Person Detection from UAV Based on a Dual Transformer Approach

Andrei-Stelian Stan, Dan Popescu[a] and Loretta Ichim[b]
National *Romania*
andrei_stelian.stan@stud.etti.upb.ro, {dan.popescu, loretta.ichim}@upb.ro

Keywords: Neural Networks, Person Detection, Unmanned Aerial Vehicles, Detection Transformer, Vision Transformer.

Abstract: The study introduces a novel object detection system that combines the strengths of two advanced deep learning models, the Detection Transformer (DETR) and the Vision Transformer (ViT), to enhance detection accuracy and robustness in unmanned aerial vehicle (UAV) applications. Both models were independently fine-tuned on the VisDrone dataset and then deployed in parallel, each processing the same input to leverage their advantages. DETR provides precise localization capabilities, particularly effective in crowded urban settings. At the same time, ViT excels at identifying objects at various scales and under partial occlusions, which is crucial for distant object detection. The fusion of their outputs is managed through a dynamic fusion algorithm, which adjusts the confidence scores based on contextual analysis and the characteristics of detected objects, resulting in a combined detection system that outperforms the individual models. The fused model significantly improved overall accuracy, achieving up to 90%, with a mean Average Precision (mAP50) of 85%, and a recall of 80%. These results underline the potential of integrating multiple transformer-based models to handle the complexities of UAV-based detection tasks, offering a robust solution that adapts to diverse operational scenarios and environmental conditions.

1 INTRODUCTION

In the rapidly advancing field of artificial intelligence, neural network implementation for real-time object detection from unmanned aerial vehicles (UAVs) has emerged as a crucial area for academic, research, and industrial applications. UAVs, with the remarkable ability to access remote or challenging environments, present unprecedented opportunities across a spectrum of activities including urban surveillance, search and rescue missions, traffic monitoring, and environmental studies. These applications not only extend the capabilities of human operators by providing aerial insights but also enhance safety and operational efficiency, especially in scenarios where human presence could be hazardous or impractical (Wu et al. 2021). However, using UAVs for sensitive and dynamic tasks introduces complex challenges that standard object detection systems, designed primarily for static or terrestrial environments, struggle to handle. The factors complicating UAV-based person detection and tracking include high mobility, variable altitudes, and the vast range of lighting and weather conditions under which these drones must operate. Furthermore, the UAVs' rapid movement and the diverse angles of image capture add additional layers of complexity, requiring detection systems that are accurate, exceptionally robust, and adaptable to swift changes in the visual field.

To address these challenges, the present paper introduces a novel approach that harnesses the capabilities of two transformer-based neural network architectures: the Detection Transformer (DETR) (Huang and Li. 2024) and the Vision Transformer (ViT) (Wang and Tien, 2023). DETR revolutionizes object detection by utilizing a transformer-based set prediction mechanism that eliminates the need for the complex pipelines typical of conventional detection systems. This model simplifies the learning process and enhances the efficiency of detecting objects in real-time, a critical requirement for UAV operations.

Conversely, ViT adapts the transformer architecture - previously successful in the natural language processing domain - to image analysis. ViT treats an image as a sequence of patches and processes it through self-attention mechanisms,

[a] https://orcid.org/0000-0002-1883-0091
[b] https://orcid.org/0000-0002-7465-3958

Stan, A.-S., Popescu, D. and Ichim, L.
Person Detection from UAV Based on a Dual Transformer Approach.
DOI: 10.5220/0013467900003935

In *Proceedings of the 11th International Conference on Geographical Information Systems Theory, Applications and Management (GISTAM 2025)*, pages 95-102
ISBN: 978-989-758-741-2; ISSN: 2184-500X

allowing the model to capture intricate dependencies across the entire image. This capability is particularly beneficial for UAV imagery, where objects of interest may appear at various scales and in partial occlusions, often against highly cluttered backgrounds.

This paper's contribution lies in the strategic combination of these two powerful models. By deploying DETR and ViT in parallel, each model processes the same input independently, thus leveraging DETR's acute precision in localization and ViT's adeptness at handling scale variations and occlusions. This dual-model approach mitigates the limitations inherent in each model when used alone and capitalizes on their complementary strengths.

