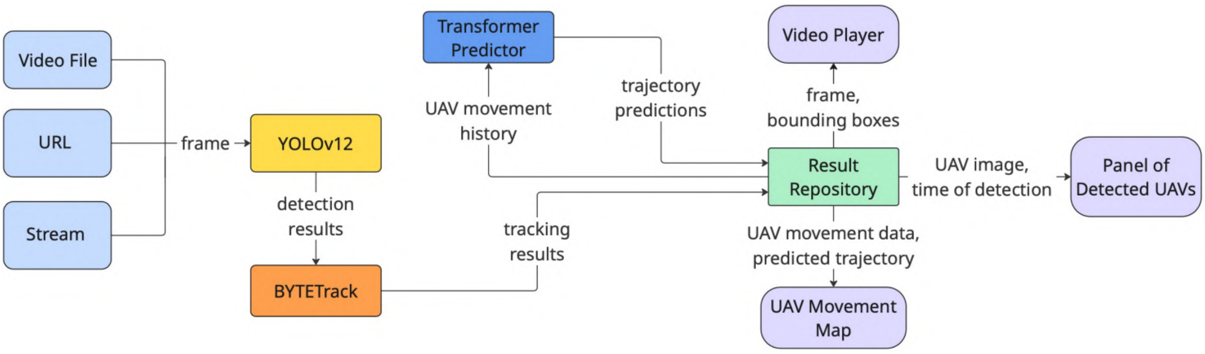

Figure 2: Pipeline of UAV detection and flight path forecasting in real-time.

The received detections are transmitted to BYTETrack, which performs associative tracking of objects in a time sequence of frames, forming permanent track identifiers and the history of the movement of each UAV. For each formed track, a sequence of coordinates is accumulated, which is used as an input for the Transformer Predictor prediction module. The transformer model generates a predicted trajectory of movement, taking into account both the current position of the drone and its historical movement.

Both the actual coordinates and the predicted trajectory points are transmitted to the results storage, which acts as a central data buffer. It accumulates a photo of the drone, the time of detection, the geometry of the trajectory and identification information for further reproduction.

Data from the storage is transmitted in parallel to two output components: a video player, which displays the original video frame with superimposed bounding boxes and a drone motion map with the actual/predicted trajectory, where the spatial movements of the object are visualized on a coordinate grid or cartographic substrate.

Additionally, a panel of found drones is provided, which accumulates a list of active objects, their characteristics and the time of the last observation, which allows the operator to monitor in real time or analyze the history of events.

Thus, the above pipeline provides a full cycle of video stream processing from receiving raw data to forming predicted trajectories and their visual visualization. The integration of detection, tracking and transformer forecasting modules into a common architecture allows not only to record the presence of the UAV, but also to predict its further movement, which increases the applied value of the information system in monitoring and operational response tasks.

3.3. Modified position-coded transformer architecture

The proposed trajectory forecasting model (figure 3) is based on a modified transformer architecture [25], adapted to work with coordinate sequences, and not with discrete tokens, as in typical NLP solutions [26]. The input data is a sequence of thirty two-dimensional points, reflecting the previous dynamics of the UAV movement. Since the coordinates in their original form do not contain sufficient representativeness to detect latent dependencies, each element of the sequence is projected into a high-dimensional space using a linear transformation, which ensures the formation of consistent features for further processing.

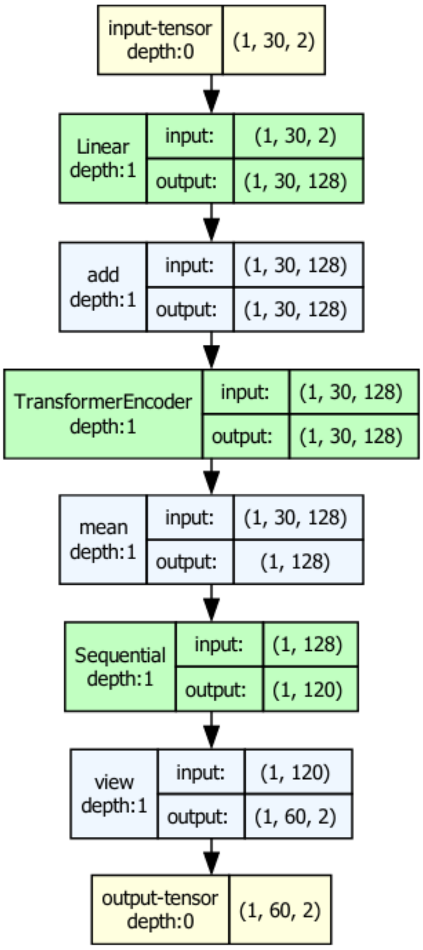

Figure 3: Generalized modified neural network architecture.

Temporal ordering of the sequence is ensured by the introduction of sinusoidal positional encoding, thanks to which the network stores information about the temporal context of each point without additional recurrent components. The basic computational core is a three-layer Transformer Encoder [27], which combines a multi-head self-attention mechanism with fully connected projections and regularization. This configuration provides simultaneous consideration of local changes and long-term trends, which reduces the sensitivity of the model to shunting distortions or partial track losses. After processing by the encoder, a latent representation is formed in the form of a three-dimensional tensor, which is reduced to a context vector by averaging over the time axis. The rejection of a full-scale decoder block in favor of such aggregation reduces computational complexity and reduces processing latency, which is critical for operation in real-time modes. The final transformation of coordinates is performed by a multilayer perceptron with normalization and nonlinear activation, after which a vector of the forecasting of the next positions is formed, which reproduces the future trajectory of the object.

