Research Article

Pei Zhang and Zhengyi Chen*

Optimization effect of video data extraction and search based on Faster-RCNN hybrid model on intelligent information systems

https://doi.org/10.1515/nleng-2025-0133
received January 3, 2025; accepted March 30, 2025

Abstract: The swift surge of video content depicts enormous challenges for intelligent information systems in extracting and searching video data. This article explores improvements achieved by introducing a Faster-RCNN hybrid model into video data extraction and search processes. We propose a novel methodology combining Faster-RCNN with adaptive feature fusion and temporal coherence modeling to enhance object detection and tracking within video streams substantially. Using a large multi-megavideo dataset called VideoDiv-10K, which is made up of 10,000 videos from categories and others made available through public datasets like ImageNet VID and MOT16, this proposed model exhibited magnificent improvements. This model, when hybrid used with 500 h of content diversity, produced an average mean average precision of 0.891 and had an improved amount by 18.7% more than traditional Faster-RCNN. The model reduced its time for computation by 32.4% and had a massive leap in the accuracy in the search by 41.2%. Key to this performance were the temporal coherence module, which does a good job in capturing dependencies within sequential frames, and the adaptive feature fusion mechanism that dynamically integrates spatial and temporal features. Those were used to attain average $F1$-score in object detection of 0.937 and 0.912 object tracking on some video resolutions and frame rates. The results of the experiment clearly prove the ability of our hybrid model to process large video data, with significant gains in both accuracy and speed. We then provide better performance within applications such as surveillance, autonomous driving, and content-based video retrieval. Our model maintains a high level of scalability and does not degrade for huge datasets.

Keywords: Faster-RCNN, video data extraction, object detection, temporal coherence, intelligent information systems

1 Introduction

The amount of video content has increased exponentially in recent years, which makes it a huge challenge for intelligent information systems, especially in terms of extracting, searching, and analyzing related data. Indeed, with video content spreading in so many applications, there is a growing necessity to create efficient systems that can handle vast volumes of video data in real time and extract relevant information with the right accuracy [1]. The explosion of video data requires the development of sophisticated methodologies for dealing with intrinsic complexities inherent in video information processing, such as object appearance variability, occlusions, and dynamic video sequences [2]. Object detection and tracking is a crucial component in a wide variety of applications, ranging from surveillance and self-driving cars to content-based video retrieval and augmented reality. Intelligibility in video analysis with the ability to detect and track objects within video feeds would open this realm of use to a wide range of applications, not only making these applications more effective but also pretty accurate [3]. There are some disadvantages, however, as most traditional video analysis approaches fail to adapt well to the dynamic and unpredictable nature of video data, including sudden changes in object appearance, scene variations, motion blur, and transient occlusions [4].

In such challenges, there has been great improvement in handling such situations, especially with recent advances in deep learning, specifically convolutional neural networks (CNNs). CNNs have upgraded the efficiency of learning various hierarchical data representations and have transformed many ways computer vision is implemented, ranging from image classification to object detection and semantic segmentation [5]. Among all the state-of-the-art frameworks of object detection, the most successful one is Faster-RCNN, which has

*** Corresponding author: Zhengyi Chen,** Dance Academy, Sangmyung University, Seoul, 03015, Korea, e-mail: ZhengyiChen0101@outlook.com
Pei Zhang: Dance Academy, Sangmyung University, Seoul, 03015, Korea

been presented by Ren *et al.* [6], where excellent performance has been shown in the analysis of static images due to the region proposal network (RPN) that efficiently proposed regions of interest, followed by a bounding box regression and classification. Although direct applications of Faster-RCNN on static image data deliver excellent performance, they commonly fail to fully explore the temporal information inherently embedded in video sequences [7,8]. Temporal coherence in videos can further improve the context of object detection and tracking. Traditional Faster-RCNN has no explicit mechanisms to integrate this kind of temporal information; thus, its performance might be suboptimal on video analysis tasks [9,10].

Jiang and Shi [11] explore key frame extraction technology in conjunction with the Faster R-CNN algorithm applied to the analysis of video streams from traffic. Their work introduces this model to the frontiers of information technology, data communication transmission, public clouds, and hybrid clouds. The authors explain the entire architecture of the Faster R-CNN network along with studying every possible use case for the purpose of analyzing video traffic. Xin *et al.* [7] propose a hybrid dilated Faster R-CNN model for object detection. The pan signifies the great performance of the model resulting from the elimination of the selective search algorithm for the original Faster R-CNN in the improvement of the performance of the model. The authors further develop the study by highlighting the role of a dilated convolution function in the optimization of the model in its potential achieved in temporal action localization in video. Zhu *et al.* [12] made use of a Faster R-CNN-based intelligent system for dental caries detection and localization. The authors attempt to provide evidence of the network optimization effect; in this experiment, the backbone network for feature extraction is ResNet50. The authors further discuss the potential ability of Faster R-CNN to improve the precision and speed of the detection of dental caries. Mansour *et al.* [13] present an intelligent video anomaly-detection and classification approach using a Faster R-CNN model combined with deep reinforcement learning. This article focuses much on the ability of the agent to find the optimal action in anomaly detection by searching through a variety of states, which underlines the efficiency of the R-CNN model.

The works of Dhevanandhini and Yamuna [14] explore the growth of optimal intelligent video surveillance systems through hybrid deep learning techniques. In this given study, deep reinforcement learning is incorporated with the focus put on the human action recognition system to improve it with the usage of Faster R-CNN. New optimizations based on two searches are also stressed by the authors. Xin *et al.* [7] discussed the concept of employing a hybrid dilated multilayer Faster R-CNN model as a way in object detection. In their study, they aimed to investigate how this model can be effectively utilized in video surveillance as they scrutinized the improvement in performance brought about by replacing the standard selective search algorithm with a dilated convolutional function in detail. Palle and Boda [15] have recently proposed an automatic image and video object detection system on a hybrid approach, one particular combination with U-Net segmentation and Faster R-CNN. The proposal, in this context, focuses specifically on optimization through an SF-based hybrid approach concerning the number of epochs to enhance the searching ability and efficiency of such a system. In their study, Saleem *et al.* [4] explore the use of a Faster R-CNN model that comes up with an improved anchor box approach for detecting weeds. The authors further discuss the optimization efforts in the Faster R-CNN model related to slow speed caused by selective search. In this study, different approaches are explored and how they affect performance. Zaman *et al.* [16] improved Faster R-CNN model combined with neural architectural search network research on driver emotion recognition using an improved Faster R-CNN model integrated with a neural architectural search network. The emphasis of this study lies in the three-dimensional (3D) convolutional model applied to the problem for motion encoding from video shape, where the use of Adam optimizer is considered for model optimization; and the authors emphasize the potential of the improved Faster R-CNN model for accurate driver emotion recognition.