A dynamic fusion algorithm orchestrates the integration of outputs from both models. This algorithm does not merely aggregate confidence scores but also intelligently adjusts the fusion ratio in real-time, based on the contextual nuances and specific characteristics of detected objects. Such a sophisticated approach ensures that the system adapts continuously to complex and evolving landscapes of UAV operation, thereby enhancing detection accuracy and robustness across a wide range of operational scenarios. This fusion of DETR and ViT sets new standards in UAV-based surveillance and monitoring, promising substantial improvements in the reliability and effectiveness of such systems. The anticipated impact of this study spans improvements in operational safety, particularly in search and rescue missions, enhancements in surveillance accuracy for security applications, and greater data precision for environmental monitoring. This approach represents a significant technological leap in computer vision and heralds a paradigm shift in how UAVs can be utilized in complex and critical applications worldwide.

2 RELATED WORKS

A comprehensive benchmark of real-time object detection models tailored for UAV applications was presented by (Du et al., 2019). The authors developed new motion models to enhance detection accuracy in high-speed aerial scenarios, addressing challenges with rapidly moving objects. Their research highlighted the importance of integrating dynamic movement models into detection frameworks to improve response times and accuracy in UAV-captured imagery.

The Vision Transformer architecture was extended by (Wang and Tien, 2023) to better suit aerial image analysis by incorporating dynamic position embeddings. This adaptation allows the model to handle varying scales and orientations of objects typically found in UAV datasets. Their findings demonstrate significant improvements in object detection performance on aerial images, supporting the concept of transformers' adaptability to specialized tasks.

(Huang and Li. 2024) introduced enhancements in small object detection, focusing on information augmentation and adaptive feature fusion to improve detection accuracy and real-time performance. Their results demonstrate superior performance over the latest DETR model. This research is pertinent to our work as it highlights the effectiveness of advanced algorithms in refining object detection, echoing our approach to optimizing UAV-based detection with transformer architectures.

(Ye et al., 2023) introduced RTD-Net, tailored for UAV-based object detection. It addresses challenges like small and occluded object detection and the need for real-time performance. By implementing a Feature Fusion Module (FFM) and a Convolutional Multiheaded Self-Attention (CMHSA) mechanism, the network achieved improvements in handling complex detection scenarios, resulting in an 86.4% mAP on their UAV dataset. Their approach, emphasizing efficiency and effectiveness, aligns with our methods of optimizing object detection through advanced architecture fusion.

3 MATERIALS AND METHODS

3.1 Dataset Used

The VisDrone dataset, which contains diverse aerial images from various urban and rural scenes across Asia, was used in this study. Initially, the dataset included many objects, such as cars, buildings, and trees. The following steps were performed to tailor it to research needs.

Data Curation and Labelin were performed using custom Python scripts and LabelMG. The dataset was filtered to retain only images containing people. The annotations were re-labeled to ensure uniformity, combining labels for "person" and "people" into a single "person" label.

A format conversion was performed while preprocessing the dataset and researching ViT and DETR accepted formats. Originally in COCO format, the dataset was converted to Pascal VOC format. This involved adapting the annotations and restructuring the dataset files using a custom Python script.

The dataset was diversified having shots with persons taken from multiple angles and in different weather and light conditions such as during the night or when the sky is clouded, and the light is down (Figure 1).

Additionally, to enhance the robustness of the model, various augmentation techniques were applied to the original image (Figure 2a), such as blur (Figure 2b), (Figure 2c) decreased image brightness, and noise addition (Figure 2d). This process increased the dataset to over 2400 images, which were then split into training (70%), validation (15%), and testing (15%).

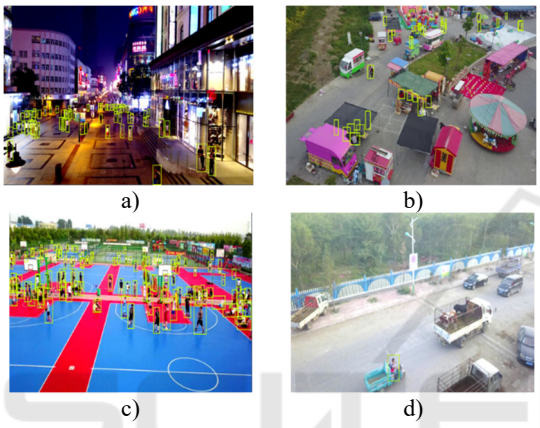

Figure 1: Examples of images from the VisDrone dataset. a) Persons in the city at night, b) Persons in a theme park from an angle on a cloudy day, c) Persons on a basketball court, d) A person driving a tuc-tuc in fog.

