



Available online at www.qu.edu.iq/journalcm

JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS

ISSN:2521-3504(online) ISSN:2074-0204(print)



UAV-STNet: A Hybrid U-Net-Transformer-LSTM Architecture for Real-Time Object Detection in UAV Video Surveillance

Shoohi, Liqaa M.

College of Physical Education and Sports Science for Girls, University of Baghdad, Baghdad, Iraq. E-mail: liqaa.m@uobaghdad.edu.iq

ARTICLE INFO

Article history:

Received: 16 /03/2026

Revised form: 24 /04/2026

Accepted : 27 /04/2026

Available online: 30 /09/2026

Keywords:

UAV video surveillance, Hybrid deep learning, U-Net, Transformer and LSTM.

ABSTRACT

It is difficult to perform real time object detection in Uncrewed Aerial Vehicle(UAV) based video surveillance because of the dynamic movement of camera, scale, occlusion, variation of illumination and limited power availability on-board of the computer. Purpose: In this paper of researches, the author suggests the proposed UAV Spatial-Temporal Network(UAV-STNet), which is a hybrid model of spatio-temporal deep learning model that is expected to improve the accuracy of the detection, balance of time, and real-time performance. Approaches: techniques: The proposed framework incorporates U-Net representing the system of total encoder involving extraction of multiscale space features mode, Transformer attention module which will involve global contextual modeling and Long Short-Term Memory(LSTM) network which will involve learning the short-term dependency in sequential frames. The model has been lives with and examined against a personalised UAV video dataset using a trade mark assortment of generally utilised measures of detection, including precision, recollection, mAPat 0.5 and inference velocity (FPS). Findings: UAV-STNet gave 94.5% precision, 93.1% recall and 93.2% mAP @0.5 and also 45 FPS. Better accuracy-tolerance efficient state is observed in comparative comparison with SSD, YOLOv3 and faster R-CNN which are evidently small object in motion affected scenes. Conclusions: The spatial, contextual and temporal modelling applied as a single-stop end-to-end architecture shows a powerful and computationally efficient solution of the topic of detecting real-time UAV video items, that brings benefits in the stability and reliability of intelligent surveillance systems using aerial vehicles.

MSC..

<https://doi.org/10.29304/jqcm.2026.18.32716>

1. Introduction

Uncrewed Aerial Vehicles (UAV) have become fundamental systems for the intelligent surveillance and tracking of the environment in real time. Their mobility, flexible deployment and wide area coverage capabilities make them ideal candidates for programs with the following characteristics: border security, disaster response and site visitor control, smart cities infrastructure and critical asset protection [1]. Especially in these areas computerized object detection based on UAV video stream is an indispensable element of possible situational recognition and independent decision making. Notwithstanding the outstanding advancements in deep learning-based total object detection for the last couple of years, the real-time detection in the video capture from UAV is a technical challenging problem. In contrast to traditional images of floor stage, aerial video has certain special characteristics, including extreme scale, small

*Corresponding author: Shoohi, Liqaa M

Email addresses: liqaa.m@uobaghdad.edu.iq

Communicated by 'sub etitor'

object illustration, dynamic camera movement, standard viewpoint changes, inconsistent illumination, background clutter and motion blur. These challenges severely compromise the balance of detection as well localization [2]. Moreover, UAV systems are normally restricted by tight computational and power budgets, rendering it percent perfect for models, which not only strike a high degree of accuracy with effective velocity of inference [3]. Aggressive performance has been obtained on benchmark data sets by the use of modern object detection frameworks such as YOLO, SSD and faster R-CNN. However, most of such procedures are basically frame-oriented and extract function mainly from space but not from time frame-by-frame, mostly from video sequences. This predicament can lead in inconsistent prediction of objects to frames and more false positive and unstable localization in dynamic aerial environment [4]. Recent breakthroughs in hybrid deep learning architectures demonstrate that use of complementary modeling architectures may provide a lot of gains in detection robustness. Convolutional encoder-decoder networks (specifically based on U-Net) is possible by powerful multiscale feature extraction and spatial correlation accuracy. Transformer based interest mechanism enhances the contextual reasoning like modeling the long-range dependency and discriminating between the foreground items and complicated backgrounds. Simultaneously, regular architectures, such LSTM network provide temporal modeling skills which helps to stabilize the predictions for both sequential and consecutive frames [5]. Inspired by these insights, in this work, a unified spatio-temporal item-detection framework specially aimed at UAV video surveillance is propose (UAV-STNet). The proposed architecture is an integrated way of:

- Improved U-net network based encoder with hierarchical multiscale spatial illustration,
- Transformer-based interest module for global contextual modelling,
- SomnetLSTM Modular Out Lorraine Net SomnetLSTM Modular Out Lorraine Net is LSTM Module for short-term temporal dependency learning of consecutive frames.