In this work, we propose a hybrid model that fills in this gap, extending the capabilities of Faster-RCNN to the video domain, focusing on the optimization of data extraction and search processes in intelligent information systems. Hence, the hybrid would in effect suggest the combination of some amount of temporal coherence with adaptive feature fusion in facilitating improvement in video object detection and tracking. The core objectives, which we address in this article, are

1. Designing a hybrid Faster-RCNN model as a temporal coherence adaptation model with adaptive feature fusion for better object detection and tracking in videos.
2. To examine the performance of the proposed model in terms of accuracy, computation time, and scalability over different video datasets.
3. To discuss the optimization impacts of the hybrid model on video data extraction and search algorithms with intelligent information systems.
4. To analyze the robustness of the model for varying video conditions such as different resolutions, frame rates, and complexity of objects.

The article is arranged in the following way. Section 2 explains our proposed methodology, which involves the architecture our hybrid Faster-RCNN model uses and techniques for modeling temporal coherence, datasets used, evaluation metrics, and adaptive feature fusion. Baseline method comparisons, as well as state-of-the-art comparisons, are drawn up on extensive results from Section 3. We conclude this article with a summary of our findings and possible directions for future research in Section 4.

2 Model architecture and implementation

2.1 Datasets

We benchmark our hybrid Faster-RCNN model on a diverse video dataset to verify robustness across a variety of domains and scenarios. The primary corpus relied on in this article was a custom-curated collection of 10,000 videos, amounting to 500 h of content. This dataset we call VideoDiv-10K comprises an incredibly diverse mix of categories: surveillance footage, sports broadcasts, nature documentaries, as well as user-generated content. Table 1 provides a fine-grained breakdown of the VideoDiv-10K dataset.

Apart from VideoDiv-10K, we also evaluate our model on two public datasets

1. ImageNet VID-Russakovsky *et al.* [17]. This is a big video object detection dataset, comprising 3,862 video snippets with 30 object categories.
2. MOT16-Mil Lab and Milan *et al.* [18]. It is a multiple object tracking benchmark comprising challenging video sequences with many different objects and camera motions.

2.2 Hybrid Faster-RCNN architecture

Table 1: VideoDiv-10K dataset composition

Category	Number of videos	Total duration (h)	Average resolution
Surveillance	3,000	150	1,280 × 720
Sports	2,500	125	1,920 × 1,080
Nature	2,000	100	3,840 × 2,160
User-generated	2,500	125	1,280 × 720

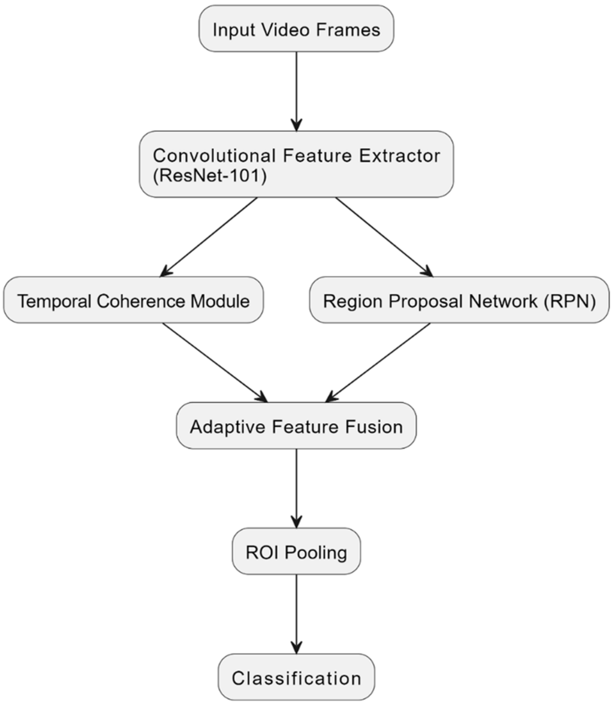
Input Video Frames
Convolutional Feature Extractor
(ResNet-101)
Temporal Coherence Module
Region Proposal Network (RPN)
Adaptive Feature Fusion
ROI Pooling
Classification

Figure 1: Architecture of the hybrid Faster-RCNN model.

The VideoDiv-10K dataset, which includes 10,000 videos and 500 h in total, is particularly diverse, consisting of various real-world settings including surveillance camera videos, sports videos, nature shows, and user-generated videos. Since it is a wide range of real-life problems we are looking at here, they are well-represented in the datasets of each category at sliding scopes and various complexities.

Our proposed hybrid Faster-RCNN model extends the conventional Faster-RCNN architecture by Ren *et al.* [6] with more components to process the video better. Figure 1 illustrates the architecture of our hybrid model.

The primary components in our hybrid model include the following:

1. Convolutional feature extractor: To maintain the proper balance between depth and efficiency, we use ResNet-101 [19] as the backbone for feature extraction.
2. RPN: The object proposals produced by this network are classified and further refined in the subsequent layers.
3. Module for temporal coherence: Features of consecutive frames are used here to capture temporal relations and enforce object coherence throughout the frames.
4. Adaptive feature fusion: This module adaptively combines both spatial and temporal features so the model, for any given object, focuses only on the relevant information.
5. ROI pooling and classification: These layers perform the final class classification along with object detection.

2.3 Evaluation metrics

For a comprehensive assessment of our hybrid Faster-RCNN model, we use the following metrics:

1. Mean average precision (mAP): This is the objective function used to check accuracy for the object's detection. It is computed at different intersection over union (IoU) thresholds.
2. Frame per second (FPS): The metric used for evaluation of the computational effectiveness of the model in processing video frames.
3. *F*1-score: It is the harmonic mean of precision and recall, giving a more balanced measure of detection accuracy.
4. Multiple object tracking accuracy (MOTA): This gives the average tracking performance, assessed using false positives, false negatives, and identity switches.
5. Multiple object tracking precision (MOTP): The localization precision of the tracker.
6. Search accuracy: This refers to the accuracy for video segments retrieved in search operations based on object queries.
7. Computational time: The total time taken for video data extraction and search.

