2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Muti-Scale Encoder and Temporal Queries Decoder for Video Object Detection

Hongxiao Yang, **Xi Chen***

Huazhong University of Science and Technology, Wuhan 430074, China
chenxi@hust.edu.cn

Abstract. Video Object Detection (VOD) leverages temporal information across adjacent frames in video datasets, enabling the identification and localization beyond single-frame image object detection. Transformer-based detectors have achieved remarkable performance in static image object detection. However, their application to video object detection lacks sufficient exploration, particularly in aggregating spatial features and temporal features effectively. Recent research has replaced handcrafted components in traditional optical flow models and association networks with novel designs to integrate spatial features across frames, thereby incorporating temporal information. Nevertheless, these methods often introduce significant computational overhead or complex processing pipelines. Moreover, the integration of multi-scale spatial features and temporal features into a unified framework remains challenging, making it difficult to process both small and large objects simultaneously. To address these issues and enhance detection efficiency, we propose a novel method that aggregates multi-scale spatial features and contextual temporal information. Specifically, we propose a strip attention mechanism for intra-scale feature interaction, utilize pyramid network to fuse spatial features across scales and construct temporal associations across video frames through decoder structures. Our end-to-end approach aggregates target queries progressively from coarse to fine, striking a balance between performance and efficiency. Extensive experiments on the ImageNet VID dataset demonstrate that our method significantly improves video object detection.

Keywords: video object detection, temporal queries, muti-scale features, transformer.

1 Introduction

Video Object Detection (VOD) extends static image object detection to video scenarios, identifying every object in a video sequence. Traditional static image object detection tasks take an image as input and output the objects present in the image along with their categories and bounding boxes. While it is possible to treat video streams as a series of independent frames and process them with single-frame object detectors, this approach overlooks the unique temporal information inherent in video stream. Moreover, video object detection often involves challenging scenarios, such as object

appearance changes, motion blur, occlusion, defocus, and rare poses, where single-frame object detection methods struggle to perform effectively.

To tackle these challenges, existing research explores three primary strategies to incorporate temporal information in video detection. The first involves post-processing methods [1-5], where a single-frame object detector generates detection results, and the outputs of the current frame and its neighboring frames are associated during post-processing. However, the decoupled nature of the detector and the post-processing pipeline prevents joint optimization, often resulting in suboptimal performance.

The second strategy employs temporal feature aggregation models [6-14], which introduce feature similarity metrics[13,15,17,18] or relation networks[13,19,20] to weight and fuse features between the current and reference frames. While these approaches effectively address motion blur and appearance changes, they often rely on two-stage detectors such as Faster-RCNN[21] or R-FCN[22], which involve complex designs, handcrafted components, and post-processing steps, complicating the pipeline.

With the introduction of Transformers [23], DETR [24] showcased exceptional performance in object detection, inspiring numerous improvements [25,26]. Building on DETR, some studies [27-30] address video object detection by simultaneously processing spatial and temporal information. These methods use attention mechanisms or learnable weights to establish associations on objects in different frames and aggregate image features or object queries. While effective, these approaches often overlook the importance of image-level features in guiding query initialization and learning, leading to slower convergence and miss performance gains. Furthermore, their heavy emphasis on associations in different frames associations often neglects backbone-extracted multi-scale features, resulting in subpar detection of small objects.

In this work, we reexamine the connection between the backbone-extracted spatial features and object queries, highlighting the critical role of raw spatial features in guiding the initialization of object queries. Randomly initializing queries ignores the quality of the queries, which inspired us to optimize and enhance the Backbone-Transformer pipeline in DETR. To this end, we propose a novel framework, featuring a Temporal Query Aggregation Decoder (TQAD) module designed to generate and aggregate queries across frames. Unlike existing methods that rely on weighted averages to aggregate queries, our approach aggregates temporal queries from coarse to fine through two parallel pipelines. Additionally, to solve the common blurring and deformation problems in video, we introduce a multi-scale feature fusion encoder (MSFE) based on strip attention, which fuses three scales of feature maps extracted by the backbone network, avoiding the common reliance solely on deep-layer features.

Our contributions are summarized as follows:

- We propose a Transformer-based end-to-end video object detector that balances efficiency and performance, addressing key challenges in VOD tasks.
- We design a TQAD module that leverages raw image features to guide the generation of high-quality queries, enhancing detection performance. The module is decoupled from the core model and can be easily integrated into other Transformer-based video object detectors.

- We develop a lightweight MSFE, which significantly improves deformed and small object detection performance without increasing considerable computational overhead.

2 Related Works

2.1 Video Object Detection

Compared to static images, video streams present unique challenges such as object appearance deformation, motion blur, occlusion, defocus, and background clutter. To address these issues, earlier studies primarily relied on optical flow models combined with various post-processing or feature aggregation strategies to capture temporal information. For instance, TCNN[3] segments video sequences into multiple frames and processes each frame independently by using static image object detectors while leveraging optical flow to associate detection results across frames. Similarly, methods like FGFA[17], THP[31], and MANet[15] aggregate spatial features from adjacent frames based on optical flow. However, training optical flow models requires extensive data and significant computational costs.