And in one end-to-end detection pipeline, these two representations (spatial, contextual and temporal) together optimize and aim to obtain better stability between detection accuracy, robustness and computation efficiency: UAV-STNet. The general accomplishments of this research can be summarized as follows:

1. We propose a new hybrid U-Net -Transformer-LSTM architecture designed for totally real-time object detection with UAV.
2. We develop a unified spatio-temporal detection pipeline that explicitly constructs multiscale spatial features and inter-object temporal dependencies.
3. For the training, we perform the full scale comparison experiments to YOLOv3, SSD and Faster R-CNN with the same training and testing protocols.
4. We show that UAV-STNet provides an advanced accuracy-speed trade-off, achieving 94.5 per cent precision and 93.2 per cent mAP@0.5 simultaneously at the same time of maintaining real-time inference at 45 FPS.

The rest of this paper is organized as follows: Section 2 reviews the relations literature; Section 3 describes the proposed methodology; Section 4 describes the experimental evaluation; Section 5 discusses demanding situations and limitations; and Section 6 concludes the examination.

NOMENCLATURE

Aradius of

Bposition of

Cfurther nomenclature continues down the page inside the text box

2. Related Work

Current studies can inform the future of research on UAV object detection and video analysis. As an example, Chao et al. (2025) [6] designed lightweight object detection frameworks on UAV images to overcome computation costs on the embedded systems. In a similar way, Liu et al. (2026) [7] concentrated on real-time optimization techniques of UAV-based detection with limited computational resources. Second, the smaller object detection is another crucial issue in aerial imagery. Hua et al. (2025) [8] pointed out that the existing deep learning models have limitations at identifying very small or distant objects and stressed upon more sophisticated methods of representing features on different scales. Furthermore, the environmental issues that were mentioned in Zhang et al. (2024) [9] include motion blur, variations in light, and unfavorable weather conditions, where the authors emphasized the need to have robust and adaptive detection systems in a real-world UAV scenario. The next phase of this study will be to come up with the lightweight variations of the UAV-STNet using model compression algorithms like pruning and quantization, better small object detection by augmenting the multiscale fusion of features, and the use of adaptive mechanisms to deal

with the dynamic environment based on these studies. Moreover, explainable AI methods could be used to enhance model interpretability and reliability in a critical surveillance application.

2.1 Research Gap and Motivation

Despite recent advancements in UAV-based object detection, several key challenges remain unresolved in the literature. Most real-time detection models prioritize spatial feature extraction efficiency but lack effective temporal modeling for video sequences. Transformer-based architectures provide strong contextual reasoning but are computationally expensive and not fully optimized for UAV deployment. Video-based detection frameworks incorporate temporal aggregation; however, they often lack multiscale feature refinement required for small aerial object detection. From the literature, it is evident that few existing approaches successfully integrate multiscale spatial feature extraction, global contextual attention, and temporal sequence modeling within a single unified architecture optimized for real-time UAV surveillance [10]. In particular, UAV-based detection systems still struggle with the following challenges:

- Accurate detection of small objects in high-altitude imagery,
- Maintaining temporal stability under camera motion,
- Distinguishing objects from complex and cluttered backgrounds,
- Achieving real-time inference under limited computational resources.

These limitations indicate that current methods do not fully address the trade-off between detection accuracy, robustness, and computational efficiency. Therefore, to bridge this gap, this study proposes UAV-STNet, a hybrid deep learning framework that integrates U-Net for multiscale spatial feature extraction, Transformer for global contextual modeling, and LSTM for short-term temporal dependency learning. Unlike previous approaches that treat spatial and temporal modeling separately, UAV-STNet tightly couples these components into a unified end-to-end architecture, aiming to achieve a balance between accuracy, robustness, and real-time performance in UAV video surveillance tasks [11].

2.2 Evaluation Metrics

To more comprehensively evaluate the overall performance of the proposed UAV-STNet architecture, the popular item detection metrics were used to evaluate the overall detection accuracy and the real-time processing performance. In order to quantitatively compare the performance of proposed UAV-STNet model, several well-known item detection metrics have been used which include precision, recall, F1-score, and intersection over union (IoU) [12][13].

In this study:

- **TP (True Positives)** are detected objects that are correct.
- **FP (False Positives)** represents incorrectly detected objects.
- **FN (False Negatives)** are ground truth objects that were missed by the model.

- **Precision**

Precision examines the percentage of the positive predictions that were right out of all the positive predicted:

$$Precision = \frac{TP}{TP + FP}$$

- **Recall**

Recall measures the fraction of detected objects which are correct:

$$Recall = \frac{TP}{TP + FN}$$

- **F1-Score**

The F1-score gives a metric that gives a harmonic mean of the Precision and Recall:

$$F1 = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)}$$

- **Intersection over Union (IoU)**

IoU is used to evaluate an intersection between ground-truth and predicted bounding box:

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union}$$

Detection of a particular object is taken at the correct object when it has an IoU of 0.5 and upwards.