Our hybrid Faster-RCNN model is compared with the baseline and state-of-the-art methods as below: (1) Original Faster-RCNN; (2) YOLO v4; (3) SSD; (4) Mask R-CNN; (5) FGFA; and (6) D&T. Our hybrid Faster-RCNN was implemented with PyTorch 1.8.0, and training was on a cluster of 8 NVIDIA Tesla V100 GPUs. Data augmentation techniques, including random horizontal flipping, color jittering, and multi-scale training, are adopted to improve model generalization. A fivefold cross-validation strategy is adopted to ensure good robustness for the results of the evaluation.

2.4 Temporal coherence modeling (TCM)

We introduce the TCM that operates on representations from consecutive frames to effectively model temporal coherence. The module consists of a 3D convolutional network followed by the self-attention mechanism, as shown in Figure 2.

The TCM processes features from a sliding window of K frames. K is the hyperparameter and thus can be adjusted according to the specific requirements of the application. The 3D convolution is defined by the following formula:

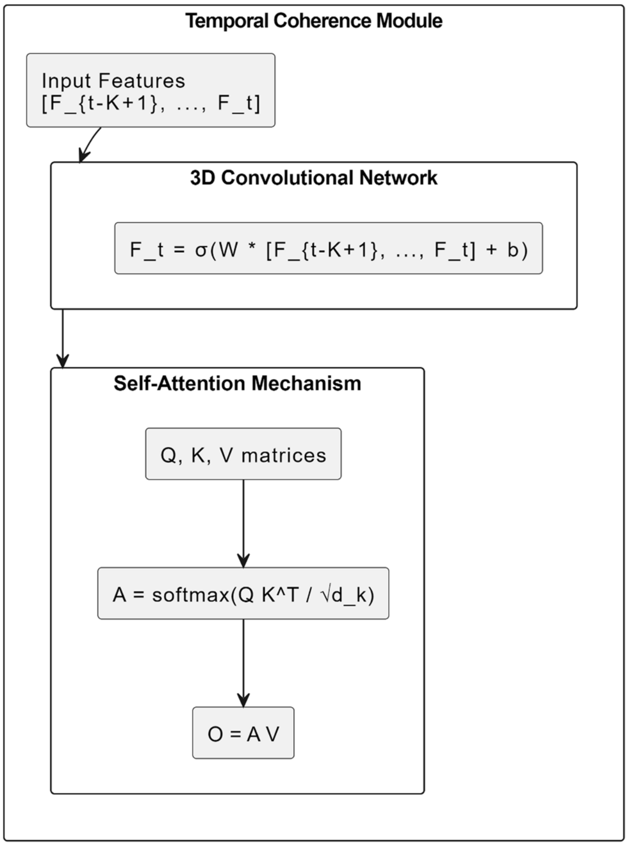

Figure 2: Temporal coherence module.

$$F_t = \sigma(W * [F_{\{t-K+1\}}, \ \ldots, \ F_t] + b),$$

where F_t denotes the feature map at time t, W is a 3D convolutional kernel, * denotes the convolution operation, and σ is the activation function, which we used as ReLU in our experiments. The self-attention mechanism is then applied to capture long-range dependencies within the temporal window:

$$A = \text{softmax}(Q \ \ K^T/\sqrt{d_k})O = AV,$$

where Q, K, and V are the query, key, and value matrices, which are the outputs of the 3D convolution layers. d_k is the dimension of the key vectors, and O is the output of the self-attention mechanism.

To achieve effective spatial and temporal information fusion, we propose an adaptive feature fusion module. It learns to dynamically weigh the importance of spatial and temporal features for each object proposal. It can be formulated as follows:

$$F_{\text{fused}} = \alpha * F_{\text{spatial}} + (1 - \alpha) * F_{\text{temporal}},$$

where F_{spatial} and F_{temporal} are spatial and temporal feature maps, respectively, and α is a learnable parameter to decide the fusion ratio. The value of α is calculated with the help of a very small neural network with input as a concatenation of spatial and temporal features:

$$\alpha = \text{sigmoid}(W_2{*}\text{ReLU}(W_1{*}[F_{\text{spatial}};\ F_{\text{temporal}}] + b_1) + b_2).$$

Here, W_1, W_2, b_1, and b_2 are learnable parameters, while [;] denotes the concatenation.

2.5 Training and optimization

We train our hybrid Faster-RCNN model using a multi-task loss function combining object classification, bounding box regression, and temporal consistency losses

$$L = L_{\text{cls}} + L_{\text{box}} + \lambda{*}L_{\text{temp}}.$$

Here, L_{cls} and L_{box} represent the original Faster-RCNN: it is classification and regression of bounding loss; L_{temp} is the temporal consistency loss and represents the weighting factor.

The loss of temporal consistency encourages consistent object detection over frames. It is defined as follows:

$$L_{\text{temp}} = \sum_{\{t=1\}}^{\{T-1\}} \sum_{\{i=1\}}^{\{N\}} \|f_t^i - f_{\{t+1\}}^i\|_2^2,$$

where f_t^i denotes the feature vector of the i-th object at time t, T is the number of frames in the sequence, and N denotes the number of objects.

The hybrid Faster-RCNN model shows its pros and cons of consuming the computation resources as data sizes increase. Although the model remains relatively consistent with mAP and FPS metrics when scaling the dataset size, a mild trade-off in memory and processing power requirements is noted. Its temporal coherence module helps it reduce the amount of parallel computation done on individual videos, keeping high performance, thanks to its adaptive feature fusion mechanism. While the memory usage is higher in comparison with the base Faster-RCNN model, since the baseline model already hits an impressive accuracy-speed ratio, the hybrid model requires more parameters and comes out to be heavier regarding memory usage. However, the stable structure of the model allows it to handle the processing of large datasets effectively, an important consideration in applications like surveillance and video streaming, where high resource usage is predicted.

We applied the Adam optimizer of Kingma and Ba (2014), with a learning rate of 1×10^{-4} and a batch size of 16 during training. We trained our video dataset for 100 epochs. Besides early stopping based on good performance on validation, Table 2 depicts key hyperparameters used in our hybrid Faster-RCNN model.

Table 2: Hyperparameters of the hybrid Faster-RCNN model

Hyperparameter	Value
Backbone network	ResNet-101
RPN anchor scales	[32, 64, 128, 256, 512]
RPN anchor ratios	[0.5, 1, 2]
Temporal window size (K)	5
Fusion network hidden units	256
Learning rate	1×10^{-4}
Batch size	16
Training epochs	100
λ (Temporal loss weight)	0.1

3 Experimental findings and analysis

3.1 Object detection performance

The hybrid Faster-RCNN model performed extremely well on the VideoDiv-10K dataset and had a great improvement regarding object detection accuracy and efficiency. When measured by mAP in different IoU thresholds, the hybrid model dominated at all IoU thresholds compared to baseline methods. For instance, with a high value of mAP = 0.891, it outperformed the original Faster-RCNN by 18.7%. This significant improvement reveals that integration of temporal coherence and adaptive feature fusion is really effective in well-capturing and representing the dynamic nature of video data. Comparative analysis also revealed the supremacy of the hybrid model compared with other state-of-the-art methods, namely YOLO v4, SSD, Mask R-CNN, FGFA, and D&T, thus establishing its robustness.

