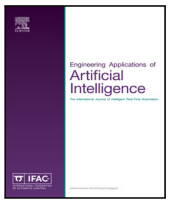
Engineering Applications of Artificial Intelligence 139 (2025) 109606

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Hybrid multi-attention transformer for robust video object detection

Sathishkumar Moorthy [a,b], Sachin Sakthi K.S. [a], Sathiyamoorthi Arthanari [a], Jae Hoon Jeong [a], Young Hoon Joo [a,*]

[a] School of IT Information and Control Engineering, Kunsan National University, 558 Daehak-ro, Gunsan-si, Jeonbuk 54150, Republic of Korea
[b] Smart Vision Tech Inc, Daeryung Techno Town 6, 648, Seobosaet-gil, Geumcheon-gu, Seoul 08504, Republic of Korea

ARTICLE INFO

Keywords:
Video object detection
Vision transformers
Attention mechanism
Target-background embeddings

ABSTRACT

Video object detection (VOD) is the task of detecting objects in videos, a challenge due to the changing appearance of objects over time, leading to potential detection errors. Recent research has addressed this by aggregating features from neighboring frames and incorporating information from distant frames to mitigate appearance deterioration. However, relying solely on object candidate regions in distant frames, independent of object position, has limitations, as it depends heavily on the performance of these regions and struggles with deteriorated appearances. To overcome these challenges, we propose a novel Hybrid Multi-Attention Transformer (HyMAT) module as our main contribution. HyMAT enhances relevant correlations while suppressing flawed information by searching for an agreement between whole correlation vectors. This module is designed for flexibility and can be integrated into both self- and cross-attention blocks to significantly improve detection accuracy. Additionally, we introduce a simplified Transformer-based object detection framework, named Hybrid Multi-Attention Object Detection (HyMATOD), which leverages competent feature reprocessing and target-background embeddings to more effectively utilize temporal references. Our approach demonstrates state-of-the-art performance, as evaluated on the ImageNet video object detection benchmark (ImageNet VID) and the University at Albany DEtection and TRACking (UA-DETRAC) benchmarks. Specifically, our HyMATOD model achieves an impressive 86.7% mean Average Precision (mAP) on the ImageNet VID dataset, establishing its superiority and practicality for video object detection tasks. These results underscore the significance of our contributions to advancing the field of VOD.

1. Introduction

Computer vision is a branch of AI that aims to give computers the ability to see and understand the world like humans do. Within the domain of computer vision, various tasks are undertaken. These tasks encompass classification (Gu et al., 2024), which entails categorizing an image into one or more predefined classes; segmentation (Zhang et al., 2024a), which involves the extraction of regions of interest from images; and tracking, where the goal is to monitor objects of interest across a continuous video stream (Pan et al., 2023). Among these tasks, object detection holds a pivotal position due to its foundational role in understanding and interacting with visual data (Li et al., 2023a; Qi et al., 2024; Chen et al., 2024). Object detection is centered on the precise localization and recognition of objects within an image, typically involving the prediction of object coordinates and subsequent determination of their respective class labels. This field of study is vital in the development of computer vision systems capable of comprehending and interpreting visual data, with applications spanning autonomous vehicles, medical imaging, surveillance, road event detection, robotics, and more. Accurate object detection is essential for enabling higher-level vision tasks and ensuring robust performance in real-world scenarios.

VOD aims to detect the objects consistently across a sequence of consecutive video frames. Advancements in storage and communication technologies have ushered in a new era where video has emerged as a prominent medium for conveying a wealth of information. In contemporary times, video-based analysis has become ubiquitous, with applications extending to areas such as action recognition, autonomous driving, robotic navigation, surveillance systems, agriculture, and healthcare. However, challenges such as occlusion, motion blur, out-of-focus cameras, and uncommon object poses in recorded videos can significantly degrade the quality of detections. Using image-based object detectors individually for each frame within a video frequently leads to suboptimal outcomes. Fortunately, videos inherently contain valuable temporal information, including rich cues about the continuous movement of objects. Leveraging this temporal context through

* Corresponding author.
E-mail address: yhjoo@kunsan.ac.kr (Y.H. Joo).

https://doi.org/10.1016/j.engappai.2024.109606
Received 10 March 2024; Received in revised form 30 September 2024; Accepted 1 November 2024
Available online 15 November 2024

Table 1
Major contributions on papers related to CNN-based models and Transformer.