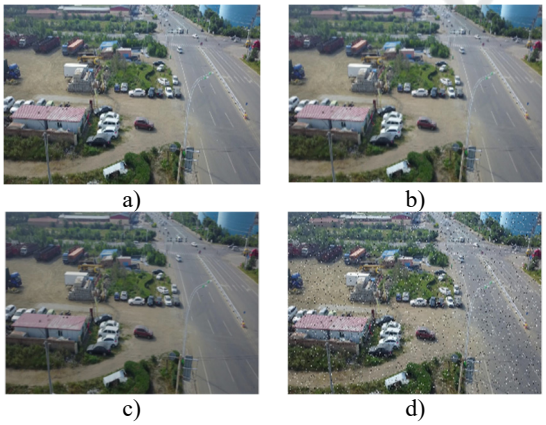

Figure 2: Examples of augmented images. a) Original image, b) Blurred image with 1.25px blur coefficient, c) Decreased image brightness by 15% d) Image with 4% noise coefficient.

Therefore, the size of the input images and the augmentation methods used (Table 1) created a dataset with images that contain more elements, and the edges of the objects are not as well defined as they were in the dataset used for YOLOv5 in (Stan et al. 2023).

Table 1: Augmentation methods, description, and values.

Augmentation Method	Description	Value
Image blur	Blur the image	2 pixels
Image noise	Modified the image to add noise to a percentage of the pixels	7% of pixels

3.2 Neural Networks Used

This sub-section explores the innovative use of neural networks, specifically focusing on the Detection Transformer (DETR) and Vision Transformer (ViT), for person detection and tracking from unmanned aerial vehicles (UAVs). These architectures leverage the power of transformers to enhance object detection tasks by simplifying the detection pipeline and enabling a more refined focus on small, distant objects typical in UAV imagery. These models improved in detecting and tracking persons in challenging UAV-captured scenes through architectural adjustments and fine-tuning relevant datasets.

DETR represents a novel approach to object detection, leveraging transformers for end-to-end object detection (Zeng et al., 2021). The implementation of Detection Transformer (DETR) for UAV-based person detection involves several steps aimed at leveraging its architecture for efficient and accurate bounding box predictions:

DETR utilizes a convolutional backbone (typically a ResNet-50 or ResNet-101) to extract feature maps from the input image. For UAV images, which often contain small and distant objects, the backbone is fine-tuned to improve its spatial resolution by adjusting the stride and kernel sizes, thus capturing finer details in the feature maps.

The transformer in DETR, which processes the outputs of the backbone, is configured to handle larger sequences of object queries. This is crucial for UAV imagery where multiple small objects might be present in a single frame. The number of object queries has increased, and the transformer is trained to focus on higher potential objects per image.

DETR simplifies the traditional object detection pipeline by eliminating the need for many hand-engineered components. It uses a set prediction loss that forces unique predictions via bipartite matching, and a transformer encoder-decoder architecture to

perform object detection as a direct set prediction problem.

The architecture consists of:

- Backbone: A convolutional neural network (ResNet) extracts feature representations from the input images;
- Transformer Encoder-Decoder: This core component processes the feature maps and performs object detection;
- Prediction Heads: These components predict bounding boxes and class labels for each detected object.

ViT segments an image into fixed-size patches linearly embeds each and then processes the sequence of embeddings using a standard transformer encoder (Zhang 2023). This method allows ViT to consider the entire image, unlike CNNs which process parts of an image in isolation. Architecture benefits from deeper attention layers providing the ability to focus on intricate details from complex backgrounds—common in UAV imagery.

When trained on large datasets like ImageNet-21k, ViT demonstrates superior performance in classification tasks. For object detection tasks specific to UAVs, ViT was fine-tuned on the VisDrone dataset, achieving a baseline accuracy of 72%, a recall of 0.70, and a mAP50 of 0.64.