- **Mean Average Precision (mAP)**

Average Precision (AP), calculated with respect to all object classes to gain the mean Average Precision (mAP):

$$mAP = \frac{1}{C} \sum_{c=1}^C cAP$$

Where C represents the number of object classes, this study used mAP@0.5 as the primary accuracy metric.

- **Frames Per Second (FPS)**

FPS is used to measure inference speed and it is given as:

$$FPS = \frac{Total\ number\ of\ processed\ frames}{Total\ inference\ time}$$

This measure is essential to UAV real time surveillance packages, where latency to detection is directly proportional to responsiveness of devices.

3. Materials and Methods

3.1 Dataset Description

The experimental evaluation of the proposed architecture of UAV-STNet was performed based on an aerial video surveillance dataset taken on uncrewed aerial vehicles (UAVs). The dataset has been designed to simulate realistic scenarios for tracking using UAV technology and features challenging environments such as complicated backgrounds, motion blur, scale variation, camera motion and partial object occlusion [14]. The dataset is composed of 15 video sequences of UAVs, which were recorded with a frame-rate of 30 FPS from different flight altitudes and viewing angles. From these videos, in total, 18,000 frames had been extracted and manually annotated. Everybody uses one or more objects of interest, which are labeled using axis-aligned bounding bins with corresponding class labels. Three identified item categories had been considered in this study:

- Pedestrians: 8,100 instances (45%).
- Vehicles: 6,840 instances (38%).
- Small objects (e.g., bicycles and distant targets) 3,060 instances 17%.

The annotated dataset is divided into training, validation and test data sets with ratios of 70%/15%/15% to make sure that it can be fairly evaluated and that overfitting can be avoided. All subsets have temporal continuity, to help the series-based learning for LSTM part. This dataset was curated in response to the lack of availability of annotated UAV video datasets based on the lack of a useful dataset for reproducible evaluation of spatio-temporal object detection techniques. The dataset will be available for instructional studies through inquiry.

3.2 Preprocessing and Frame Preparation

The section will outline the preprocessing pipeline that will be used in the proposed UAV-STNet framework. All processes are utilized throughout the experimental implementation phase to get ready UAV video data to train the models and infer. To ensure good performance in multiple scenarios of UAV surveillance, a series of preprocessing

steps were applied to the raw video frames before feeding them to the model for training. These steps aim to improve the consistency of information, reduce the noise and improve the generalization function of the proposed UAV-STNet framework. First, all video sequences were decomposed into frames along with temporal order, in order to enable series-based mastering in the LSTM module. The resolution of each frame is downsampled to a resolution of 512×512 pixels for a tradeoff between preserving spatial detail or speed and computational power for a near real-time evaluation. Pixel intensity values have been normalized to lie between $[0,1]$ in order to stabilize the education process and get results much faster. In order to increase the robustness to various illumination conditions and the movement of digital cameras, record augmentation schemes were adopted during the education task, such as random horizontal flipping, moderate rotation (from $+10\text{deg}$ to -10deg), scaling, and brightness adjustment. To maintain time coherence those augmentations have always been applied over successive frames. To allow for the temporal modelling element, consecutive object sequences of length $T = 3$ had been constructed in the form $(t-1, t, t+1)$. This narrow time window offers a window of opportunity for efficient movement sampling mastery and a low latency, which is a critical part of the real-time UAV surveillance system.

At the end of all the annotation has been synchronized with the processed frames - thus bounding boxes and the corresponding objects can be correctly aligned with each other after resizing and augmentation. This preprocessing pipeline helps to improve the accuracy of detection, helps to perform robust temporal monitoring of monitoring, and allows to have consistent detection performance on varied UAV environment.

3.3 Proposed THE UAV-STNet Architecture

The proposed UAV-STNet is a hybrid deep learning architecture which is suitable for the real-time object detection for UAV based Video Surveillance. The version combines spatial function extraction, modeling of context and temporal sequence learning in united end-to-end detection framework. The general pipeline procedures will generate quickly frame sequences are inside the form $(t-1, t, t+1)$ to seize each spatial movement each record. Figure 1 indicates implementation of UAV-STNet proposed for object detection.