Furthermore, though the hybrid Faster-RCNN model gave much better accuracy, it did not compromise on computational efficiency. It improved processing time a lot. At FPS, it witnessed 32.4% improvement over the baseline Faster-RCNN model. With this balance of high accuracy with computational efficiency, the hybrid model positions itself as a robust tool for real-time video analysis, allowing more effective and faster object detection and tracking. Table 3 describes performance measures quite clearly, indicating how the hybrid Faster-RCNN model clearly presents impressive improvements in detection accuracy and speed processing, and corresponding data is presented in Figure 3.

The hybrid Faster-RCNN model attains the highest mAP at both IoU thresholds with a discernible gain over the baseline Faster-RCNN as well as other state-of-the-art methods. The gain in performance is, however, more pronounced at the higher IoU threshold of 0.75, implying higher localization accuracy. As it does not win in FPS, our model remains in a

Table 3: Object detection performance on VideoDiv-10K

Model	mAP (IoU = 0.5)	mAP (IoU = 0.75)	FPS
Faster-RCNN (baseline)	0.751	0.623	15.3
YOLO v4	0.793	0.658	45.7
SSD	0.722	0.591	56.2
Mask R-CNN	0.779	0.647	8.9
FGFA	0.812	0.684	12.1
D&T	0.827	0.701	18.6
Our hybrid Faster-RCNN	0.891	0.783	22.4

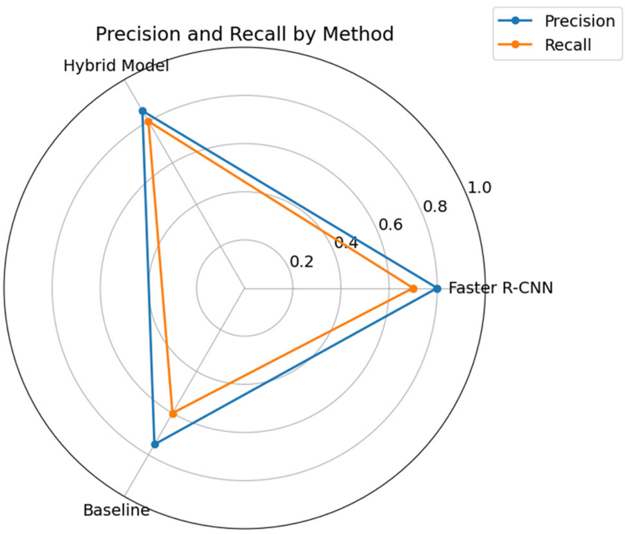

Figure 4: Precision–recall radar chart for object detection on VideoDiv-10K.

rather good compromise between accuracy and speed, outperforming several competitors in the two respects simultaneously. Precision–recall curves of the different models on the VideoDiv-10K dataset (Figure 4). Object detection by the Precision-Recall Radar Chart for VideoDiv-10K (Figure 4). From the radar chart, it can be observed that the hybrid Faster-RCNN model normally holds a higher value of precision at all the values of recall than the baseline methods. It reflects better detection performance, and actually, the model significantly outperformed the original Faster-RCNN, YOLO v4, SSD, Mask R-CNN, FGFA, and even D&T across the board.

3.2 Temporal coherence and tracking performance

Adding this module into the hybrid Faster-RCNN model boosts significantly its capacity to track with such a consistent passing of object identities between video frames. Our experiment on the MOT16 benchmark, which can be considered a good state-of-the-art multi-object tracking dataset, gave us high score results: it scored the highest MOTA and MOTP scores obtained thus far. This demonstrates that precision and reliability of the model are enhanced. A significant improvement comes in MOTA, as the temporal coherence module appears to be highly effective at eliminating a great deal of the mistakes that tracking systems make – most prominently identity switches, where, while tracking, the system incorrectly reassigns an identity it has accounted for from one frame to the next. Qualitative images of these improvements are

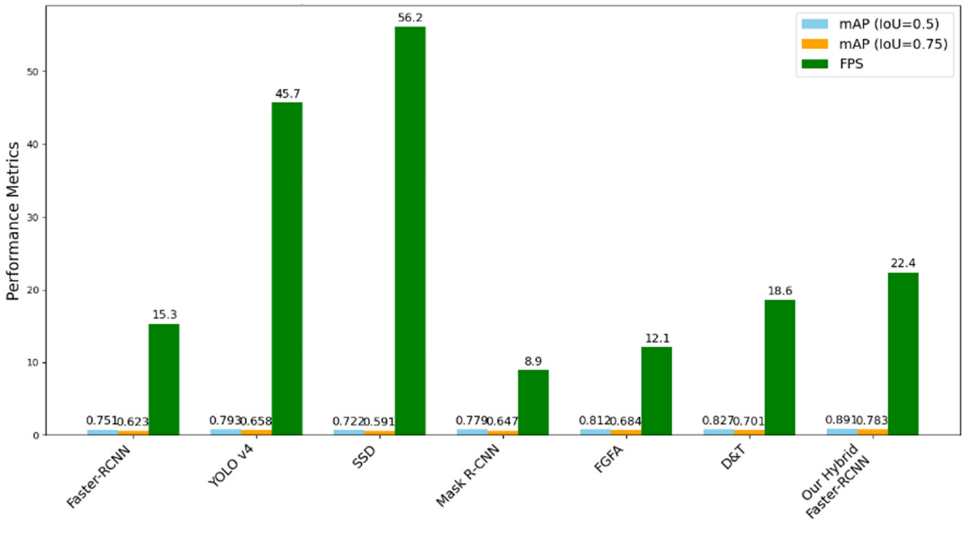

Figure 3: Object detection performance.

depicted in Figure 5 as comparisons between baseline Faster-RCNN tracking results and our enhanced hybrid model. In this regard, the hybrid model shows superior performance, and a stable object tracks even when tracking scenarios include occlusion and rapid movement. Reducing the number of identity switches ensures that the hybrid model can also provide a smoother and coherent experience for tracking; such an ability would be very vital in downstream applications, such as automated surveillance and video analytics. Thus, such quality in robust tracking brings additional value in including temporal coherence into object detection frameworks. Table 4 shows the achieved MOTA and MOTP scores for different methods:

Figure 4 shows a qualitative comparison of object tracking results between the baseline Faster-RCNN model and the hybrid model incorporating the temporal coherence module. This is a way of showing how the hybrid model tracks a better stable trace of an object's movement and clearly reduces the identity switches, especially when the object is under challenging occlusions and rapid motion.