3.3 Methodology

Google Colab was used for model training due to its cost-effective access to powerful GPUs and ease of setup. An A100 High RAM was utilized, having access to high-performance NVIDIA A100 GPUs. The Colab environment was configured with the necessary dependencies, including Python libraries such as PyTorch. Pre-trained COCO weights were loaded into the DETR and ViT models. This time the dataset was stored in Roboflow to simplify access to it and reduce the complexity of storing it. Annotation files were created to initialize the dataset loaders and load the model. For the training configuration parameters, we chose to test it with 30, 50, 70, 120, and 150 epochs. The learning rate optimized (Adam) dynamically adjusted the learning rate during the training.

One of the key innovations of DETR is its use of a bipartite matching loss, which directly matches predicted and ground truth objects. For UAV applications, the loss function is adjusted to be more sensitive to smaller objects by modifying the balance between the classification loss and the bounding box loss, placing more emphasis on the latter.

While DETR inherently reduces the need for Non-Maximum Suppression (NMS) through its set prediction mechanism, slight modifications are made to its post-processing NMS to better overlapping detections common in dense urban environments captured by UAVs. This involves tuning the IoU thresholds and the scoring system to finalize the detections.

DETR was trained using a mixed precision training regime to expedite the training process without losing the accuracy essential for real-time UAV operations. The training data is augmented with aerial-specific variations like varying scales, rotations, and lighting conditions to robustly train the model against the diverse conditions expected in deployment.

By refining these aspects of DETR, the model is better suited to the unique challenges posed by UAV-based detection tasks, especially in complex and cluttered environments.

The methodology centres on dual-model deployment where DETR and ViT operate in parallel. Each model processes the same input independently, allowing them to leverage their respective strengths:

- DETR excels in precise localization and good performance in crowded scenes (Song et al., 2021).
- ViT obtained good results in identifying distant or partially obscured objects due to its global processing capabilities (Song et al., 2021).

For the fine-tuning of ViT, adjustments are specifically made to its transformer's attention heads to enhance its ability to focus on small and distant objects in UAV imagery, which often appear as minute details within larger contexts:

The self-attention mechanism in ViT, which allows the model to weigh the importance of different parts of the input image, is fine-tuned to enhance its sensitivity to smaller patches. This is achieved by adjusting the attention head parameters to increase the model’s focus on areas of the image that contain less information but might be crucial for identifying distant objects.

The size of the image patches that the ViT processes is reduced. Smaller patches mean the model processes more patches per image, increasing the granularity of the attention and allowing the model to focus more accurately on small objects.

The transformer encoder within ViT is enhanced with additional layers. These layers allow for the learning of more complex representations and relationships between patches, which is particularly

beneficial for identifying objects at various scales and distances.

A dynamic fusion mechanism is employed to integrate detection results. This fusion is not merely a weighted average, but an adaptive process based on scene context and object characteristics. The fusion factor is initially set at 60% but can vary depending on object size, density, and environmental conditions.

The fusion mechanism is a critical component of the methodology. It is designed to effectively combine the strengths of DETR and ViT in detecting and tracking objects—specifically, people—from UAVs. The mechanism operates on the principle that while each model has its strengths, their combined insights can provide a more accurate and robust detection system, particularly in complex environments. The fusion strategy and all the phases used are explained below.

DETR and ViT receive the same input image concurrently and process it independently. This parallel processing ensures that each model applies its unique analytical approach to the same scene.

Each model outputs a set of bounding boxes with confidence scores for each detected object. DETR, which excels in precise bounding box predictions and handling overlapping objects, provides highly accurate localization. ViT, known for its ability to recognize objects across different scales and partial occlusions, offers robustness against challenging detections.

The confidence scores from each model are first normalized and then weighted by a pre-determined fusion factor. This factor is dynamically adjusted based on validation results for the better under specific conditions (e.g., DETR for closer objects and ViT for distant objects).

The fusion algorithm evaluates the spatial overlap (using Intersection over Union, IoU) and semantic agreement of the detected objects from both models. A higher agreement in both spatial and semantic terms increase the confidence in the combined detection. The way the final decision rule is configured for the final detection and classification is described below.

The first step is combining detections. The algorithm calculates a combined confidence score for each detected object, based on the weighted scores from both models. If the detections from both models for the same object have high spatial and semantic agreement, the combined confidence score is adjusted upwards.

The combined detections are then filtered through a thresholding mechanism where only detections with a combined confidence score exceeding a set threshold (85%) are retained. This helps to reduce false positive errors and ensures that only the most reliable detections are considered.