Subsequent studies incorporated attention mechanisms and deformable convolutions to model temporal context. For example, SELSA[16] and TROI[32] aggregate features across frames based on spatial semantic similarities. STSN[33] employs deformable convolutions to sample spatial features from reference frames, enhancing feature representation for the current frame. Advanced approaches like MEGA[13], OGEMN[34], and LWDN[35] leverage memory mechanisms to integrate local and global features, propagating and updating them across frames. Building on MEGA, TF-Blender[36] introduces learnable fusion networks to capture temporal context.

The above approaches are generally based on two-stage object detection frameworks such as Faster-RCNN[21] and R-FCN[22], which require numerous handcrafted components and post-processing steps, making the VOD pipelines overly complex and difficult to train. Recently, with the emergence of Transformer-based architectures, there has been a shift toward building end-to-end VOD frameworks. TransVOD[27,28] introduces modules such as TDTE, TQE, and TDTD to aggregate video frame features and object queries. To address the spatial information limitations of TransVOD, PTSE-Former[29], proposes the STAM module to fuse spatial features across frames. However, most of these methods aggregate queries after the decoder stage, which neglect the importance of query initialization. FQA[30] addresses this by dynamically generating and aggregating object queries using a learnable network. Our approach focuses on multi-scale feature maps extracted from backbone networks and emphasizes that they guide the generation of object queries.

2.2 Vision Transformer in Object Detection

Currently, the main research is divided into two directions. One is to directly employ the Transformer throughout the entire pipeline for feature extraction and downstream

tasks, replacing the CNN backbone network, like ViT. The other is to retain CNN as the feature extraction backbone network and utilize Transformer for instance understanding through image features and object queries, like DETR. Regardless of which direction, we can establish end-to-end vision models. In the field of object detection, works based on DETR often follow the second approach, and our work is no exception.

DETR[24] leverages the Hungarian algorithm for bipartite matching, establishing one-to-one label assignments between ground truth boxes and learnable object queries, eliminating the need for handcrafted components(e.g. anchor generation) and post-processing steps (e.g. NMS). However, the flattened feature sequences in DETR require extensive self-attention operations, leading to high computational complexity. Deformable DETR[25] addresses this by introducing deformable attention, which focuses on a few learnable reference points around each pixel, significantly reducing computational costs. To tackle DETR's slow convergence issue, DN-DETR[37] introduces denoising training to mitigate the instability of bipartite matching. DAB-DETR[38] represents each positional query explicitly as an anchor box, providing location priors to accelerate convergence. DINO[39] enhances the denoising approach by focusing on negative samples and refining query initialization. RT-DETR[40] studies the efficiency of attention across multi-scale feature maps and proposes IoU-aware query generation to improve query quality. These advancements inspires our model design, ensuring that high-quality detections are achieved for individual frames before incorporating temporal information.

3 Method

3.1 Overview

Similar to recent video object detection pipelines, our method is an end-to-end video object detector based on a Backbone-Transformer architecture. The overall structure is illustrated in **Figure 1**. Given a current frame $I_t$and its reference frames$[I_{t+i}]_{i=-L:L}$, we first extract multi-scale feature maps F_t and $[F_{t+i}]_{i=-L:L}$by using a backbone network, such as ResNet, and feed them into the Multi-Scale Feature Fusion Encoder (MSFE). The encoder produces memory outputs M_t and $[M_{t+i}]_{i=-L:L}$, which are subsequently used to initialize object queries. In MSFE, multi-scale features are fused, as detailed in Section B. Based on our experiments and insights from TransVOD, performing spatio-temporal attention operations across frame memories using a Temporal Transformer Encoder only marginally improves temporal learning while significantly increasing computational costs. Therefore, we propose TQAD to construct temporal associations among frame memories through two parallel pipelines. Finally, TQAD outputs the aggregate query $\mathcal{Q}_{fus}$ of the current frame, and the query is sent to the detection head to output the classification confidence and bounding box coordinates through the feedforward neural network.