The coordinated interaction of the transformer encoder with coordinate-oriented decoding ensures the preservation of the spatio-temporal structure of the data without the need to calculate the derivative features of the movement. The combination of positional encoding with a compact latent representation allows achieving a balance between accuracy and speed, which makes the architecture suitable for implementation in airspace monitoring and operational response systems.

Thus, the improved architecture based on the positional-coded transformer provides full processing of coordinate sequences without the need for intermediate calculation of the derived parameters of the movement. Context aggregation through a simplified encoder allows maintaining high predictive ability with minimal computational delay, which makes the approach compatible with the requirements of real-time systems.

4. Experiment

4.1. Experimental dataset

The «Anti-Unmanned Aerial Vehicle» dataset [28] is designed to study the detection and tracking of unmanned aerial vehicles in a video stream and contains recordings obtained in real-world surveillance conditions. It covers a wide range of scenes with various types of backgrounds, lighting levels, and shooting angles, which ensures representativeness for practical airspace monitoring scenarios. Each frame in the dataset has an accurate annotation in the form of bounding boxes indicating the position of the drones, which allows this dataset to be used for both detection and tracking. A feature of the dataset is the presence of recordings in both the visible and infrared spectra, however, within the framework of this work, only a subset of videos in the visible range is used.

Structurally, the dataset is divided into training, validation, and test parts, containing 160, 67, and 91 video recordings, respectively – a total of 318 video fragments, which together number 593,802 frames. Such a number of observations provides the possibility of full-fledged training of deep models without additional artificial expansion of the sample [29].

An important property of the dataset is the purposeful presence of complex observation scenarios in it, in particular cases of temporary disappearance of the object from the field of view, changes in scale and illumination, partial or complete overlaps, low target resolution and rapid movements, leading to significant inter-frame shifts. This creates conditions close to the real deployment of monitoring systems and makes it possible to assess the stability of algorithms not only by average accuracy indicators, but also in critical situations where typical tracking loses its effectiveness. Thus, the «Anti-UAV Dataset» can be considered as a full-fledged base for training and objective testing of methods for detecting, tracking, and predicting UAV trajectories.

4.2. Experimental setup

The experimental setup is formed as an integrated software and hardware complex, in which each module performs interconnected functions within a single video stream processing pipeline. The primary data capture is carried out in the format of a sequence of frames that are fed to the detection system in real time. The YOLOv12 model is used to identify small-sized aerial objects, the training of which is focused on maintaining accuracy under conditions of a changing background and a moving camera. Detections are automatically transferred to the association module, where the BYTETrack algorithm is implemented, capable of maintaining the continuity of the object's identity even with partial losses, temporary overlap or fragmentation of the trajectory.

The resulting coordinate sequences are converted into a format compatible with the transformer forecasting model. The forecasting architecture is built in such a way as to provide short-term trajectory approximation with minimal time delay and without intermediate extraction of operational motion parameters. The data flow between the detector, tracker, and prediction module is coordinated through an intermediate cache, which minimizes frame synchronization losses and allows for the reuse of previous calculations. The experimental setup is shown in figure 4.