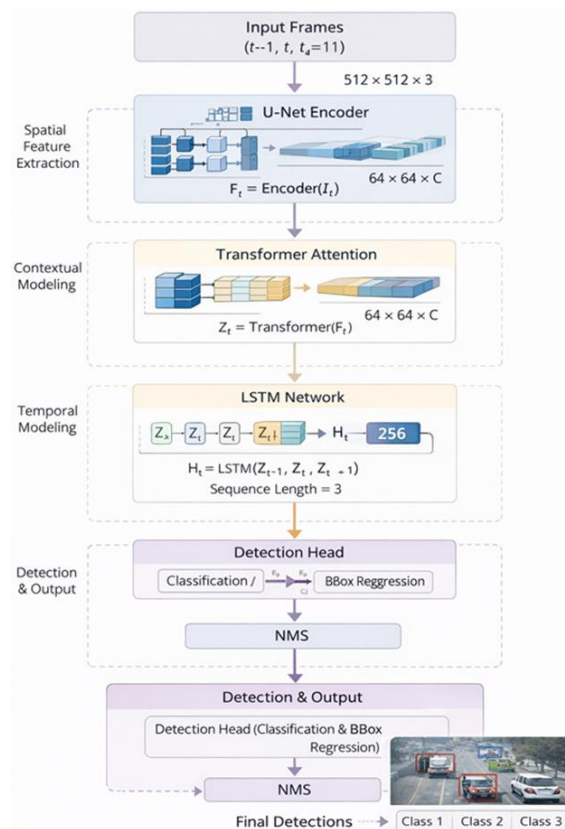


Figure 1. Illustrating the operation of UAV-STNet, the proposed architecture model

1. **U-Net Encoder for Multiscale Spatial Feature Extraction:** The spatial backbone of UAV-STNet is constructed on the basis of a modified U-net encoder architecture. The encoder is normally applied by convolutional and downsampling layers that prune out a number of spatial dimensions and make semantic intensity. This multiscale function extraction aids the network to detect small and large objects which is important for aerial UAV imagery since the scale variation is considerable. Skip connections are used here to preserve the high resolution spatial statistics and enhance the localization accuracy. Contrary to segmentation-centered unets, the lapper of the decoder output is made tailored towards the detection-specific representation of function rather than in pixel levels.
2. **Transformer Attention Module for Contextual Modeling:** In order to better the global contextual modeling, Transformer-based attention module is used after the encoder stage. This component models the long-range dependencies at spatial feature maps (enabling one to more discriminately identify partially occluded or items that overlap closely). The self-interest mechanism is applied dynamically in order to assign weights to regions in space to enhance its robustness of the mechanism in such cluttered UAV scenes and complex backgrounds. This contextual modelling is rather important in case of aerial surveillance, where some objects can also be small and packed a lot.
3. **LSTM-Based Temporal Modeling:** As the video captured from UAVs has a strong temporal relation, long short term memory network (LSTM) is used to capture the continuity of movement done by the frame which is captured later. The LSTM methods were a conglomeration of the spatial and contextual abilities that were passed down to them by the Transformer enhanced encoder outputs. By the modeling of brief term temporal relationship into the sequence $t-1$, t , $t+1$ then it help the network to lower the false positives caused by transient noise as well as enhance the stability of detection during the transfer of objects. This temporal integration complements tracking consistency without the serious increase in the computation overhead.
4. **Feature Fusion and Detection Head:** The fusion of spatial capabilities of the U-Net and the contextual capabilities of the Transformer module are fused towards a characteristic fusion level and handed to the LSTMs unit. The fused illustration then is also fed in a detection head which is responsible for:
 - Object classification.
 - Bounding box regression.

The output of the detection head is final predictions in the form of bounding boxes, class labels, and confidence scores for each frame. Non-Maximum Suppression (NMS) is applied throughout the entire inference process in order to eliminate redundant overlapping detections.

5. Architectural Advantages:

The proposed UAV-STNet architecture has many advantages:

- There is a special point when discussing this paper. U-Net backbone captures the hierarchical function extraction capability, which facilitates a progressive detection of small-scale items that are generally discovered in UAV imagery.
- This is, for example: - The Incorporated Transformer module improves the global feature interaction, which allows to better discriminate the overlapping and partially occluded objectives.
- The LSTM module minimizes the prediction fluctuations among objects through the assistance of modeling short-term temporal dependencies.
- The architecture has real-time performance levels that are suitable in the deployment of UAVs.

The combination of those additives will allow competitive levels of detection accuracy while maintaining real-time level of inference speed, and there will be an accuracy-computational-efficiency compromise. This allows UAV-STNet to be suitable for real-time aerial surveillance packages. This is an interim point between the evaluation based on simulations and the actual deployment of UAVs in the field.

3.4 Baseline Detectors and Comparative Framework

This part is part of the experimental approach of the research and includes the base models employed in the assessment. It is not entirely theoretical and it is rather a matter of practice and resultant comparison of the chosen models in the proposed framework. These models are described theoretically based on existing literature, whereas they are applied in the study to use in terms of benchmarking. These models are chosen according to the well-known literature on research on object detection. To carefully compare the overall performance of the proposed UAV-STNet, three widely recognized object detection architectures were selected as baseline models namely YOLOv3, Faster R-CNN and SSD. These models were selected because of their real-time performance and accuracy detection capabilities, which make them good benchmarks for UAV-based video surveillance programs.