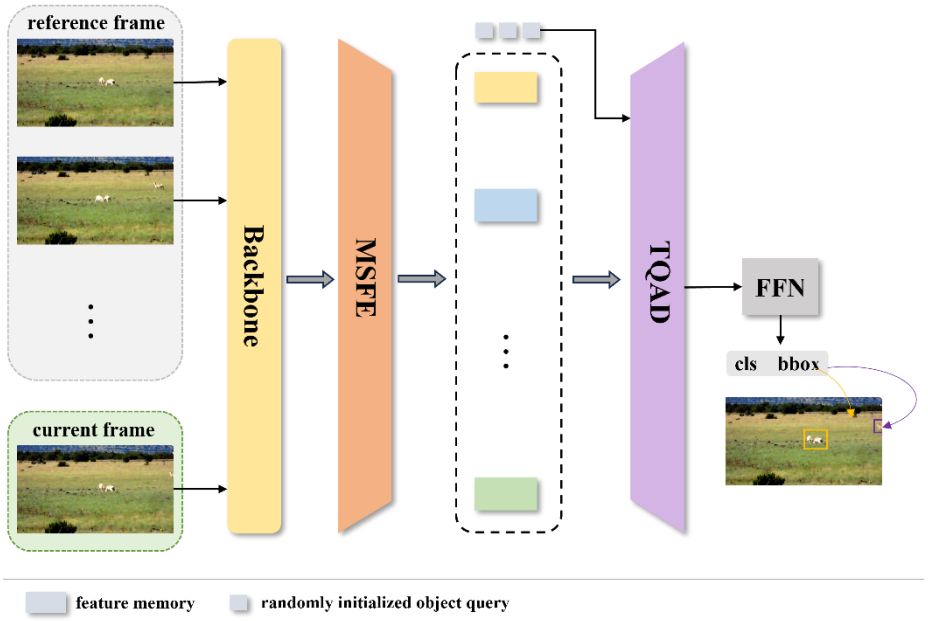

Figure 1. Framework of the proposed method.

3.2 Strip Attention

Video shot in dynamic scenes often has blurred frames, and blur and deformation may have different directions and sizes. To model this directivity, we propose strip attention, which extracts blured and deformable features by combining intra-strip and inter-strip attention. The tokens in the strip sequence have the blured feature of local single direction pixels, and the tokens that constitute the attention between the strip sequences have the fuzzy feature of the global region.

The intra-strip sequence attention module consists of two branches: horizontal in-strip sequence attention and vertical in-strip sequence attention. Let the input feature map of the attention block in the strip sequence $X \in \mathbb{R}^{H\times W\times C}$, where H, W, C represent the height, width and number of channels of the input feature map, respectively. First enter X into a LayerNorm layer for normalization, then process it with a 1×1 convolution layer, and then divide it into two along the channel dimension:

$$(X^h, X^v) = \mathrm{Chunk}\left(\mathrm{Conv}\left(\mathrm{Norm}(X)\right)\right) \tag{1}$$

where X^h and $X^v \in \mathbb{R}^{H\times W\times D}$ represent the input features of the horizontal and vertical branches of the attention block in the strip sequence, $D = C/2$.

For attention in horizontal strip sequence, the input feature X^h is divided into H non-overlapping horizontal strip sequences$X_i^h \in \mathbb{R}^{W\times D}$ along the height direction of the feature map. Each horizontal strip sequence has a length of W, which means W tokens of dimension D. X_i^h is mapped to three vector Spaces by linear projection layer, and then

the result is split to generate its query, key and value vector, and output the result through multi-head attention mechanism. The process can be described as:

$$\left(Q_i^h, K_i^h, V_i^h\right) = \mathrm{Split}(X_i^h P) \tag{2}$$

$$M_{ij}^h = \mathrm{Softmax}\left(\frac{Q_{ij}^h \left(K_{ij}^h\right)^T}{\sqrt{D/m}}\right) V_{ij}^h \tag{3}$$

$$M^h = \mathrm{Stack}(M_{ij}^h) \tag{4}$$

where m represent the number of the head in multi-head attention, i and j index the height of the feature map and the head of multi-head attention, respectively. The output of the attention branches in the vertical strip sequence M^v can be obtained by the same method. By means of residual connection, M^h and M^v are connected and sent into the 1×1 convolution layer for fusion, and the output $M'_{intra} \in \mathbb{R}^{H\times W\times C}$ of the attention layer in the strip sequence is obtained:

$$M'_{intra} = \mathrm{Conv}\left(\mathrm{Concat}(M^h, M^v)\right) + X \tag{5}$$

M'_{intra} is sent to the feedforward neural network to obtain the output memory feature $M_{intra} \in \mathbb{R}^{H\times W\times C}$ of the intra-strip attention:

$$M_{intra} = \mathrm{FFN}\left(\mathrm{Norm}(M'_{intra})\right) + M'_{intra} \tag{6}$$

Inter-strip attention is the interaction of attention based on the sequence dimension, treating each strip sequence as a token, and the module also includes two branches of horizontal and vertical. The calculation method of inter-strip attention is the same as Intra-strip attention.

3.3 Multi-Scale Feature Fusion Encoder

Deformable attention has been introduced to reduce the computational complexity of attention mechanisms while incorporating multi-scale features. However, this also increases sequence lengths, leading to negligible reductions in overall computational costs. Instead of considering cross-scale feature interactions, TransVOD replaced spatial scales with temporal sequences, but found the TDTE module's effectiveness to be limited. Some researchers[40] confirm that cross-scale feature interaction is as inefficient as cross-temporal sequences interaction. Therefore, we only use strip attention method on the deepest features.