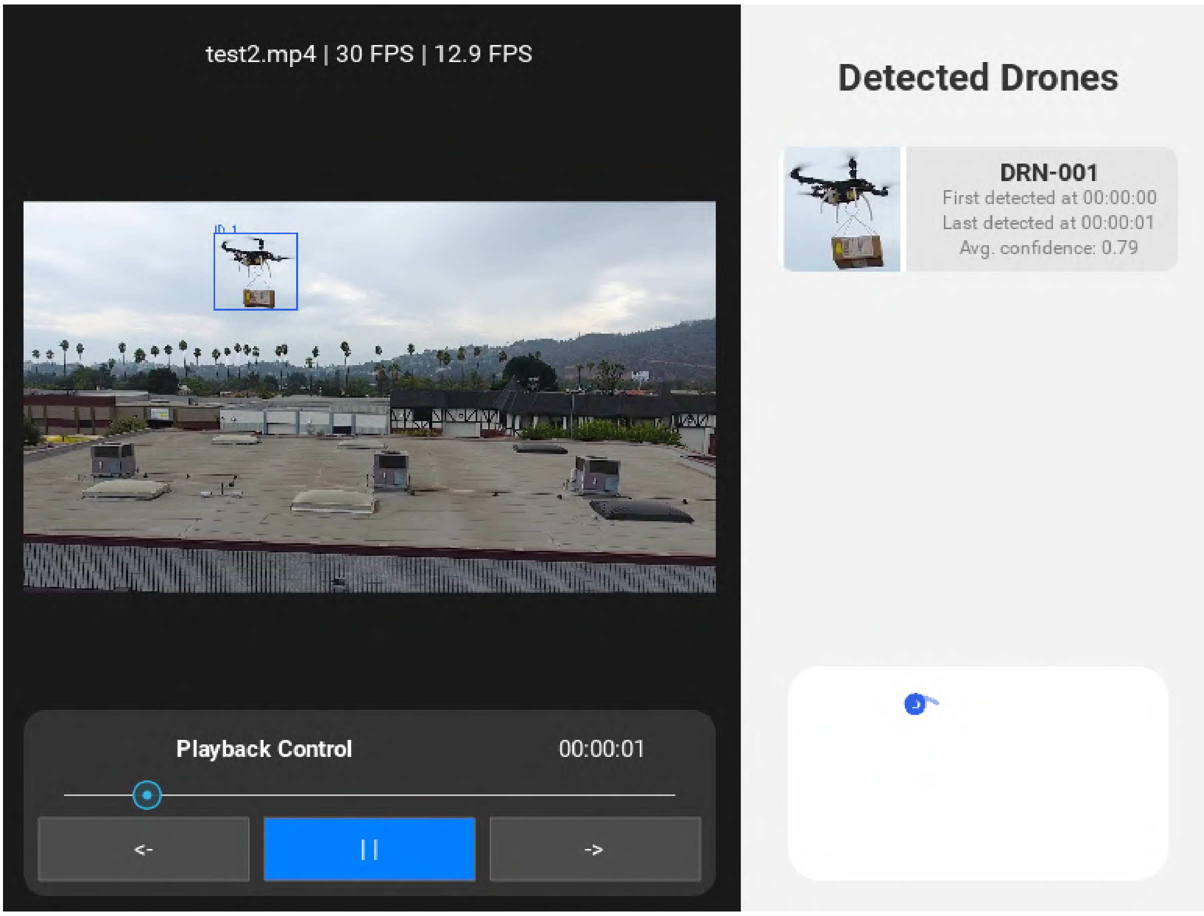

Figure 4: Experimental application interface.

For visualization and control of the interaction of components, a graphical environment based on Kivy [30] is used, which provides rendering of actual and predicted trajectories, as well as control over the operation of each element of the system without interrupting the main process. Such an organization allows at the same time to evaluate performance, interpret errors and adapt parameters in a mode close to real time. The installation is built taking into account the possibility of further deployment on embedded systems or mobile computing platforms.

5. Results and discussion

Evaluation of the proposed detection approach based on YOLOv12 showed that the model is able to consistently recognize small aerial objects even in the presence of noise, scale changes, and a variable background. The high value of mAP50 on the test set indicates accurate positioning of the bounding boxes, while the mAP50-95 ratio demonstrates that the model retains adequate sensitivity even with increased matching criteria (table 1). Robustness to small targets is confirmed by the fact that the number of false negatives does not increase significantly in cases where the bounding box area does not exceed several hundred pixels.

Table 1
Drone detection metrics by the YOLOv12 model.

mAP50 (%)			mAP50-95 (%)	FPS (CPU)	FPS (GPU)	GFLOPs
training	validation	test				
99.05	94.65	91.04	62.76	13	55	6.3

The classification results matrix (figure 5) reflects a typical situation for object detection tasks in which the background is not considered as a separate class. Because the potential variability of background structures in a single image is too high for detailed labeling, insignificant parts of the scene are traditionally not annotated, but only present in the set as context. Accordingly, cases that formally look like false background classifications to the "drone" class do not necessarily indicate a false response of the model.

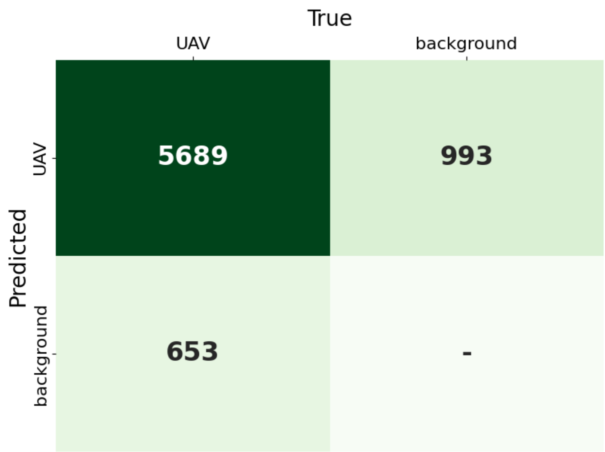

Figure 5: Confusion matrix of the YOLOv12 model on the test set.

A significant portion of such observations correspond to objects with similar structural or silhouette features, in particular aircraft or drone-like structures, which in real conditions may also be of interest to the monitoring system. Examples of such false positives are shown in figure 6.

Under these conditions, the behavior of the model demonstrates its sensitivity to potentially relevant targets rather than random error generation. Further refinement of the boundaries between target and secondary objects can be achieved by including a separate group of "adjacent" classes in the training set, but even in its current form, the model exhibits a balanced ability to generalize without a tendency to systematically overtrain on specific background patterns.

The mAP50 indicator (figure 7) demonstrates a stable growth during the first tens of training epochs. At the initial stage, a sharp increase in accuracy is observed, after which the curve gradually levels off. Around 30–40 epochs, the metric value reaches a plateau and subsequently changes only with minor fluctuations. This indicates that the model achieves stable detection quality without signs of overtraining, and further training has practically no effect on the result.

The diagram of the loss function value change shows a rapid decrease in the error at the initial stages of training (figure 8).

After the curve stabilizes on the training sample, a gradual smoothing of fluctuations on the validation sample is observed, which indicates the achievement of a balanced operating mode of the model without obvious signs of overtraining. A slight difference between the loss-sample curves indicates the variability

Figure 6: Examples of false alarms.