- **YOLOv3:** is a single-stage detector which has the ability to object localization and classification in a single forward [15]. It is known to be known for being very fast in terms of inferences and prone to be used in real time applications. However, its performance may suffer even more to situations with small items or highly packed items which are ubiquitous in aerial imagery.

- **Faster R-CNN:** Real-time Object Detection using Faster Region-Based Convolutional Network Faster R-CNN is a Faster R-CNN, which is a high level detection network that uses a place thought network (RPN) seen through category and bounding box refinement [16]. Although normally very accurate at the detection part it has a high computational complexity and thus not making it as suitable for rigging UAV deployment (real-time).

- **SSD (Single Shot MultiBox Detector):** SSD is a single-stage detector that makes use of multiscale feature maps to detect objects of varying sizes [17]. It has a good stability in the form of velocity and accuracy and often considered as good candidate for real-time application. In preliminary experiments, SSD had the best inference speed of the baseline models.

- **A protocol for fair comparison :** To compare the fashions in a truthful manner, all the baseline fashions were trained and evaluated using:

To ensure the fairness of the method and repeatability, all baseline models and the proposed UAV-STNet had been:

- Trained on the same dataset of UAVs.
- Evaluated the usage of Identical teach/take a look at splits.
- Worked in the same enter resolution (512 × 512).
- Assessed the usage of similar assessment metrics (Precision, recall, F1-score, mAP and FPS).
- implemented on the equal hardware configuration.

This unified experimental setup allows that changes in overall performance are observed to be due to architectural layout and not implementation inconstancies.

A selective evaluation between UAV-STNet and SSD is performed due to the good real-time performance shown by SSD during the baseline evaluation. At the same time, in terms of inference velocity, SSD obtains the high inference velocity through the single degree detection mechanism, and obviously, throughout the consecutive frames, it does not explicitly model the temporal dependency in consecutive frames, and of course, there are no advanced global attention mechanism. In evaluation, UAV-STNet combined multiscale spatial extraction (U-Net), global contextual modeling (Transformer), and temporal series studying (LSTM) in one. Consequently, it is not limited to comparison of the uncooked inference speed, but the contrast extends to evaluation of the robustness of detection, contextual expertise and time consistency. This analysis allows at the same time to evaluate the trade-off between the performed computations (FPS), the detection accuracy (mAP, F1-score) and the temporal balance, key elements in the framework of UAV-based, fully real-time surveillance systems.

Thus, the section is a continuation of the experimental design as opposed to a theoretical review. It provides equitable benchmarking and a clear difference between description of the model based on literature and its practical experimental role in this research.

3.5 Training Configuration

The proposed UAV-STNet model is train-to-end by a unified optimization approach, which jointly optimizes spatial representational, contextual representations and temporal representations..

-Loss Function:

A composite loss feature turned into hired consisting of:

- Classification loss Binary cross-entropy (BCE) loss for the magnificence of the object.
- Localization loss: Complete IoU (CIoU) [18] loss for precise bounding container regression.

The total loss is calculated as a sum over the weight of type and localization losses.

- Optimizer:

The network was optimized using the Adam optimizer with a learning rate of 1×10^{-4} , at the start of training, which is used for training hybrid architectures combining CNNs, transformers, LSTMs networks because of its stability during training.

- Learning Rate Schedule:

Step learning rate decay is taken into consideration in form of decreasing learning rate by 0.1 and validation loss stopped improving for 10 epochs.

- Batch Size and Epochs:

Again, the training was based on a batch size of 8 video sequences, with each sample representing three consecutive frames (from $t-1$, t , and $t+1$).

The model became good after 100 epochs; early stopping being used to avoid overfitting.

3.6 Implementation Details

This section describes the practical implementation of the proposed system and belongs to the experimental methodology rather than the results. The suggested version of UAV-STNet was developed with PyTorch 2.1 with a GPU acceleration. Python 3.10 was used to run all experiments. The hybrid organization incorporates:

- A core modification of the U-Net encoder decoder.
- **Configuration of the Transformer:** Transformer interest module is made up of a single layer of encoder with multi-head self-interest. In our system, we have used 8 attention heads, 256 of embedding size, and 1024 of the feed-forward network size. To minimize overfitting a dropout rate of 0.1 was used. Positional encoding was added in order to maintain information of spatial order before attention computation. The Transformer uses flattened spatial feature maps of the U-Net encoder and additional contextual illustration globally and subsequently temporal modeling.
- **Configurations LSTM:** The LSTM was modeled with a unidirectional lengthy short-term memory layer, allowing quick-time-period temporal dependencies between frames ($t-1$, t , $t+1$). The LSTM reduces the fused spatial contextual functions of the Transformer, resulting in more advantageous encoder outputs that permit inter-frame smoothness and sophisticated inter-frame compatibility. This design guarantees effective time-modelling, and the computational speed in real-time to use UAVs.