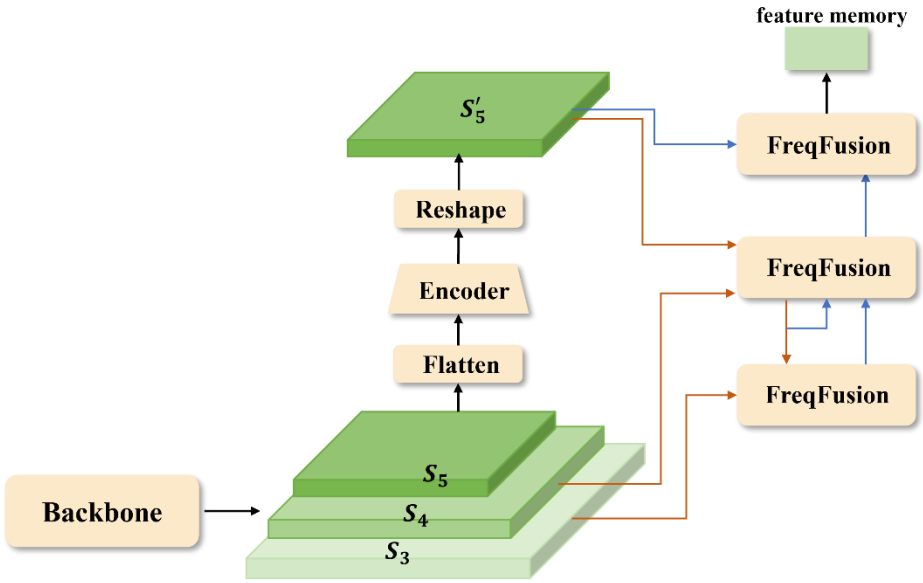

Figure 2. Details of MSFEncoder.

Based on the above analysis, we propose the Muti-Scale Feature Fusion Encoder (MSFE), illustrated in **figure 2**. It consists of an intra-scale feature interaction encoder based on strip attention and a cross-scale feature pyramid based on convolutional neural network.. Specifically, three maps which are extracted by backbone are fed into MSFE. The deepest feature map S_5, with the smallest resolution, is fed into the intra-scale feature interaction encoder for self-attention computation and reshaped to S_5' which is the same shape as S_5:

$$S_5' = \text{Rashape}(\mathcal{N}_{enc}(S_5)) \tag{7}$$

Then we send S_3, S_4 and S_5' into the multi-scale feature fusion module. Inspired by PANet, the multi-scale feature fusion module fuses features through both top-down and bottom-up pathways, with intermediate layers facilitating cross-scale interactions. The fusion module is based on FreqFusion to achieve superior fusion results. The process can be simplified as:

$$M = \text{MSFM}(S_3, S_4, S_5') \tag{8}$$

where M is called feature memory which is the final output of MSFE.

3.4 Temporal Query Aggregation Decoder

With Transformer being used for object detection, it replaces traditional handcrafted components like anchor boxes with object queries, enabling end-to-end object detection. In DETR, these object queries are randomly initialized and learn image features through attention mechanisms, allowing for object classification and localization.

However, this initialization limits queries to current-frame information, ignoring temporal context. Prior methods associate object queries across different frames to capture temporal information but often overlook the quality of these randomly initialized queries. Given the similarity and continuity between frames, object queries across frames are inherently correlated. Randomly initialized queries reduce the utility of shared temporal information, while a single shared query set fails to capture frame-specific differences. To address these issues, we propose a simple yet effective query generation and aggregation strategy for inter-frame interaction.

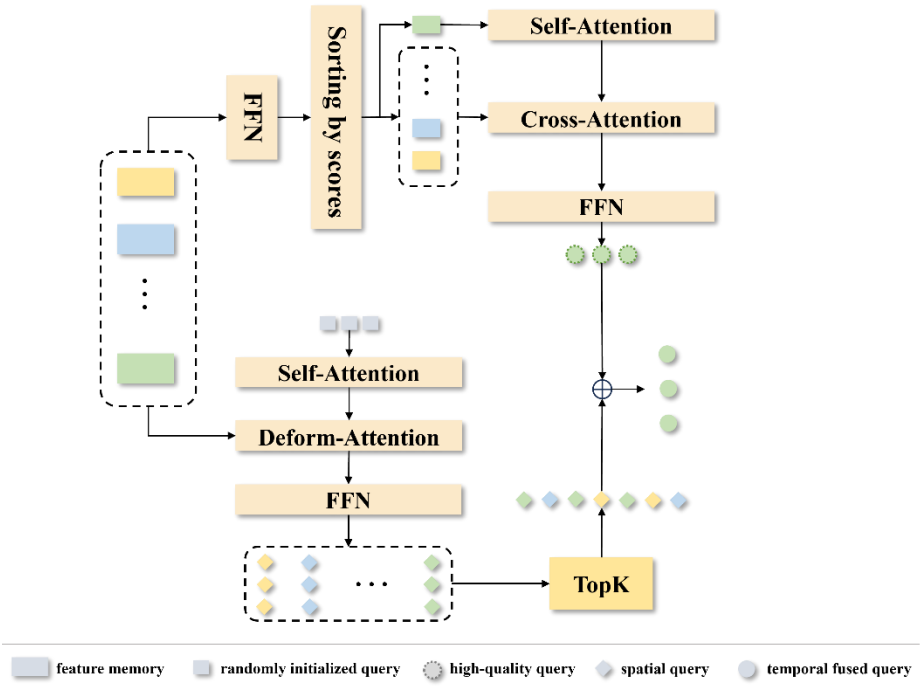

Figure 3. Details of TQADecoder.