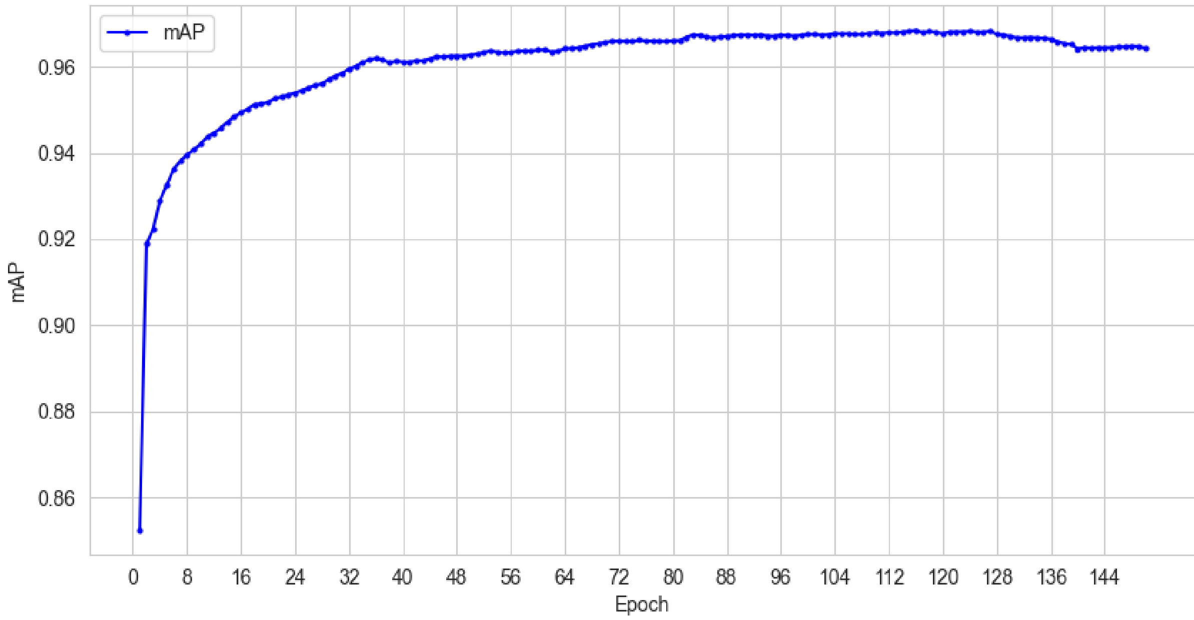

Figure 7: Diagram of the mAP50 metric change on the validation sample by epochs.

of visual conditions inherent in the task, but does not significantly affect the overall tracking quality.

The next stage is to study the possibilities of tracking detected objects. The results of evaluating the performance of the BYTETrack tracker on the test sample demonstrate a sufficiently high consistency between detections and formed trajectories. The MOTA value = 0.79 indicates that most objects retain their identity during observation even in the presence of partial overlaps or changes in scale (table 2).

The IDF1 indicator = 0.64 confirms a moderate level of accuracy of comparison between detections

Table 2
Metrics of tracking drones of the test sample by the BYTETrack algorithm.

MOTA	FIDF1	FPS (CPU)
0.79	0.64	332

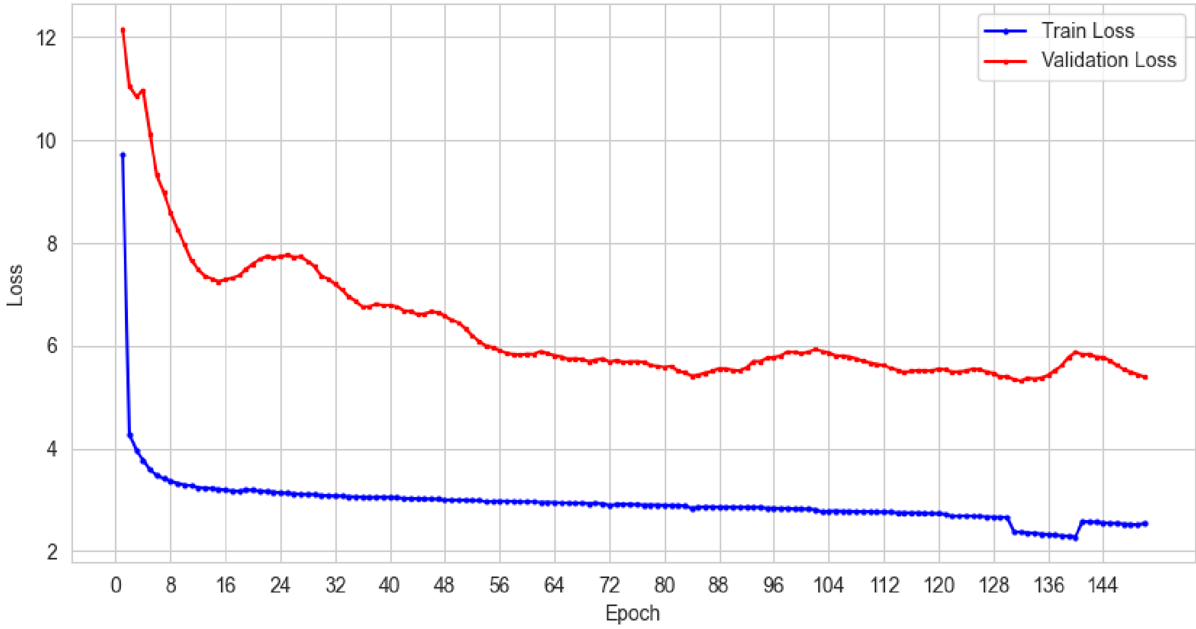

Figure 8: Diagram of the change in the value of the loss function on the training and validation samples by epochs.

and existing tracks, which is typical for scenarios with uneven speed of movement. High performance (332 frames per second on CPU when running on an Intel i5 processor, without video playback, the tracker receives bounding box data and frame number from the set) gives reason to believe that the selected algorithm configuration can be used in real-time systems without additional optimization.