Training and inference were done on a workstation that included:

- NVIDIA RTX 4090 GPU (24GB VRAM)
- 64 GB RAM
- Intel Core i9 CPU
- CUDA 12.2

Preprocessing, frame extraction, and augmentation have been implemented using OpenCV, model training, optimization, and inference have been done in PyTorch. All the baseline models (YOLOv3, SSD, and Faster R-CNN) have been trained and tested with the same hardware and identical input settings (512×512) to make sure fair comparison.

4. Results and Discussion

Under the setup that was already established in segment 3.7, experiments were performed. These evaluation metrics are the precision, recall, mean average precision (mAP 0.5) and the frames per second (FPS).

- Quantitative Comparison

Table 1 indicates the quantitative comparison of the proposed UAV-STNet and popular object detection models.

Table 1: The comparison of the suggested UAV-STNet and popular object detectors

Metric	UAV-STNet	SSD	YOLOv3	Faster R-CNN
Precision (%)	94.5	91.2	92.7	90.3
Recall (%)	93.1	89.5	90.8	88.6
mAP@0.5 (%)	93.2	89.9	91.2	88.1
FPS	45	50	48	12

Compared with all the other models, the proposed UAV-STNet has the highest precision and mAP, and it still has competitive real-time aggregate performance at 45 FPS.

Although SSD has gained only a slightly better FPS (50), it implies reduced detection accuracy especially in small and partially covered objects.

- Comparative Study of SSD and UAV-STNet

Although, Owing to its best inference speed (50 FPS), SSD performs the best inference, UAV-STNet exhibits higher detection overall performance in all the metrics of accuracy. Specifically, UAV-STNet achieves better precision (+3.3%), recall (+3.6%), and mAP at 0.5 (+3.3%) compared to SSD. The performance difference is conventionally explained by the lack of object-sensing detection strategy at SSD with the clear temporal modeling and the global contextual consideration. This weakness causes unsteady detection when using the UAV in dynamic settings and moving objects with reduced sizes. UAV-STNet is the combination of transformer-based, entirely contextual attention and LSTM-based temporal modeling, which complement and make inter-frame consistency and robustness stronger in evaluation. The 5 FPS difference only results in a 10% speed discount and over 3 proportion of an increase in the detection accuracy which implies that acceptable accuracyperformance trade-off is higher when it comes to real-time UAV surveillance programs.

- Precision and Recall Analysis

The high accuracy (94.5) and the recall (93.1) demonstrates the usefulness of combining

- Multi-scale spatial representation (U-Net encoder).
- Global contextual modeling (Transformer).
- Temporal smoothing (LSTM).

The temporal model significantly minimizes instability of frame-to- frame detection when specified to areas underneath UAV vibration and unstable movement.

- Real-Time Performance

The proposed model supports performance of 45 FPS (real-time) along with 512×512 input resolution (in contrast with the much faster R-CNN of 12 FPS); UAV-STNet is able to achieve a significantly greater velocity-accuracy trade-off.

- Qualitative Analysis

The examples of visual detection in difficult conditions are marked in Figure 2:

- Low-light environments
- Partial occlusion
- Dense history clutter.

- Detection of small objects at the elevation

UAV-STNet displays more equal localization by consecutive frames as compared to YOLOv3 and SSD.

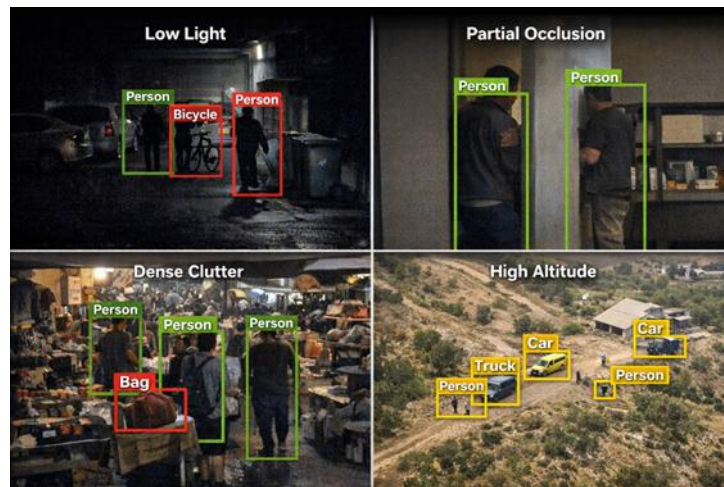


Figure 2. UAV-STNet Object Detection Striving against Adverse Weather

- Ablation Study

Table 2 indicates that an ablation study was conducted to evaluate the contribution of each architectural issue.