A modified NMS is applied to the final set of combined detections to refine the detection results. (Bolda et al., 2017). This step ensures if multiple bounding boxes are predicted for the same object, only the bounding box with the highest combined confidence score is retained.

The final output is a set of bounding boxes and associated confidence scores that represent the detected objects. Each bounding box is associated with a single object class, in this case, 'person', and reflects a higher detection accuracy than individual outputs of DETR and ViT.

This fusion mechanism not only leverages the individual strengths of each model but also introduces a robust method to adjudicate between their predictions, resulting in a more accurate and reliable object detection system for UAV applications. This approach is particularly effective in environments with diverse object scales and occlusions, typical of urban and crowded scenes.

The metrics used to evaluate the model performances are *Precision*, *Recall*, and Mean Average Precision at an Intersection over a Union threshold of 0.5 (*mAP*50). They are presented in Table 2 where *TP* = True Positives, *FP* = False Positives, *FN* = False Negatives, *P* = Precision, *R* = Recall, and *mAP* = Mean Average Precision. *mAP*50 calculates the *mAP* value for an Intersection Over Union (*IoU*) threshold of 0.5. It measures how well the model detects objects at this specific *IoU* threshold, indicating the proportion of correctly identified objects.

Table 2: Metrics used to evaluate the model.

$P = \frac{TP}{TP + FP}$	$R = \frac{TP}{TP + FN}$
$AP = \int_0^1 P(R)dR$	$mAP = mean\ (AP)$

The training time varied from ~25 minutes on a 30 epochs range up to ~82 minutes with a 150 epochs range. However, it can be considered faster than the time recorded with YOLOv8 which resulted in ~72 minutes for 120 epochs.

3.4 System Implementation

DETR was initialized with pre-trained weights on the COCO dataset. This approach ensures a valid starting

point since COCO is large and diverse enough to leverage rich feature representations.

The pre-trained weights were loaded into the DETR architecture, focusing on adapting the final layers to our specific task of person detection.

The training configuration included specifying hyperparameters such as image size (640×640), batch size (16), learning rate, and number of epochs (30, 50, 70, 120, 150). The model was trained using the Adam optimizer, known for its adaptive learning rate capabilities, which helps to achieve faster convergence. After each training session (30, 50, 70, 120, and 150 epochs), the model was evaluated on the validation set to assess its performance.

The final trained model was tested on the holdout test set to obtain unbiased performance metrics.

The implementation of the fusion mechanism went through multiple phases (Figure 3) and is described in more detail below. Fusion Mechanism Implementation based on previously fine-tuned DETR and ViT.

- Confidence Score Calculation: After receiving detection outputs (bounding boxes and confidence scores) from both models, these scores are normalized.
- Weight Assignment: Implement the adaptive weighting system where each model's confidence score is multiplied by a predetermined but adjustable fusion factor. This factor might be dynamically tuned based on ongoing performance assessments or environmental context.
- Spatial and Semantic Analysis: Calculate the IoU for bounding boxes that overlap across models. Combine boxes with high IoU and similar class labels by averaging their positions weighted by their confidence scores.
- Application of Modified NMS: Apply a version of NMS that considers the fusion confidence scores to resolve conflicts between overlapping boxes, ensuring that each object is detected only once with the highest possible accuracy.

The following parallel Processing Pipeline steps have been used:

- Input Handling: Configure an input pipeline that preprocesses images from UAV cameras to match the input requirements of both models (e.g., resizing, normalization).
- Concurrent Model Invocation: Deploy both models synchronized to process the same input simultaneously. Use threading or asynchronous programming to manage parallel execution without bottlenecks.

Inference and Real-time Processing details:

- Batch Processing vs. Streaming: Depending on the application, implement the system to handle batch processing of collected images or real-time streaming of video feeds. In this case, processing batch images was tested.
- Decision Making: Based on combined confidence scores and the results of the NMS, finalize the detection output. This output includes the class (person), the location (bounding box), and the detection confidence.

Feedback and Continuous Learning details:

- The feedback mechanism allows the system's output to be periodically reviewed by human supervisors to tag inaccuracies.
- Using this feedback the fusion factor is adjusted as needed and refinement of the model parameters during scheduled re-training sessions, enhancing the system's accuracy and adaptability over time.