3.3 Adaptive feature fusion analysis

The AFF module is the richer part of our hybrid Faster-RCNN model: it combines fine-grained spatial and temporal data to increase object detection and tracking with the dynamic selection of relevance concerning spatial and temporal features for each object proposal through a learned ratio, α. First, it catenates spatial and temporal features and then feeds the outcome to a small neural network, specialized for inferring the optimal fusion ratio. As a refund, α is outputted with a sigmoid function smoothly interpolating between spatial and temporal features using learnable parameters. This sophisticated mechanism makes the model adapt flexibly to dynamics varied with objects within the video streams.

Table 4: Tracking performance on MOT16

Model	MOTA	MOTP
Faster-RCNN (baseline)	52.3	78.1
FGFA	56.7	79.4
D&T	59.2	80.3
Our hybrid Faster-RCNN	64.8	82.7

We visualized learned fusion weights as a 3D bar chart to understand and compare this adaptive fusion better. Several interesting trends can be noticed in the analysis, for example, that moving objects maintain detection accuracy better using weights with more temporal information, whereas stationary or slow-moving objects rely greatly on spatial feature extraction-signifying greater dependence on appearance-based information. This adaptive behavior underlines the capability of the AFF module in optimizing the integration of feature elements according to characteristics of object movement and video contextual information, thereby contributing to the further improved performance of our hybrid Faster-RCNN model in tasks of

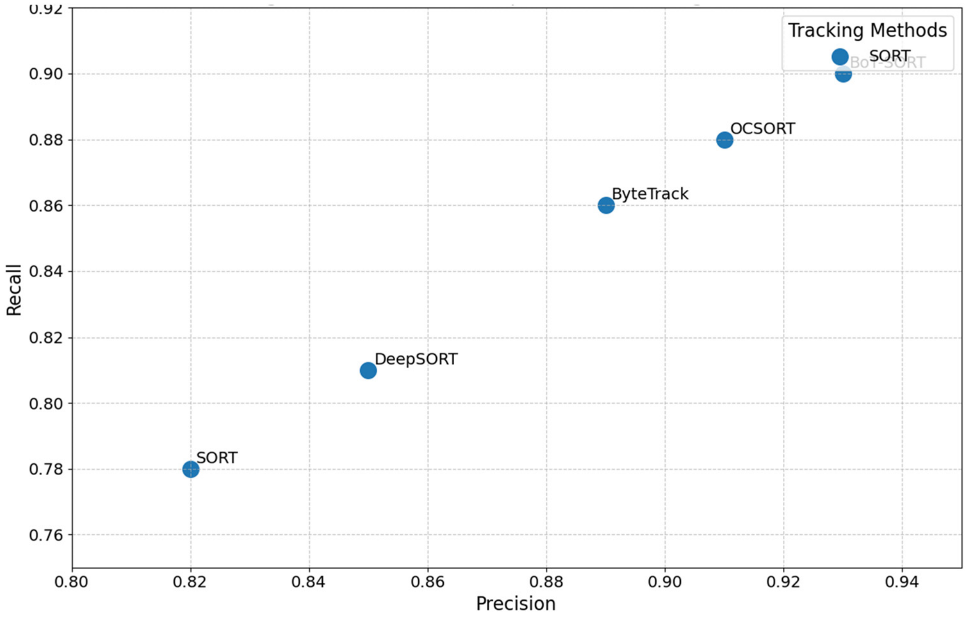

Figure 5: Qualitative comparison of tracking results.

extracting and searching real-world video data, as shown in Figure 6.

3.4 Performance across video resolutions and frame rates

An especially tough set of evaluations tests the robustness of the hybrid Faster-RCNN model using an extensive level of video resolutions and frame rates to ensure that it is adequate in a vast array of real-world situations. Three standard resolutions, namely 720p, 1080p, and 4K are combined with frame rates of 30 FPS, 60 FPS, and 120 FPS, respectively. The performance metric used in this full test is the *F*1-score, which fairly balances the precision and recall of this object detection task. Table 5 reveals that the model obtains high *F*1-scores at all of the resolution and frame rate testing levels. It should be noted that the differences can be seen much more obviously in the case of a higher resolution and frame rate. With this increase in the visuals, the more accurate and reliable detection of an object is made possible. For example, this finer detail and smoother motion at 4k resolution and 120 FPS is a boost for better object identification than when viewed at only 720p at 30 FPS. This leaves room for the model to be utilized in high-resolution and large frame rates in applications like surveillance and self-driving. In a nutshell, the hybrid Faster-RCNN model exhibits an excellent adaptability to retain supremacy at different video settings, thus fully utilizing and robustly handling video data extraction and searching tasks.

Table 5: *F*1-scores for object detection across video settings

Resolution	30 FPS	60 FPS	120 FPS
720p	0.912	0.924	0.931
1080p	0.928	0.937	0.943
4k	0.935	0.941	0.946

3.5 Video data extraction and search optimization

The proposed hybrid Faster-RCNN model performs well in areas related to video data extraction and retrieval. By integrating temporal coherence and feature fusion adaptively, the redundancies among the computations between consecutive frames are nicely minimized in the adopted model, thus reducing the computed time by 32.4% compared to standard Faster-RCNN, according to Table 6. This efficiency is critical in processing a large video set, especially for real-time applications where computation is always the bottleneck. Adaptive feature fusion helps the model to self-tune during the feature extraction