The structure of the Temporal Query Aggregation Decoder is illustrated in **Figure 3**. In this setup, one pipeline guides randomly initialized queries using the raw features of video frames to learn inter-frame differences, while aggregating the queries most likely to map to targets. This approach ensures that the queries accurately capture information from each frame while also grasping the differences between frames, which can reflect the blurring characteristics of the images. Another pipeline performs a linear projection on the encoder's feature memory to learn the essential information from the raw features of the video frames. It filters and retains the more critical information, associating the important details between the current frame and reference frames through cross-attention to learn inter-frame differences. Finally, the results of the two pipelines are summed, fusing the similarities and differences between video frames. This enables the model to select important features from reference frames to guide the detection of the current frame.

Our main idea is to make full use of encoder memory features to guide query generation. Traditional detection pipeline is to feed the encoder memory and randomly initialized object queries into the decoder, and then output the category confidence and bounding box coordinates through FFN. This means that the object query learns the encoder's memory features from the decoder. After understanding that the original memory contain rich semantic information, we consider directly using the memory to generate target queries.

The first difference-pipeline ensures that the spatial information of each frame is thoroughly learned by the network by providing a set of randomly initialized queries $Q_{init} \in \mathbb{R}^{l\times n\times d}$ for each frame, where l, n, and d represent the number of video frames, the number of queries, and the depth of the queries, respectively. Q_{init} and all frame memories M are fed into the decoder. The output of the final decoder layer passes through a feedforward network(FFN) and the sigmoid activation function to obtain class confidence scores. These scores are then ranked, retaining only the top k highest values, as higher confidence indicates a greater likelihood of being a target. Based on this, the temporal blur queries $Q_1 \in \mathbb{R}^{k\times d}$ from the first difference-pipeline are obtained.

To ensure that the queries focus on the feature memory, the second similarity-pipeline generates and initializes queries via a fully connected (FC) layer, leveraging cross-attention to aggregate temporal information. Specifically, the feature memory M_t of the current frame and the feature memories $\{M_{t+i}\}_{i=-L:L}$ of the reference frames from the encoder output are first fed into the FC layer for linear projection. A sigmoid activation function is then applied to obtain class confidence scores s:

$$s = \text{sigmoid}[\text{FFN}(M)] \tag{9}$$

The confidence scores s are sorted, and the top k values are retained for each frame. Unlike the Top-k processing in the first difference-pipeline, which globally ranks the combined s values across all frames, the second pipeline independently sorts the results for each frame and retains the top k values per frame. Here, k denotes the number of queries per frame.After Top-k selection, high-quality queries Q_t (for the current frame) and $Q_t^r = \{Q_{t+i}\}_{i=-L:L}$ (for the reference frames) are obtained, which encapsulate rich semantic information from each frame. To establish temporal dependencies among the queries, self-attention is applied to the current frame's queries Q_t, enhancing focus on the current frame:

$$Q, K, V = Q_t \tag{10}$$

$$\text{self\_attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{11}$$

Cross-attention is employed between the current frame's queries Q_t and the reference frame queries Q_t^r to interactively fuse their features:

$$Q = Q_t \tag{12}$$

$$K = V = Concat(Q_t^r) \tag{13}$$

$$\text{cross\_attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right) V \tag{14}$$

In this way, the second pipeline thus generates, initializes, and aggregates high-quality query Q_2 across frames, embedding temporal context into the object queries for further processing.

Finally, the output of the two pipes is aggregated as Q_{dec} through element-wise addition:

$$Q_{dec} = Q_1 \oplus Q_2 \tag{15}$$

4 Experiment

4.1 Experimental Setup

Dataset. In order to make a fair comparison with other state-of-the-arts methods, we conduct experiments on ImageNet-VID[41] and evaluate our approach. ImageNet-VID is a large-scale public dataset for video object detection, which contains 30 object categories with 3862 training videos and 555 validation videos. Following the existing common training protocol for transformer-based video object detectors, we mix ImageNet-VID and ImageNet-DET datasets to train and validate our model. Mean Average Precision (mAP) is used as the evaluation metric.

Training details. In this work, we use a ResNet-50[42] pretrained on ImageNet as the backbone of the network. Our model is trained with the AdamW optimizer on four RTX3090 GPUs, each of which trains a batch containing one current frame and its corresponding reference frame. As in the previous methods, we perform bidirectional uniform sampling for the reference frames, where reference frames are evenly sampled from both sides of the current frame.

4.2 Comparison with State-of-the-art Methods

We use ResNet-50 as the backbone to compare our method with recent state-of-the-art video object detectors. The performance of these models on the ImageNet VID validation set is presented in Table 1. Since such studies rarely focus on small object detection alone and there is no publicly available performance metric APS, we reimplement and train the methods which lack publicly available pre-trained weights to get their mAP and APS.

It is not difficult to see our method achieves an mAP of 84.6%, surpassing existing Transformer-based methods by 1.1% to 5.7%. Notably, our method demonstrates significant improvements in small-object detection, with APS. reaching 16.5%, an increase of 1.5% to 5.2% over competing methods.