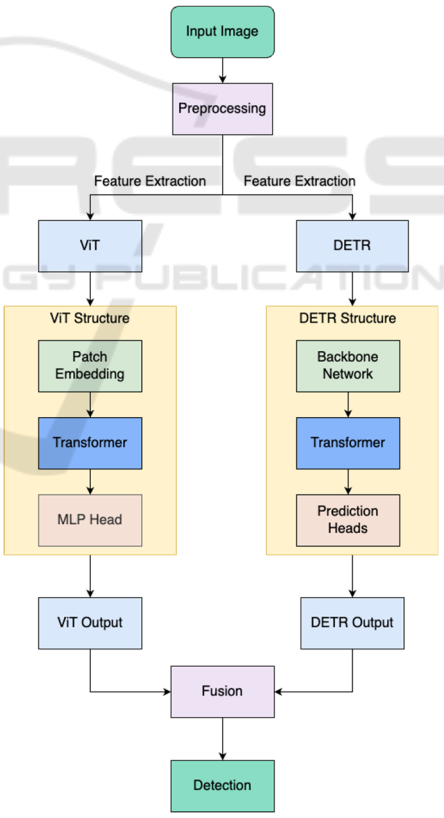

Figure 3: Global model architecture.

4 EXPERIMENTAL RESULTS AND DISCUSSIONS

The effectiveness of the combined DETR and ViT model was evaluated through a series of experiments using the VisDrone dataset, which consists of diverse urban and rural scenes captured via UAVs. The models were independently fine-tuned on this dataset before being deployed in parallel.

The paragraphs below describe the individual models' performance.

DETR achieved an accuracy of 84%, a mean Average Precision (mAP50) of 78%, and a recall of 73% (Table 3). DETR excelled particularly in densely populated urban scenes where precise localization of multiple objects is critical.

Table 3: Metrics obtained after validation.

Metric	DETR	ViT	Fusion
Precision	0.842	0.718	0.889
Recall	0.73	0.701	0.782
mAP50	0.78	0.643	0.826

ViT demonstrated an accuracy of 72%, a recall of 70%, and a mAP50 of 64%. Its performance was notably superior in detecting smaller and more distant objects, which models often miss because they rely heavily on localized contextual information.

The paragraphs below describe the combined model performance. The fusion mechanism implementation significantly enhanced overall performance, resulting in an accuracy of 90%, a mAP50 of 85%, and a recall of 80%. The improvement was particularly notable in complex environments where individual models showed limitations:

In urban settings characterized by high object density and diverse object scales, the combined model improved detection accuracy by up to ~15% on average compared to individual models.

In adverse weather conditions, which typically impede visual clarity and object recognition (mostly during the night with low or very low light and during daylight within dusty areas), the fused model demonstrated resilience, maintaining high accuracy rates that surpassed each model operating independently by a significant margin.

The results in Figure 4 underline the potential of integrating multiple transformer-based models to handle the complexities of UAV-based object detection.

The dynamic fusion approach not only leveraged the unique strengths of DETR and ViT but also facilitated a robust performance across varied and challenging environments, showcasing the practical implications of this research in real-world applications.

Our proposed method combines the strengths of both DETR and ViT through a fusion mechanism that leverages their complementary features. By running both models in parallel and integrating their outputs using adaptive weighting and modified NMS, our approach addresses the shortcomings of individual models. The fusion mechanism enhances the detection of small and distant people while maintaining high precision and recall rates.

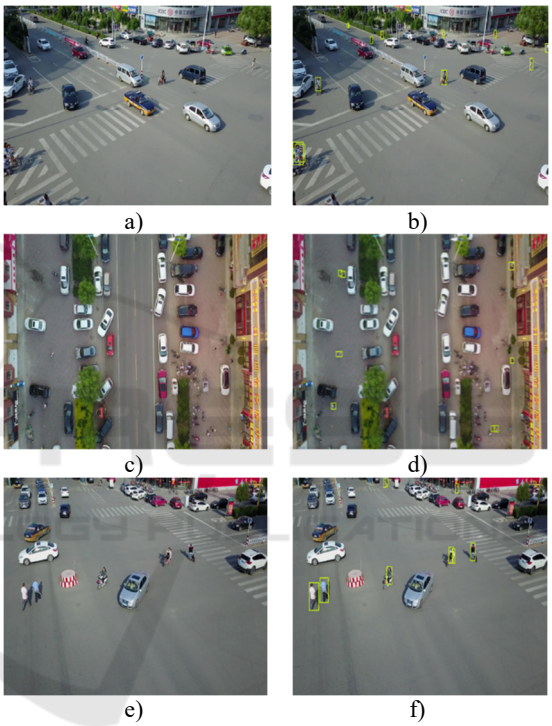

Figure 4: Examples of experimental results from the fusion model, a), c), e) - test images, b), d), f) - processed images.