Figure 9 shows the curve of training and validation losses during 200 epochs of training. A rapid decrease in the value of the loss function is observed at the initial stages of training (0–20 epochs), after which the dynamics gradually stabilizes. The training sample demonstrates a monotonic decrease in the error up to a value of ≈ 0.05, which indicates effective learning of patterns in the trajectories. The validation curve after the initial decrease remains stable within 0.18–0.22, without significant fluctuations or sharp jumps, which confirms the absence of overtraining. The early stopping mechanism was not activated prematurely, which means stable dynamics of improvements up to the final epochs. It should be noted that the trend of loss convergence is interpreted in the normalized coordinate space: all positional values were previously scaled to the interval [0,1] relative to the width and height of the image, which provides a unified measurement system and allows comparing the dynamics of learning regardless of the real scale of the scene. That is why the error values in the range of 0.05–0.20 at this stage correspond to small deviations in the normalized coordinates, while in the following sections, where the results are presented in pixels, the absolute values may look significantly larger, although they actually describe the same level of accuracy.

The summary metrics, shown in table 3, show an average root mean square error (RMSE) of 12.66 pixels, which corresponds to an accuracy level of ~1–2% of the typical frame width. The metrics ADE = 14.1 and FDE = 17.4 pixels indicate that the model is able to not only generally correctly estimate the direction of motion, but also predict the final position quite accurately in a 2-second horizon. The computational complexity of the model is only 0.03 GFLOPs, making it suitable for deployment on resource-constrained devices and integration into real-time systems.

Table 3
Transformer metrics in pixels on the test set.

RMSE	ADE	FDE	GFLOPs
12.66	14.1	17.4	0.03

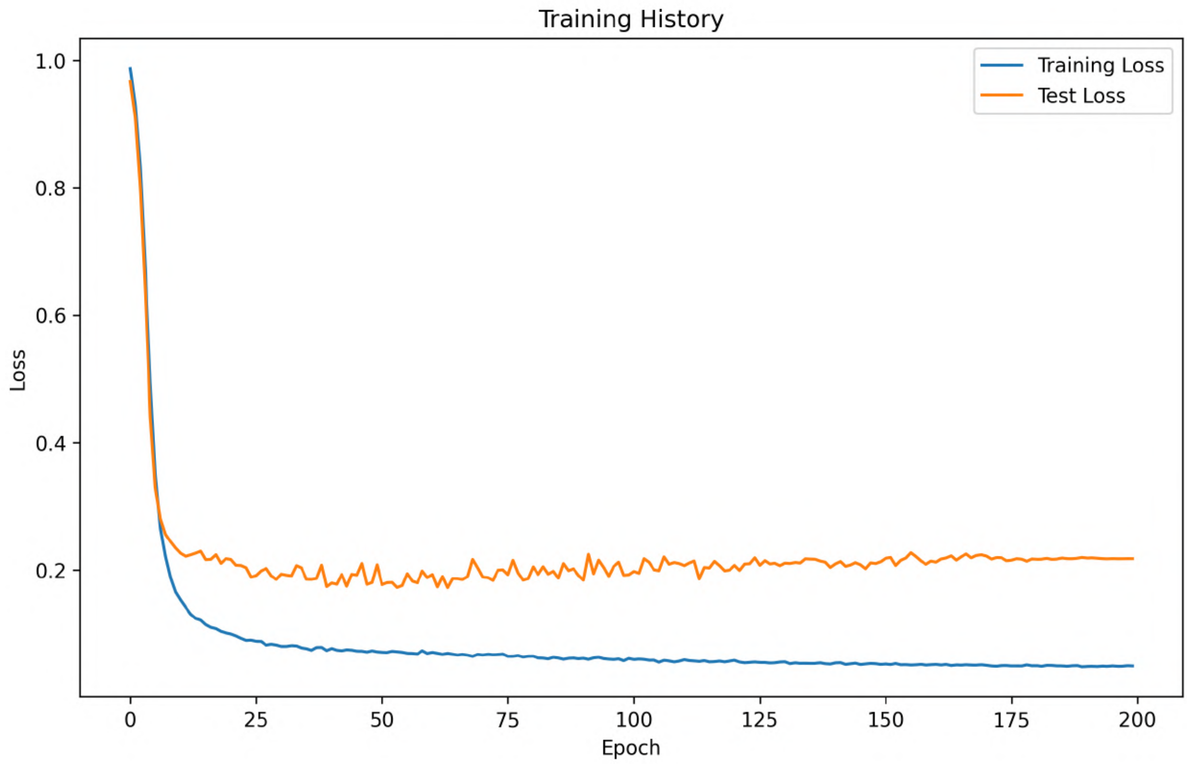

Figure 9: Transformer training diagram.

Thus, the results obtained confirm that the modified transformer architecture is capable of stable short-term forecasting of UAV trajectories, maintaining a balance between accuracy, stability and performance. Compared with the results of classical statistical filters (IMM/Kalman), which are presented in table 4, the model demonstrates significantly lower error in maneuvering trajectories, which makes it promising for practical use in situational awareness systems.

Table 4
IMM/Kalman metrics in pixels on the test set.

RMSE	ADE	FDE
126.27	161.1	250.95

The information system integrates a trajectory forecasting module that uses the transformer to predict the drone trajectory 60 frames ahead. The result is displayed on the drone motion map. Examples of the approach to forecasting UAV trajectory are shown in figure 10.