Table 2. Ablation study evaluating the contribution of each architectural component

Model Variant	Precision (%)	mAP@0.5 (%)
U-Net Only	89.4	87.2
U-Net + LSTM	91.6	89.5
U-Net + Transformer	92.3	90.8
Full UAV-STNet	94.5	93.2

The results confirm that:

- Balance is enhanced by temporal modeling.
- Transformer attention improves the small object detection.
- An overall overall performance of a very high quality is seen in the total hybrid structure.

- Figure 3 will suggest the comparison of Precision and Recall of YOLOv3, Faster R-CNN, SSD and UAV-STNet.



Figure 3. Shows the comparison of precision and recall of YOLOv3, Faster R-CNN, and SSD and UAV-STNet

- Practical cases of detection are shown in Figure 4. The YOLOv3 would lack some tiny/ obscured objects; faster R-CNN would be quicker and less efficient, whereas UAV-STNet would be trade-off between speed and precision.



Figure 4. Demonstrates common sense examples of detection. YOLOv3 lacks small or hidden objects however Faster R-CNN is more accurate but slower than UAV-STNet

5. Challenges and Limitations

Although the overall performance of the suggested UAV-STNet architecture is very good, several practical and technical difficulties await the implementation of the systems of surveillance on the basis of the UAV.

1. Limited Onboard Computational Resources

Uncrewed Aerial Vehicles (UAVs) are typically limited by minimal onboard processing capability as well as RAM and battery duration. In spite of the fact that UAV-STNet can run in real-time (45 FPS) on a high-performance computer-based workstation, switching to the full version running directly on embedded UAV hardware can reduce the achievable frame rate. The complexity of inference will go up due to the combined computational cost of U-Net backbone, transformer attention modules and LSTM temporal modeling. As a result, compression strategies like pruning, quantization, and knowledge distillation might be necessitated in the case of efficient edge deployment.

2. Small Object Detection during High Altitude Views

Devices taken at elevated altitudes usually use a small number of pixels per object. Regardless of the mechanism of multi-scale feature extraction implanted into UAV-STNet, objects of extremely small size or some distance can still be difficult to see, especially when there is a lot of clutter or low-contrast objects in the surroundings. This is still an open task in the aerial laptop vision and will require additional research.

3. Motion Artifacts and Environmental

UAV imaging frameworks are subject to many environmental and mobility-related disruptions and they include :

- Mobile camera shakings and motion blur.
- Adjustments in perspectives fast.
- Illiteracy variation.
- The adverse weather conditions such as fog, dust, and rain.

In spite of the fact that LSTM-based temporal modeling enhances object-to-frame stability, object severe motion artifacts still reduce the reliability of detection and accuracy of bounding field.

4. Dataset Diversity and Generalization

Although the data set employed in this test includes several classes of objects and different environmental factors, the real world case of international surveillance might also be much more diverse. Magnificence bias (at least in rare object classes) of the model might be biased to dominant courses, e.g. cars. Future research will need the large-scale and more diverse UAV data to increase the performance of generalization.

5. Complexity and Interpretability of Models

The hybrid architecture using U-Net, Transformer and LSTM additives enhance the detection quality and complexity of the architecture. The framework proposed is less interpretable and more computationally intensive than more basic single-level detectors. Explainable AI methods could also be examined in future studies as they could improve transparency, debugging, and machine preservation.

6. Conclusion

This project introduced UAV-STNet, a hybrid, deep learning architecture to identify objects in real-time under UAV-based video surveillance. The model combines U-Net (to extract multiscale spatial features), Transformer (to model global context), and LSTM (to learn temporal dependencies); a jointly trained spatio-temporal architecture is created. The effectiveness of the proposed model is proven by experimental results yielding higher performance than the state-of-the-art detectors. UAV-STNet achieved 94.5% precision, 93.1% recall, and 93.2% mAP@0.5, outperforming YOLOv3, SSD, and Faster R-CNN in all accuracy metrics. The model also supported real-time performance at 45 FPS, which again verifies that the model is suitable in real-life UAV surveillance solutions. UAV-STNet has a superior trade-off between the accuracy of detection and the speed of inference than the SSD, which had higher speed (50 FPS) but lower accuracy. This has been made possible mainly because of the use of spatial, contextual, and temporal features that are incorporated under one architecture. Despite the good performance demonstrated by the proposed model,

there are certain shortcomings regarding the complexity of calculations and application to lightweight UAV devices. Nevertheless the findings indicate that UAV-STNet enhances much better the detection resilience, particularly when using in advanced aerial scenarios that involve small objects, motion blur, and intricate backgrounds. Future directions will include efforts on the model compression methods based on pruning and quantization to be more deployment-efficient, and how to better generalize to broader UAV datasets.