Table 1. Performance comparison with the recent state-of-the-art video object detection approaches on ImageNet VID validation set with ResNet50 backbone.

Method	Base Detector	mAP_S	mAP_{50}
DFF [18]	R-FCN	5.0	70.4
FGFA [17]	R-FCN	5.8	74.0
SELSA [16]	Faster-RCNN	7.6	75.8
TROI [32]	Faster-RCNN	7.1	76.5
MEGA [13]	Faster-RCNN	8.2	77.3
TransVOD [27]	Deformable DETR	11.3	79.9
TransVOD++ [28]	Deformable DETR	13.4	80.5
FAQ [30]	Deformable DETR	13.2	81.7
Ours	Deformable DETR	16.5	85.6

4.3 Ablation Studies and Analysis

MSFE. To evaluate the efficiency of the Multi-Scale Feature Fusion Encoder (MSFE), we conduct ablation studies by replacing the MSFE with the original encoder in Deformable DETR. As shown in Table 2, incorporating the MSFE improves overall performance, increasing mAP from 76.0% to 79.8%. The improvement is particularly pronounced for small objects, with AP_S increasing from 10.6% to 13.7%. These results highlight the effectiveness of MSFE, especially in addressing the challenges of small-object detection in the ImageNet VID dataset. Additionally, the lightweight design of the MSFE ensures that inference speed is only reduced by approximately 10%.

Table 2 Ablation studies of MSFE.

Method	mAP_S	mAP_{50}	FPS
Vanilla Encoder	10.6	76.0	25.2
MSFE	13.7	79.8	21.4

Table 3 Ablation studies of pipeline design.

Baseline	TransVOD	FAQ	TQAD	mAP_{50}
✓				76.0
✓	✓			79.9
✓		✓		81.7
✓			✓	82.5

TQAD. Our method TQAD aggregates temporal information in a coarse-to-fine manner through the design of two pipelines. To validate the effectiveness of this design, we conducted comparisons with modules that aggregate temporal information in SOTA. As shown in Table 3, Ablation studies demonstrate the effectiveness of TQAD. The baseline model is Deformable DETR which achieves 76.0% mAP. The three modules of TransVOD have increased mAP to 79.9% and the dynamic query strategy proposed by the FAQ has improved to 81.7%. The performance of our method is further improved to 82.5%, which is 0.8% better than the current best method.

Number of encoder layers in MSFE. To analyze the impact of encoder depth on model performance, we vary the number of encoder layers in MSFE while keeping other parameters constant. As shown in Table 4(a), the result indicate that performance peaks when using 2 encoder layers. Increasing the number of layers beyond this does not yield further benefits and may even degrade performance for small-object detection.

Number of decoder layers in TQAD. We investigate the effect of varying the number of decoder layers in TQAD, fixing the number of encoder layers in MSFE to one. The results shown in Table 4(b) indicate that performance stabilizes after three decoder layers, so we adopt three layers in the final configuration.

Number of top k queries of each frame in TQAD. To assess the impact of query quality on model performance, we vary the number of retained queries *k* after FFN and sigmoid activation. As shown in Table 4(c), increasing *k* improves performance up to a threshold of 300 queries, beyond which performance plateaus or even declines due to the inclusion of low-quality queries. These results validate the importance of prioritizing high-quality queries.

Table 4 Ablation studies on hyperparameters of each component.

(a) Number of encoder layers N_{MSFE} in MSFE					
N_{MSFE}	0	1	2	3	4
mAP_{50}	77.3	79.3	79.8	79.9	79.5
mAP_S	11.2	12.9	13.7	13.6	13.5
(b) Number of decoder layers N_{TQA} in TQAD					
N_{TQAD}	1	2	3	4	5
mAP_{50}	81.7	82.4	82.5	82.5	82.6
(c) top k queries of each frame in TQAD					
k	50	100	200	300	500
mAP_{50}	81.2	81.5	82.1	82.5	82.5

4.4 Visualization

In addition to extensive experimental data results, we also aim to demonstrate the performance of our method through visualizations. Figure 4 illustrates the detection results comparison between our model and the pre-trained Deformable DETR model across two video sequences.