Visually, it can be seen that the model is able to reproduce both translational and maneuvering types of movement with a slight deviation from the real trajectory, preserving both the general direction and the local shape of the curve. In most cases, the discrepancy between the predicted and actual path grows smoothly, which indicates the stable behavior of the model even under conditions of sharp changes in direction.

As a result of comparison with known analogues, the proposed method demonstrates higher accuracy in terms of the MOTA metric than most modern solutions for UAV tracking. For example, in the conducted study, the value of the MOTA metric was obtained = 0.79, while in [18] MOTA = 0.461 is reported.

For correct comparison with works in which the prediction error is given in meters, the difference in coordinate representation formats should be taken into account. In the presented experiment, trajectories were estimated in pixels normalized to the frame width, whereas in works like [19] the

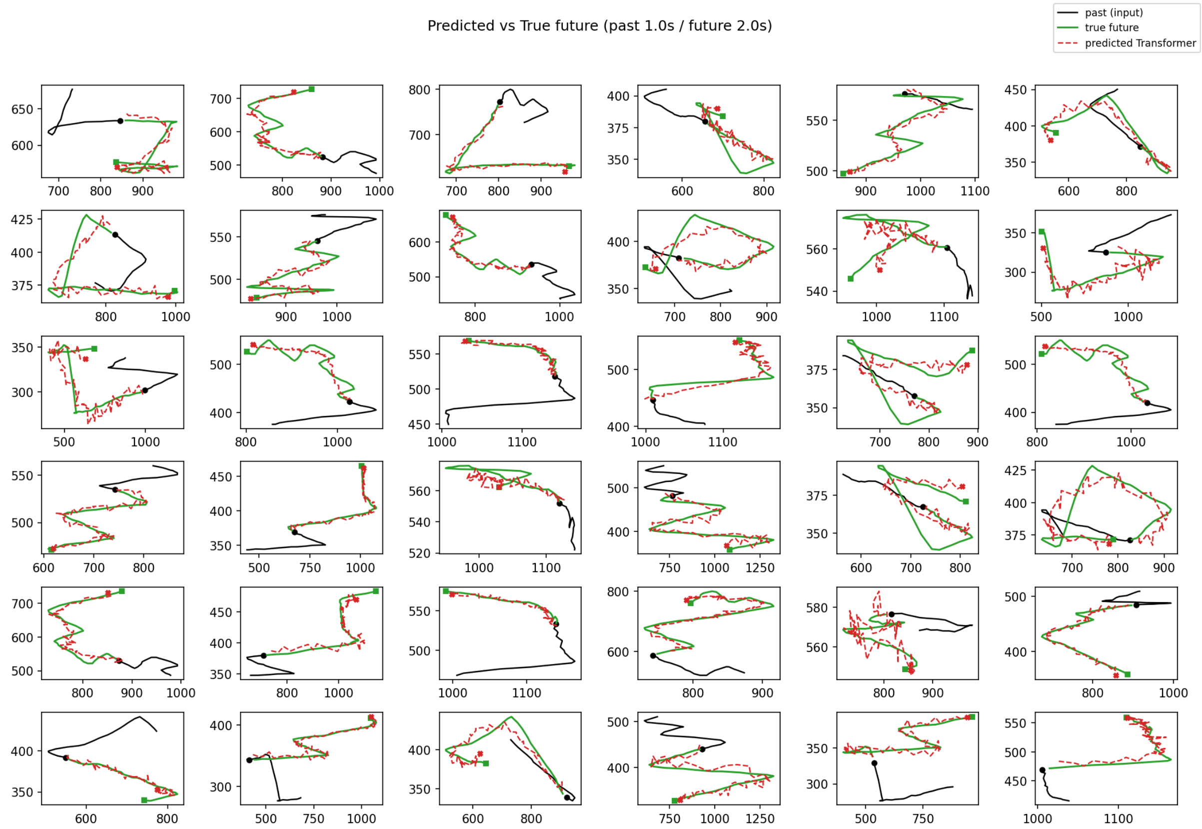

Figure 10: Examples of the approach to forecasting UAV movement trajectories.

error is measured in physical space. At a typical scene scale in the Anti-UAV set, one meter in the observation plane corresponds to approximately 40-60 pixels (depending on the flight altitude and viewing angle). Under this assumption, the RMSE value = 12.66 pixels is interpreted as approximately 0.25-0.32 m, which is comparable to the error of 0.214-0.41 m given in [19]. This indicates that even with different estimation formats (image vs. sensor coordinates), the prediction accuracy is within the range of state-of-the-art models adapted to physical space.

An additional advantage is the high performance of 332 FPS on CPU, which significantly exceeds typical performance in open multi-object tracking implementations, where performance rarely exceeds 30-50 FPS. This allows us to classify the method as suitable for deployment in real-time conditions, in particular in aerial monitoring or UAV countermeasure systems.

6. Conclusions

The paper develops a method for neural network detection and forecasting of UAV trajectories in a video stream, which combines detection, tracking, and transformer forecasting in a single processing pipeline. The constructed system is capable of forming continuous trajectories of small-sized objects in difficult observation conditions and generating a short-term forecast with an RMSE error of 12.66 pixels (about the range of 0.25-0.32 m in physical space). To increase the accuracy of the system, the use of deep forecasting architectures of the Transformer type is envisaged, which provides adaptability to different types of trajectories.