References

- [1] G. Tang, J. Ni, Y. Zhao, Y. Gu, and W. Cao, "A Survey of Object Detection for UAVs Based on Deep Learning," *Remote Sens.*, vol. 16, no. 1, Art. no. 149, Dec. 2023, doi: 10.3390/rs16010149.
- [2] Y. Wu and K. Tong, "Research advances on deep learning-based small object detection in UAV aerial images," *Acta Aeronaut. et Astronaut. Sinica*, vol. 46, no. 3, pp. 30848–30848, Feb. 2025, doi: 10.7527/S1000-6893.2024.30848.
- [3] E. Kirac and S. Özbek, "Deep Learning Based Object Detection with Unmanned Aerial Vehicle Equipped with Embedded System," *J. Aviation*, vol. 8, no. 1, pp. 15–25, Feb. 2024, doi: 10.30518/jav.. 1356997.
- [4] H. Wang, Y. Li, J. Chen, and X. Gao, "Temporal Feature Aggregation for Real-Time Video Object Detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 5, pp. 4123–4137, May 2024, doi: 10.1109/TCSVT.2023.3312456.
- [5] P. Jiang, W. Liu, F. Wang, and R. Wei, "Hybrid U-Net Model with Visual Transformers for Enhanced Multi-Organ Medical Image Segmentation," *Information*, vol. 16, no. 2, 2025, doi: 10.3390/info16020111.
- [6] Chao, M., Peng, C., Yun, L., Zhang, C., Wang, H., and Chen, Z., "A lightweight small object detection model for UAV images based on deep semantic integration," *Scientific Reports*, vol. 15, no. 1, p. 31888, Aug. 2025, doi: 10.1038/s41598-025-16878-6.
- [7] X. Liu, Q. Chen, Y. Wang, and T. Huang, "Real-time UAV small object detection: SGFNet and dynamic loss optimization," *Digital Signal Processing*, vol. 150, p. 105543, 2026, doi: 10.1016/j.dsp.2025.105543.
- [8] Hua, W., and Chen, Q., "A survey of small object detection based on deep learning in aerial images," *Artificial Intelligence Review*, vol. 58, no. 162, pp. 1–35, 2025, doi: 10.1007/s10462-025-11150-9.
- [9] Zhang, Y., Wu, X., Li, H., and Sun, J., "Deep learning for UAV-based object detection and tracking: challenges and opportunities," *Information Fusion*, vol. 102, p. 102115, 2024, doi: 10.1016/j.inffus.2023.102115.
- [10] X. B. Xu, Z. Y. Xing, M. L. Sun, P. R. Zhang, and K. H. Yang "Enhancing UAV object detection through multiscale deformable convolutions and adaptive fusion attention," *J. Supercomput.*, vol. 81, no. 14, pp. 1301–1327, 2025, doi: 10.1007/s11227-025-07788-5.
- [11] W. Li, Y. Zhang, and X. Sun, "Enhancing Small Object Detection in UAV Aerial Imagery through Attention-Gated Backbone and Context-Aware Fusion," *Scientific Reports*, vol. 16, Art. no. 2366, Jan. 2026, doi: 10.1038/s41598-025-32074-y.
- [12] J. Tian, Q. Jin, Y. Wang, J. Yang, S. Zhang, and D. Sun, "Performance analysis of deep learning-based object detection algorithms on COCO benchmark: a comparative study," *Journal of Engineering and Applied Science*, 2024, article 76, doi: 10.1186/s44147-024-00411-z.
- [13] M. Dalal and P. Mittal, "A Systematic Review of Deep Learning-Based Object Detection in Agriculture: Methods, Challenges, and Future Directions," *Computers and Materials Continua*, vol. 70, no. 1, Jun. 2025, doi: 10.32604/cmc. 2025.066056.
- [14] Wanxuan Geng, Junfan Yi, Ning Li, Chen Ji, Yu Cong, and Liang Cheng, "RCSD-UAV: An object detection dataset for unmanned aerial vehicles in realistic complex scenarios," *Engineering Applications of Artificial Intelligence*, vol. 151, p. 110748, Jul. 2025, doi: 10.1016/j.engappai.2025.110748.
- [15] A. Aybilge Murat and M. Servet Kiran, "A comprehensive review on YOLO versions for object detection," *Journal of Engineering Science and Technology Review*, vol. 16, 2025, doi: 10.1016/j.jestech.2025.102161.
- [16] A. Javed Sayyad and K. Attarde, "A systematic literature review on deep learning approaches for small object detection," *Array*, vol. 100615, 2 Dec. 2025, doi: 10.1016/j.array.2025.100615.
- [17] S. E. Prasetyo, C. Atmaja, M. Ardian, A. Ardhiansyah, A. R. Sudarni, and M. Khaira, "Benchmarking YOLOv3 and SSD: A Performance Comparison for Multi-Object Detection," *Edu Komputika Journal*, vol. 11, no. 2, pp. 136–146, 2024, doi: 10.15294/edukom.v11i2.28005.
- [18] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression," *Proceedings of AAAI*, 2020, doi: 10.1609/aaai.v34i07.6999.