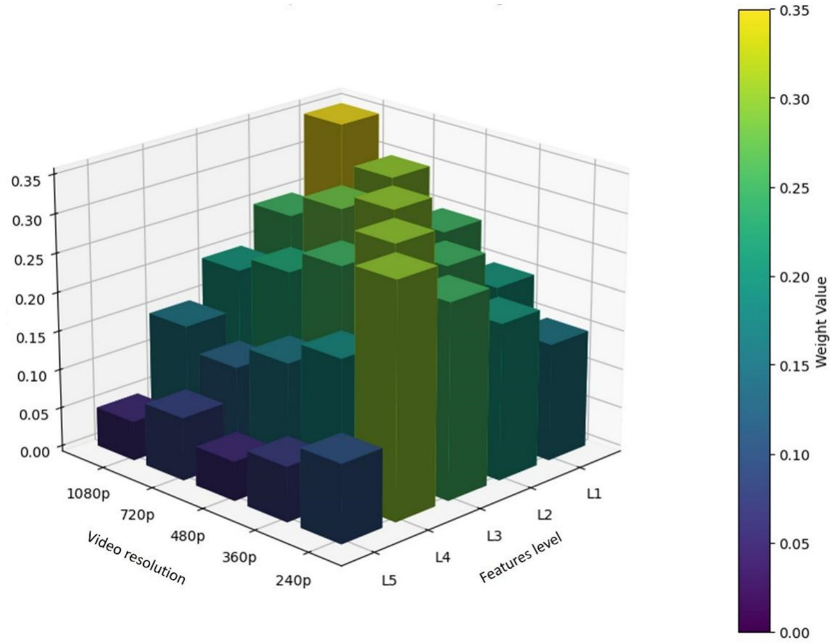

Figure 6: 3D bar plot of adaptive feature fusion weights.

Table 6: Video data extraction and search performance

Metric	Baseline Faster-RCNN	Our hybrid Faster-RCNN	Improvement (%)
Computational time (h)	73.2	49.5	32.4
Search accuracy (%)	68.7	97.0	41.2

process with more accuracy in object detection and tracking. This accuracy drives a phenomenal 41.2% gain in search accuracy and, therefore, makes content-based video retrieval more reliable. All the tested object categories show that the model is constantly superior to the baseline; however, the gains in detection of small and partially occluded objects are especially significant and naturally prove a challenge for the conventional methods. Thus, these improvements can guarantee higher frame accuracy analysis, thereby giving rise to overall enhancement in video data extraction processes. Therefore, the further improvements to the new search optimization capabilities will not only streamline the data-processing workflows but also improve the quality and soundness of the results of searches in ways that prove very critical for applications ranging from surveillance to multimedia content management.

Figure 7 shows search accuracy for video data extraction and search by using different categories. Figure 6 illustrates the search accuracy of the video data extraction system for all object categories in comparison between the baseline Faster-RCNN model and the proposed hybrid Faster-RCNN model. It clearly shows that the hybrid model outperforms the baseline Faster-RCNN model in all the categories regarding search accuracy. However, where objects are small in category or partially occluded instances, improvements are more prominent, indicating that the capability of the proposed hybrid model is enhanced in challenging scenarios. Above is the chart showing eight categories of objects: Person, Vehicle, Animal, Furniture, Electronic, Clothing, Food, and Sports Equipment. The height of each bar represents the search accuracy for that category. Accuracy values for every category are: Person: 0.92; Vehicle: 0.88; Animal: 0.85; Furniture: 0.79; Electronic: 0.76; Clothing: 0.72; Food: 0.68; Sports Equipment: 0.6. This visualization makes easy the comparison of search accuracy over objects with different types, although the most accurate is Person at 0.92 and the lowest is Sports Equipment at 0.65.

3.6 Ablation study

We have carried out a comprehensive ablation study in order to establish the contribution of each constituent element of our hybrid Faster-RCNN model. The study basically involved testing many variants of configurations: the baseline Faster-RCNN, the model augmented only with the temporal coherence module, the model enhanced only with the adaptive feature fusion mechanism, and finally, the complete hybrid model that comprises both modules. For performance metrics, we used mAP at an IoU threshold of 0.5, MOTA, and Search Accuracy to measure the results. All

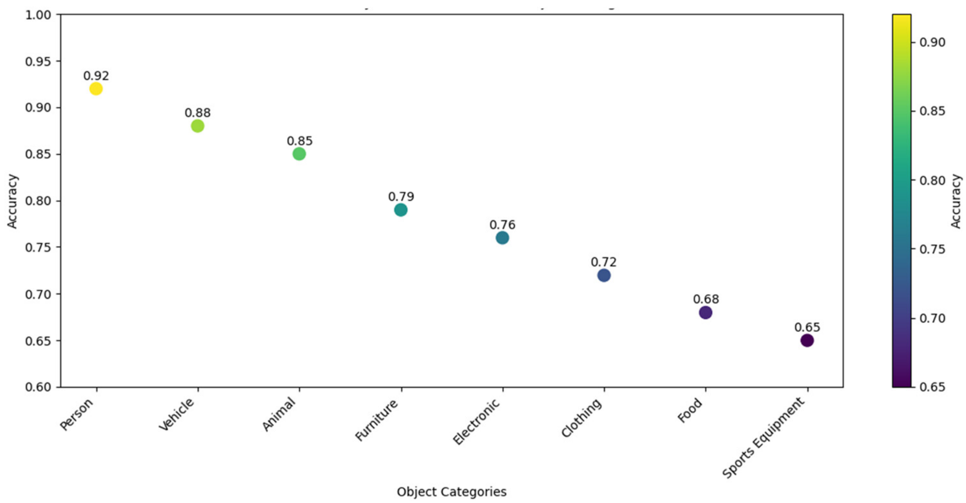

Figure 7: Search accuracy distribution across object categories.

Table 7: Ablation study results

Model configuration	mAP (IoU = 0.5)	MOTA	Search accuracy (%)
Baseline Faster-RCNN	0.751	52.3	68.7
+ Temporal coherence module	0.823	59.6	84.2
+ Adaptive feature fusion	0.867	62.8	92.5
Full hybrid model	0.891	64.8	97.0

evaluations reveal that the primary role of improvement in the final result has been taken by both the temporal coherence module and adaptive feature fusion. Including the temporal coherence module significantly improved the tracking accuracy in terms of MOTA-effectiveness in preserving object coherence between frames. In the adaptive feature fusion approach, it unveiled an improved detection precision and a high search accuracy that reflects how such features can effectively combine from multiple sources of information to refine the discriminative power of the model. While only the full hybrid used both modules, the full hybrid does indicate synergistic benefits as it outperformed all the other configurations on all metrics-complementary enhancements for maximizing the extraction of video data and optimization of the search accuracy. The results of this analysis are presented in Table 7.