The experiments conducted showed that the proposed approach provides the values of the metrics MOTA = 0.79 and IDF1 = 0.64, which exceeds the published indicators of most modern solutions based on statistical models and brings the system closer to the level suitable for real-time deployment. The speed of 332 FPS on the CPU confirms the possibility of integrating the method into practical monitoring systems.

Despite the achieved results, the proposed approach has certain experimental limitations inherent in most works of this class. The forecasting is carried out in two-dimensional image coordinates, which corresponds to the conditions of stationary observation from a fixed camera, therefore, scenarios with active platform movement require additional adaptation. The work considers only a short-term forecast horizon, such an approach is justified for operational monitoring systems, and increasing the forecast interval will require a combination with behavioral models or probabilistic uncertainty estimates. A separate direction of development is the adaptation of the model to different types of dynamics from translational motion to complex maneuvering modes, which can be implemented through preliminary classification of trajectories or multi-model forecasting. These limitations will be addressed in future studies.

Author contributions

Conceptualization, Iurii V. Krak and Olexander V. Barmak; methodology, Maksim O. Shurypa and Maryna O. Molchanova; formulation of tasks analysis, Olexander V. Mazurets; software, Maryna O. Molchanova and Olena V. Sobko; writing–original draft, Maksim O. Shurypa and Maryna O. Molchanova; analysis of results, Olena V. Sobko and Olexander V. Mazurets; visualization, Maksim O. Shurypa; reviewing and editing, Olexander V. Mazurets. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Data availability statement

No new data were created or analysed during this study. Data sharing is not applicable.

Conflicts of interest

The authors declare no conflict of interest.

Acknowledgments

The authors are grateful Program Committee members for organizing and conducting the 8th Workshop for Young Scientists in Computer Science & Software Engineering (CS&SE@SW 2025).

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] P. Singh, K. Gupta, A. K. Jain, Vishakha, A. Jain, A. Jain, Vision-based UAV Detection in Complex Backgrounds and Rainy Conditions, in: 2024 2nd International Conference on Disruptive Technologies (ICDT), 2024, pp. 1097–1102. doi:10.1109/ICDT61202.2024.10489147.

[2] A. AlKhonaini, T. Sheltami, A. Mahmoud, M. Imam, UAV Detection Using Reinforcement Learning, Sensors 24 (2024) 1870. doi:10.3390/s24061870.

[3] O. Siechko, A. Kitsera, R. Kochan, J. Józwik, A. Tofil, D. Mika, Acoustic Signal Analysis of the Targets of Distributed Airspace Monitoring System, in: 2024 11th International Workshop on Metrology for AeroSpace (MetroAeroSpace), IEEE, 2024, pp. 52–56. doi:10.1109/MetroAeroSpace61015.2024.10591580.
[4] M. Alqaraleh, M. S. Alzboon, M. S. Al-Batah, Skywatch: Advanced Machine Learning Techniques for Distinguishing UAVs from Birds in Airspace Security, International Journal of Advanced Computer Science and Applications 15 (2024). doi:10.14569/IJACSA.2024.01511104.
[5] K. Rezaee, S. M. Rezakhani, M. R. Khosravi, M. K. Moghimi, A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance, Personal and Ubiquitous Computing 28 (2024) 135–151. doi:10.1007/s00779-021-01586-5.
[6] Y. Tang, H. He, Y. Wang, Hierarchical vector transformer vehicle trajectories prediction with diffusion convolutional neural networks, Neurocomputing 580 (2024) 127526. doi:10.1016/j.neucom.2024.127526.
[7] P. Shukla, S. Shukla, A. Kumar Singh, Trajectory-Prediction Techniques for Unmanned Aerial Vehicles (UAVs): A Comprehensive Survey, IEEE Communications Surveys & Tutorials 27 (2025) 1867–1910. doi:10.1109/COMST.2024.3471671.
[8] M. Molchanova, V. Didur, O. Mazurets, O. Sobko, O. Zakharkevich, Method for Construction and Demolition Waste Classification Using Two-Factor Neural Network Image Analysis, in: N. Shakhovska, A. T. Augousti, S. Liaskovska, O. Duran (Eds.), Proceedings of the 2nd International Conference on Smart Automation & Robotics for Future Industry (SMARTINDUSTRY 2025), Lviv, Ukraine, April 3-5, 2025, volume 3970 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 168–182. URL: https://ceur-ws.org/Vol-3970/PAPER14.pdf.
[9] L. Wang, W. Huang, H. Li, W. Li, J. Chen, W. Wu, A Review of Collaborative Trajectory Planning for Multiple Unmanned Aerial Vehicles, Processes 12 (2024) 1272. doi:10.3390/pr12061272.
[10] L. Xu, Z. Luo, Anti-UAV detection and identification technology: Fundamentals, methods and challenges, Physical Communication 71 (2025) 102676. doi:10.1016/j.phycom.2025.102676.
[11] J. Ni, S. Zhu, G. Tang, C. Ke, T. Wang, A Small-Object Detection Model Based on Improved YOLOv8s for UAV Image Scenarios, Remote Sensing 16 (2024) 2465. doi:10.3390/rs16132465.
[12] Y. Li, Q. Li, J. Pan, Y. Zhou, H. Zhu, H. Wei, C. Liu, SOD-YOLO: Small-Object-Detection Algorithm Based on Improved YOLOv8 for UAV Images, Remote Sensing 16 (2024) 3057. doi:10.3390/rs16163057.
[13] B. Yao, C. Zhang, Q. Meng, X. Sun, X. Hu, L. Wang, X. Li, SRM-YOLO for Small Object Detection in Remote Sensing Images, Remote Sensing 17 (2025) 2099. doi:10.3390/rs17122099.
[14] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, X. Wang, ByteTrack: Multi-object Tracking by Associating Every Detection Box, in: S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, T. Hassner (Eds.), Computer Vision – ECCV 2022, volume 13682 of *Lecture Notes in Computer Science*, Springer Nature Switzerland, Cham, 2022, pp. 1–21. doi:10.1007/978-3-031-20047-2_1.
[15] N. Wojke, A. Bewley, D. Paulus, Simple online and realtime tracking with a deep association metric, in: 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 2017, pp. 17–20 September 2017. doi:10.1109/icip.2017.8296962.
[16] X. Li, R. Zhu, X. Yu, X. Wang, High-Performance Detection-Based Tracker for Multiple Object Tracking in UAVs, Drones 7 (2023) 681. doi:10.3390/drones7110681.
[17] J. K. Lee, Transformer-Based Vehicle-Trajectory Prediction at Urban Low-Speed T-Intersection, Sensors 25 (2025) 4256. doi:10.3390/s25144256.
[18] Y. Yuan, Y. Wu, L. Zhao, Y. Pang, Y. Liu, Multiple Object Tracking in Drone Aerial Videos by a Holistic Transformer and Multiple Feature Trajectory Matching Pattern, Drones 8 (2024) 349. doi:10.3390/drones8080349.
[19] Z. Cheng, J. Yang, J. Sun, L. Zhao, Trajectory Planning of Unmanned Aerial Vehicles in Complex Environments Based on Intelligent Algorithm, Drones 9 (2025) 468. doi:10.3390/drones9070468.
[20] Y. Ghazlane, M. Hicham, et al., Real-time lightweight drone detection model: Fine-grained Identification of four types of drones based on an improved Yolov7 model, Neurocomputing 596 (2024) 127941. doi:10.1016/j.neucom.2024.127941.