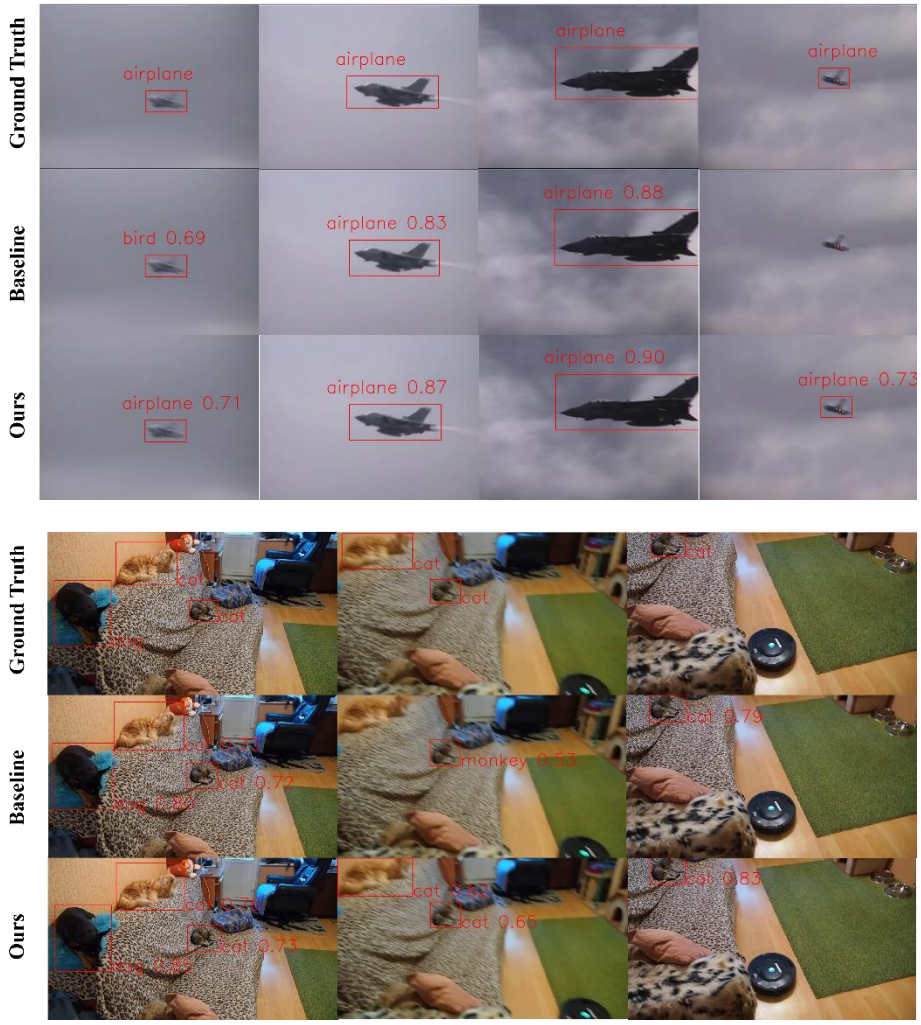

Figure 4. Visualization of the detection results. Baseline indicates pre-trained Deformable DETR model on ImageNet VID.

As shown in the figure, our method plays a decisive role in practical detection. In the first video sequence, the baseline model suffers from false positives and missed detections. Specifically, it misidentifies an airplane as a bird in the first image and fails to detect the target in the last frame. In the second scenario, similar issues arise: the baseline model misses the orange cat in the upper-left corner of the second frame and incorrectly labels the central cat as a monkey. This can be attributed to the inherent challenges of single-frame detectors, which lack temporal context to handle quality degradation caused by motion blur or occlusions. In contrast, our approach successfully detects all targets across both sequences and achieves significant improvements in

detection precision. These results clearly demonstrate the superiority of our method in leveraging temporal dependencies to enhance robustness under complex video conditions.

5 Conclusion

In this work, we summarize the limitations of existing video object detectors and analyzed the advantages of transformer-based methods. Current state-of-the-art methods often overlook the importance of query quality and struggle with small-object detection. From a novel perspective, we propose a coarse-to-fine pipeline approach to simultaneously capture spatial and temporal information in video sequences. Additionally, we designe a lightweight Multi-Scale Feature Fusion Encoder (MSFE) to enhance the detection performance for blurry, deformed and small objects, and a Temporal Query Aggregation Decoder (TQAD) to guide the generation and aggregation of high-quality queries. Extensive experiments on the ImageNet VID dataset validate the effectiveness of our approach, demonstrating a well-balanced trade-off between performance and efficiency.

References

[1] Wei Han, Pooya Khorrami, Tom Le Paine, Prajit Ramachandran, Mohammad Babaeizadeh, Honghui Shi, Jianan Li, Shuicheng Yan, and Thomas S Huang. Seq-nms for video object detection. arXiv preprint arXiv:1602.08465, 2016..

[2] Hatem Belhassen, Heng Zhang, Virginie Fresse, and ElBay Bourennane. Improving video object detection by seqbbox matching, in VISIGRAPP, 2019, 5: 226-233..

[3] Kai Kang, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Object detection from video tubelets with convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 817-825.

[4] Kai Kang, Hongsheng Li, Junjie Yan, Xingyu Zeng, Bin Yang, Tong Xiao, Cong Zhang, Zhe Wang, Ruohui Wang, Xiaogang Wang, et al. T-cnn: Tubelets with convolutional neural networks for object detection from videos. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 28(10): 2896-2907.

[5] Alberto Sabater, Luis Montesano, and Ana C Murillo. Robust and efficient post-processing for video object detection. In 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2020: 10536-10542.

[6] C.-H. Yao, C. Fang, X. Shen, Y. Wan, and M.-H. Yang, Video object detection via object-level temporal aggregation. In European Conference on Computer Vision, 2020: 160-177.