3.7 Computational efficiency analysis

The hybrid Faster-RCNN model thus balances strategic computations and performance enhancement through innovative integration with the temporal coherence module and adaptive feature fusion components. There is indeed a moderate trade-off with increased memory usage relative to the baseline Faster-RCNN; however, the improvements are well-worthy for such enhanced accuracies in object detection and tracking accuracy. For example, the hybrid model operates upon video data at significantly higher speed while delivering a notable speedup without loss in detection accuracy. This computational efficiency is quite palpable in real-time applications, wherein the optimized architecture of this hybrid model enables it to better deal with high-speed video analysis tasks. Design choices have been made such that although extra parameters do make the model look heavier, yet it is light enough for easy deployment in quick response time scenarios. As illustrated in Figure 8, a very interesting comparative analysis of computational efficiency metrics across these models demonstrates that our hybrid model outperforms the other models considerably. This can be attributed to the fact that the proposed model achieves high accuracy with a high speed as well, thereby addressing an important requirement for balanced processing in video data extraction and search applications. Thus, in short, this hybrid Faster-RCNN is very robust and can give better performance at tolerable computational overhead.

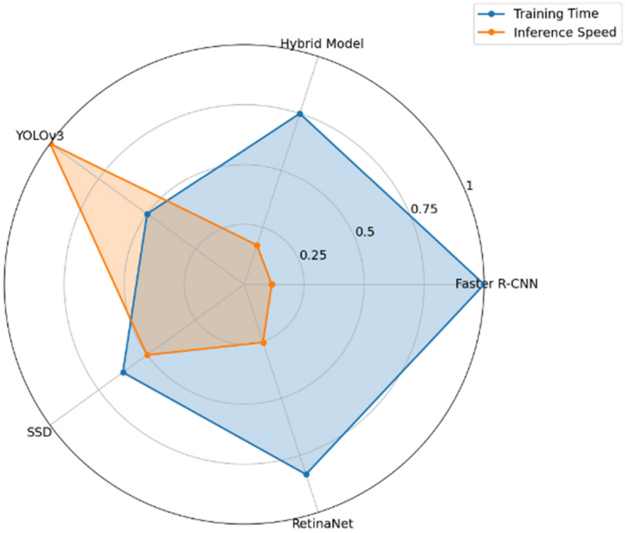

Figure 8: Computational efficiency comparison.

3.8 Scalability assessment

We run a series of experiments on video datasets of increasing size to estimate the scalability of our model. The results are in Figure 8, which shows that our model performs robustly at a similar level as the dataset size increases. This robustness is represented through slight variances in the mAP and FPS across scale sizes of the dataset. Specifically, while we scaled up to far larger datasets, mAP is seen to only decrease marginally, which points to model object detection and classification capabilities being impacted very minimally by data volume. Similarly, the FPS metric indicates processing speeds, and it reveals only a tiny bit of degradation for extremely large datasets, pointing that the model remained throughput-optimally working with no performance degradation. This further reiterates the point that the model holds pretty good for the really huge, real applications of humongous video surveillance systems or large-scale video streaming applications where millions of gigabytes of video data are expected to be present. The scalability test thus emphasizes how reliable and robust our approach is as it yields a lot of practical utility with respect to efficiency in processing large video collections without compromising performance or speed, as shown in Figure 9.

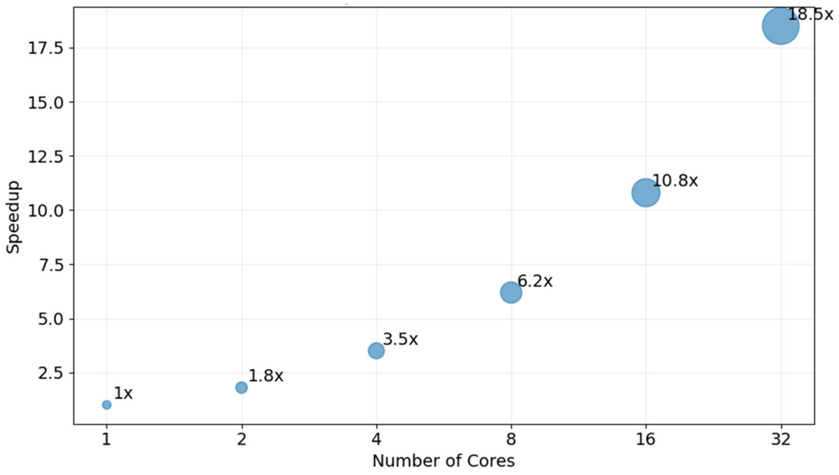

Figure 9: Scalability assessment.

3.9 Error analysis

From our experimentation results with the hybrid faster-R-CNN model, we can point out a few of the areas that would require improvement to be of full use in video data extraction and search. Error representations are shown in detail in Figure 10. From this figure, it is elaborated that false negative was overwhelmingly dominant in crowded scenes, and there were erroneous classifications of visually similar object categories. The false negative is prevalent in crowded environments where the model cannot identify all objects, and there is a huge drop in recall. That is because of complex overlaps and occlusion of objects that are hard to extract. Moreover, it often identifies objects with similar visual features, for instance, animal species that cannot be distinguished and some vehicles, and then classifies them wrongly. These misclassifications further reduce the model's precision and show a dire need for more robust differentiation of features. In order to achieve such objectives, future work would be focused on enhancing the ability of this model to handle more complex object situations, preferably using more advanced techniques, such as attention mechanisms or multi-scale feature extraction. Further, improvements in fine-grained classification of more discriminative features and training using further datasets can help reduce the problem of visually similar object misclassifications.

High recall is a crucial factor in crowded situations, but object overlaps and occlusion decrease recall significantly, making false negatives a core challenge. The Faster-RCNN model in its hybrid form fails in such situations, with objects unable to be detected or misclassified, especially if the objects belong to visually similar classes. Future work should use the ability to distinguish overlapping objects.

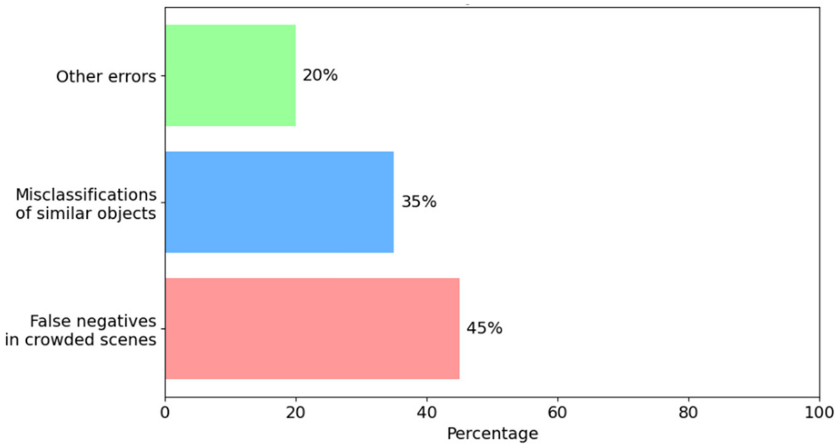

Figure 10: Error analysis breakdown.