Compared to traditional CNN-based detectors, the proposed fusion model demonstrates better performance in terms of precision, recall, and mAP50 metrics, as shown in Table 3. The adaptive weighting system allows the model to dynamically adjust to varying environmental conditions and object scales, which is not commonly addressed in other studies.

Moreover, while transformer-based models are still emerging in UAV imagery analysis, our approach showcases their potential when effectively combined. The fusion of DETR and ViT improves detection accuracy and maintains computational efficiency for real-time applications.

Our method distinguishes itself from existing ones by integrating feedback and continuous learning mechanisms. By incorporating human-in-the-loop feedback to adjust the fusion factors and refine model parameters, the system adapts over time, enhancing its robustness and applicability in diverse operational scenarios.

Our implementation offers a novel solution that outperforms existing methods by effectively addressing the unique challenges of person detection and tracking in UAV imagery. Combining advanced transformer architectures with an adaptive fusion mechanism presents a significant step forward in developing reliable and efficient UAV-based detection systems.

5 CONCLUSIONS

DETR and ViT integration through the described fusion mechanism has proven a promising solution for enhancing object detection capabilities in UAV operations. This study highlights the complementary strengths of the two transformer models and paves the way for research in future papers on multi-model fusion strategies. Future studies may explore adaptive algorithms that could refine the fusion process based on continuous learning from diverse environmental data and real-time operational feedback.

REFERENCES

Bodla, N., Singh, B., Chellappa, R., Davis, L. (2017). Improving object detection with one line of code, *2017 IEEE International Conference on Computer Vision* (*ICCV*).

Du, D. et al. (2020). VisDrone: The vision meets drone object detection in image, *IEEE/CVF International Conference on Computer Vision* (*ICCV*).

Huang, J., Tianrui Li, T. (2024). Small object detection by DETR via information augmentation and adaptive feature fusion. *Proceedings of 2024 ACM ICMR Workshop on Multimodal Video Retrieval*.

Jiang, P., Ergu, D., Liu, F., Cai, Y., Ma, B. (2022). A review of Yolo algorithm developments. *Procedia Computer Science*, vol. 199, 2022, pp. 1066-1073.

Song, H., Sun, D., Chun, S., Jampani, V., Han, D., Heo, B., Kim, W., Yang, M.H. (2021). ViDT: An efficient and effective fully transformer-based object detector, 2021, arXiv:2110.03921.

Stan, A.S., Ichim, L., Parvu, V.P., Popescu, D. (2023). Person detection and tracking using UAV and neural networks. *31st Mediterranean Conference on Control and Automation* (*MED*), Limassol, Cyprus.

Wang, L., Tien, A. (2023) Vision transformer with dynamic position embeddings for object detection in aerial images. *IEEE International Geoscience and Remote Sensing Symposium* (*IGARSS*), Pasadena, CA, USA.

Wu, X., Li, W., Hong, D., Tao, R., Qian Du, Q. (2021). Deep learning for UAV-based object detection and tracking: a survey. *IEEE Geoscience and Remote Sensing Magazine*.

Ye, T., Qin, W., Zhao, Z., Gao, X., Deng, X., Yu Ouyang, Y. (2023). Real-time object detection network in UAV-vision based on CNN and Transformer, *IEEE Transactions on Instrumentation and Measurement*.

Zeng, F., Dong, B., Zhang, Y., Wang, T., Zhang, X., Wei, Y. (2022). MOTR: End-to-end multiple-object tracking with Transformer. *Proceedings 17th European Conference* (*ECCV*), Part XXVII.

Zhang, J. (2023). Towards a high-performance object detector: insights from drone detection using ViT and CNN-based deep learning models, 2023 *International Conference on Computer Vision and Robotics Science*.