[7] Z. Jiang, Y. Liu, C. Yang, J. Liu, P. Gao, Q. Zhang, S. Xiang, and C. Pan. Learning where to focus for efficient video object detection. In European Conference on Computer Vision, 2020: 18-34.

[8] M. Han, Y. Wang, X. Chang, and Y. Qiao. Mining inter-video proposal relations for video object detection. In European Conference on Computer Vision, 2020: 431-446.

[9] L. Han, P. Wang, Z. Yin, F. Wang, and H. Li. Exploiting better feature aggregation for video object detection. In: Proceedings of the 28th ACM International Conference on Multimedia, Seattle, USA, Oct. 12-16, 2020: 1469-1477.

[10] L. Lin, H. Chen, H. Zhang, J. Liang, Y. Li, Y. Shan, and H. Wang. Dual semantic fusion network for video object detection, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020: 1855-1863.
[11] F. He, N. Gao, Q. Li, S. Du, X. Zhao, and K. Huang. Temporal context enhanced feature aggregation for video object detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020: 34(07): 10941-10948.
[12] K. Chen, J. Wang, S. Yang, X. Zhang, Y. Xiong, C. C. Loy, and D. Lin. Optimizing video object detection via a scale-time lattice, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7814-7823.
[13] Y. Chen, Y. Cao, H. Hu, and L. Wang. Memory enhanced globallocal aggregation for video object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 10337-10346.
[14] G. Sun, Y. Hua, G. Hu, and N. Robertson. Mamba: Multi-level aggregation via memory bank for video object detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(03): 2620-2627.
[15] Shiyao Wang, Yucong Zhou, Junjie Yan, and Zhidong Deng. Fully motion-aware network for video object detection, in: Proceedings of the European Conference on Computer Vision, 2018: 542-557.
[16] Haiping Wu, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Sequence level semantics aggregation for video object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 9217-9225.
[17] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-guided feature aggregation for video object detection, in: Proceedings of the European Conference on Computer Vision, Honolulu, 2017: 408-417.
[18] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei. Deep feature flow for video recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017: 2349-2358.
[19] J. Deng, Y. Pan, T. Yao, W. Zhou, H. Li, and T. Mei. Relation distillation networks for video object detection in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 7023-7032.
[20] M. Shvets, W. Liu, and A. C. Berg. Leveraging long-range temporal relationships between proposals for video object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 9756-9764.
[21] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, 2015, 28: 91-99.
[22] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. in: Advances in Neural Information Processing Systems, 2016: 379-387.
[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need, in: Advances in Neural Information Processing Systems, 2017: 5998-6008.
[24] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, in: Proceedings of the European Conference on Computer Vision, 2020: 213-229.
[25] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, Deformable detr: Deformable transformers for end-to-end object detection, in: International Conference on Learning Representations, 2020.
[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly et al. An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representation , 2021.

[27] Lu He, Qianyu Zhou, Xiangtai Li, Li Niu, Guangliang Cheng, Xiao Li, Wenxuan Liu, Yunhai Tong, Lizhuang Ma, and Liqing Zhang. End-to-end video object detection with spatial-temporal transformers, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021: 1507-1516.
[28] Qianyu Zhou, Xiangtai Li, Lu He, Yibo Yang, Guangliang Cheng, Yunhai Tong, Lizhuang Ma, and Dacheng Tao. TransVOD: end-to-end video object detection with spatial-temporal transformers. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(6): 7853-7869.
[29] Han Wang, Jun Tang, Xiaodong Liu, Shanyan Guan, Rong Xie, and Li Song. Ptseformer: Progressive temporal-spatial enhanced transformer towards video object detection. In European Conference on Computer Vision, 2022: 732-747.
[30] Yiming Cui. Feature aggregated queries for transformer-based video object detectors, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 6365-6376.
[31] X. Zhu, J. Dai, L. Yuan, and Y. Wei. Towards high performance video object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7210-7218.
[32] T. Gong, K. Chen, X. Wang, Q. Chu, F. Zhu, D. Lin, N. Yu, and H. Feng. Temporal roi align for video object recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(02): 1442-1450.
[33] G. Bertasius, L. Torresani, and J. Shi. Object detection in video with spatiotemporal sampling networks, in: Proceedings of the European Conference on Computer Vision, 2018: 331-346.
[34] H. Deng, Y. Hua, T. Song, Z. Zhang, Z. Xue, R. Ma, N. Robertson, and H. Guan. Object guided external memory network for video object detection in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 6678-6687.
[35] Z. Jiang, P. Gao, C. Guo, Q. Zhang, S. Xiang, and C. Pan. Video object detection with locally-weighted deformable neighbors, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(01): 8529-8536.
[36] Y. Cui, L. Yan, Z. Cao, and D. Liu. Tf-blender: Temporal feature blender for video object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 8138-8147.
[37] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 13619-13627.
[38] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr, in International Conference on Learning Representations, 2021.
[39] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection, in: International Conference on Learning Representations, 2022.
[40] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, Jie Chen. Detrs beat yolos on real-time object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 16965-16974.
[41] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein et al. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 2015, 115(03): 211-252.
[42] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016: 770-778.