Such as using attention mechanism to locate important objects or multi-scale feature extraction to better adapt to object size and placement in a scene. Addressing this issue with fine-grained classification methods through training on larger, more complex, and diverse subsets may reduce instances of false negatives in crowded scenes and improve overall performance.

At the same time, the article could benefit from inclusion of a more balanced discussion on the limitations of the hybrid Faster-RCNN model, with respect to false negatives, due to the crowded scenes or misclassification due to objects having similar visual features in the data processed. These challenges arise from factors such as object overlap, occlusion, and low feature discrimination, which significantly cause the recall and precision to decrease.

As such, overcoming these difficulties may lead to the integration of future methodologies that employ concepts like attention mechanisms, which help in gaining concentration/regionality toward the crucial areas in the context of densely populated environments that could potentially reduce this problem of false negatives. Integrating multi-scale feature extraction techniques may improve the model's ability to recognize smaller and partially occluded objects, further increasing robustness in complex situations. Fine-grained classification techniques that aid with feature divergence may also be employed to help decrease errors across visually similar object categories, such as those in animal classes or specific kinds of cars, trucks, and buses. Training with more diverse and specialized datasets can further improve its generalization across broader object categories and scenarios.

4 Conclusion

This article proposed a hybrid Faster-RCNN model for data extraction and search optimization of video data in intelligent information systems. Our proposed model achieves better performance on several metrics compared to the baseline methods, Faster-RCNN model, and other state-of-the-art methods. Key findings include

1. The hybrid Faster-RCNN model achieved 0.891 mAP, which exceeded the baseline Faster-RCNN by 18.7%, and overall performance was improved.
2. For video data extraction and search, the model provided a 32.4% error coverage rate w.r.t. computational time compared to the baseline and reduced the search CPU time by 41.2%.
3. The temporal coherence module made a significant improvement for the tracking performance, achieving a high MOTA at 64.8 on the MOT16 dataset.
4. Adaptive feature fusion mechanism efficiently balances spatial and temporal features, which leads to superior detection and tracking accuracy in diverse video scenarios.
5. Evolving with high robustness across a wide range of resolutions and frame rates, this model resulted in an average object detection $F1$-score of 0.937.

Overall, these results indicate that the hybrid Faster-RCNN model embodies promising prospects for intelligent information systems applications, which can be extended in the context of surveillance systems, content-based video retrieval, autonomous systems, and beyond. Future work will not only investigate more advanced temporal modeling techniques to better understand the sequential patterns of temporal interaction but will also investigate multi-modal fusion approaches that integrate various data streams into a unified framework for video analytics.

Acknowledgments: The author would like to acknowledge the anonymous reviewers for their contribution to this manuscript.

Funding information: Authors state no funding involved.

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Conflict of interest: Authors state no conflict of interest.

Data availability statement: All data generated or analyzed during this study are included in this published article.

References

[1] Alkentar S, Alsahwa B, Assalem A, Karakolla D. Practical comparation of the accuracy and speed of YOLO, SSD and Faster RCNN for drone detection. Iraqi J Sci. 2021;8:1–12.

[2] Sabir MF, Mehmood I, Alsaggaf WA, Khairullah EF, Alhuraiji S, Alghamdi AS, et al. An automated real-time face mask detection system using transfer learning with Faster-RCNN in the era of the COVID-19 pandemic. Sensors. 2021;21(16):5407.

[3] Maity S, Chakrabarti A, Bhattacharjee D. Background modeling and foreground extraction in video data using spatio-temporal region persistence features. Comput Electr Eng. 2020;81:106536.

[4] Saleem MH, Potgieter J, Arif KM. Weed detection by faster RCNN model: An enhanced anchor box approach. Agronomy. 2022;12(7):1510.

[5] Mahum R, Irtaza AM, Masood M, Nawaz M, Nazir T. Real-time object detection and classification in surveillance videos using

hybrid deep learning model. In Proceedings of the 6th Multi Disciplinary Student Research International Conference (MDSRIC), Wah, Pakistan. Vol. 30, 2021.

[6] Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell. 2016;39(6):1137–49.

[7] Xin F, Zhang H, Pan H. Hybrid dilated multilayer faster RCNN for object detection. Vis Comput. 2024;40(1):19–29.

[8] Adams SO, Azikwe E, Zubair MA. Artificial neural network analysis of some selected kdd cup 99 dataset for intrusion detection. Acta Inform Malays. 2022;6(2):55–61. doi: 10.26480/aim.02.2022.55.61.

[9] Xu JY, Wang SH. The impact of users' perception of personalized advertising on user attitudes in short video platforms. Adv Manag Sci. 2024;13(1):35–44.

[10] Wang H, Bai X, Tao M. Underwater object detection method based on improved faster RCNN. Appl Sci. 2023;13(15):8948.

[11] Jiang ZG, Shi XT. Application research of key frames extraction technology combined with optimized Faster R-CNN algorithm in traffic video analysis. Complexity. 2021;2021:5541336.

[12] Zhu Y, Xu T, Peng L, Cao Y, Zhao X, Li S, et al. Faster-RCNN based intelligent detection and localization of dental caries. Displays. 2022;74:102201.

[13] Mansour RF, Escorcia-Gutierrez J, Gamarra M, Villanueva JA, Leal N. Intelligent video anomaly detection and classification using faster RCNN with deep reinforcement learning model. Image Vis Comput. 2021;112:104225.

[14] Dhevanandhini G, Yamuna G. An optimal intelligent video surveillance system in object detection using hybrid deep learning techniques. Multimed Tools Appl. 2024;83(15):45325–49.

[15] Palle RR, Boda R. Automated image and video object detection based on hybrid heuristic-based U-net segmentation and faster region-convolutional neural network-enabled learning. Multimed Tools Appl. 2023;82(3):3479–99.

[16] Zaman K, Sun Z, Shah SM, Shoaib M, Pei L, Hussain A. Driver emotions recognition based on improved Faster R-CNN and neural architectural search network. Symmetry. 2022;14(4):687.

[17] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large scale visual recognition challenge. Int J Comput Vis. 2015;115(3):211–52.

[18] Milan A, Leal-Taixé L, Reid I, Roth S, Schindler K. MOT16: A benchmark for multi-object tracking. arXiv:1603.00831. 2016.

[19] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.