[21] S. Xie, G. Deng, B. Lin, W. Jing, Y. Li, X. Zhao, Real-Time Object Detection from UAV Inspection Videos by Combining YOLOv5s and DeepStream, Sensors 24 (2024) 3862. doi:10.3390/s24123862.
[22] Y. Sun, X. Zhi, H. Han, S. Jiang, T. Shi, J. Gong, W. Zhang, Enhancing UAV Detection in Surveillance Camera Videos through Spatiotemporal Information and Optical Flow, Sensors 23 (2023) 6037. doi:10.3390/s23136037.
[23] I. Krak, O. Sobko, M. Molchanova, I. Tymofiiev, O. Mazurets, O. Barmak, Method for neural network cyberbullying detection in text content with visual analytic, in: S. O. Semerikov, A. M. Striuk (Eds.), Proceedings of the 7th Workshop for Young Scientists in Computer Science & Software Engineering (CS&SE@SW 2024), Virtual Event, Kryvyi Rih, Ukraine, December 27, 2024, volume 3917 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 298–309. URL: https://ceur-ws.org/Vol-3917/paper57.pdf.
[24] Roboflow, Inc., YOLOv12: State-of-the-Art Object Detection Model, 2026. URL: https://yolov12.com/.
[25] I. Krak, M. Molchanova, V. Didur, O. Sobko, O. Mazurets, O. Barmak, Method of semantic features estimation for political propaganda techniques detection using transformer neural networks, in: S. O. Semerikov, A. M. Striuk (Eds.), Proceedings of the 7th Workshop for Young Scientists in Computer Science & Software Engineering (CS&SE@SW 2024), Virtual Event, Kryvyi Rih, Ukraine, December 27, 2024, volume 3917 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 286–297. URL: https://ceur-ws.org/Vol-3917/paper56.pdf.
[26] I. Krak, V. Didur, M. Molchanova, O. Mazurets, O. Sobko, O. Zalutska, O. Barmak, Method for Political Propaganda Detection in Internet Content Using Recurrent Neural Network Models Ensemble, in: I. Sinitsyn, P. Andon (Eds.), Proceedings of the 14th International Scientific and Practical Programming Conference (UkrPROG 2024), Kyiv, Ukraine, May 14-15, 2024, volume 3806 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 312–324. URL: https://ceur-ws.org/Vol-3806/S_36_Krak.pdf.
[27] J. You, J. Korhonen, Transformer for image quality assessment, in: 2021 IEEE international conference on image processing (ICIP), IEEE, 2021, pp. 1389–1393. doi:10.1109/ICIP42928.2021.9506075.
[28] N. Jiang, K. Wang, X. Peng, X. Yu, Q. Wang, J. Xing, G. Li, G. Guo, Q. Ye, J. Jiao, J. Zhao, Z. Han, Anti-UAV: A Large-Scale Benchmark for Vision-Based UAV Tracking, IEEE Transactions on Multimedia 25 (2023) 486–500. doi:10.1109/TMM.2021.3128047.
[29] O. Sobko, O. Mazurets, M. Molchanova, I. Krak, O. Barmak, Method for analysis and formation of representative text datasets, in: T. Hovorushchenko, E. Zaitseva, S. Lysenko, V. G. Levashenko (Eds.), Proceedings of the 1st International Workshop on Advanced Applied Information Technologies with CEUR-WS, Khmelnytskyi, Ukraine / Zilina, Slovakia, December 5, 2024, volume 3899 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 84–98. URL: https://ceur-ws.org/Vol-3899/paper9.pdf.
[30] Kivy: Cross-platform Python Framework for GUI apps Development, 2025. URL: https://kivy.org/.