Category	Method	Contributions and innovations
One-stage	• RetinaNet (Lin et al., 2017) • YOLOv3 (Redmon and Farhadi, 2018) • CenterNet (Zhou et al., 2019a) • ExtremeNet (Zhou et al., 2019b) • YOLOv8 (Jocher and Chaurasia, 2023)	• RetinaNet incorporates a focal loss that adapts the weighting scheme dynamically, improving the model's performance on hard-to-detect objects. • YOLOv3 uses anchor boxes to predict object locations and sizes, achieving superior results on small-scale objects. • CenterNet utilizes key-point and center-point detection paradigms, simplifying objects into single points for more efficient detection. • ExtremeNet leverages extreme points and center points to directly identify key points and bounding boxes, improving detection accuracy. • YOLOv8 introduces an anchor-free detection approach, streamlining the training process and enhancing model flexibility.
Two-stage	• Faster R-CNN (Ren et al., 2015) • Mask R-CNN (He et al., 2017) • Cascade R-CNN (Cai and Vasconcelos, 2018) • Grid R-CNN (Lu et al., 2019) • AugFPN (Guo et al., 2020)	• Faster R-CNN introduces a region proposal network (RPN) to generate proposals at multiple scales and ratios, enhancing detection accuracy. • Mask R-CNN extends Faster R-CNN by incorporating ROIAlign for better alignment and integrating an FPN network to handle multi-scale features. • Cascade R-CNN uses a cascading training process to address performance degradation with increased IoU thresholds. • Grid R-CNN divides images into grid cells to adjust grid sizes based on object scales, improving the model's accuracy. • AugFPN applies consistent supervision to narrow the semantic gap between different scales before fusion, enhancing detection performance across scales.
DETR	• Deformable DETR (Zhu et al., 2020) • Efficient DETR (Yao et al., 2021) • Sparse DETR (Roh et al., 2021) • Anchor DETR (Wang et al., 2022b) • Focus-DETR (Zheng et al., 2023)	• Deformable DETR integrates deformable attention modules to concentrate on small key points around sampled features, enhancing the model's ability to capture fine-grained details. • Efficient DETR uses both dense and sparse modules within a single detection head to improve detection speed while maintaining accuracy. • This method introduces a sparse sampling strategy that focuses on the most relevant predictions, reducing computational complexity. • Anchor DETR employs anchor points as object queries, enabling the precise prediction of multiple objects in a given region. • Focus-DETR leverages dual attention modules to capture detailed foreground information, enhancing the model's ability to identify fine-grained features.

effective integration of temporal data can significantly enhance the accuracy and reliability of object detection in videos.

In the current landscape, object detectors could be divided into several categories including two-stage models (Ren et al., 2015), one-stage models (Liu et al., 2016; Redmon et al., 2016; Redmon and Farhadi, 2018), and query-based models (Carion et al., 2020; Zhu et al., 2020). Table 1 offers a comprehensive comparison of these methodologies, highlighting their contributions and innovations. These categories offer distinct approaches, each with its unique strengths and limitations. Two-stage models, as exemplified by the Region-based Convolutional Neural Network (R-CNN) family (Ren et al., 2015; Dai et al., 2016), operate in a two-step manner. Initially, they generate a set of proposals, followed by refining the prediction results. However, a drawback of these two-stage detectors is their relatively slower inference speed, which can hinder real-time applications. On the other hand, one-stage object detectors aim to strike a balance between efficiency and performance by directly predicting object locations and categories from the input image feature maps. Notable examples include the You only look once (YOLO) series (Redmon et al., 2016; Redmon and Farhadi, 2018) and a fully convolutional one-stage object detector (FCOS) (Tian et al., 2019). More recently, the spotlight has turned to query-based object detectors. These models generate predictions using a series of input queries and eliminate the need for intricate post-processing pipelines such as Non-Maximum Suppression (NMS). Prominent models in this category encompass the DEtection TRansformer (DETR) series (Carion et al., 2020; Zhu et al., 2020).

The aforementioned VOD approaches have primarily harnessed temporal information in two distinct ways. The first method involves post-processing (Han et al., 2016) which uses the temporal information to enhance the coherence and stability of detection results. Usually, these methods use a static-image detector to acquire detection results and subsequently strive to establish temporal associations with these outcomes. However, the absence of integrated optimization between image object detectors and post-processing pipelines often results in subpar performance, especially in situations involving substantial object motion or intricate interactions. Conversely, another set of approaches (Chen et al., 2020; He et al., 2020), focuses on aggregating temporal information's features. Specifically, they enrich the features of the current frame by meticulously integrating information from neighboring frames or even entire video clips, utilizing a unique array of expertly designed operators. This approach effectively addresses challenges including video defocus, pose divergence, motion blur, target occlusion, and rapid appearance changes. Significantly, a number of these approaches (Chen et al., 2020) employ two-stage detectors such as Faster-RCNN (Ren et al., 2015) or Region-based Fully Convolutional Networks (R-FCN) (Dai et al., 2016) as their foundational static-image detectors, leveraging their high detection accuracy while compensating for their slower inference speeds with advanced temporal aggregation techniques.

Transformers have exhibited substantial performance on a broad range of vision tasks including object detection, segmentation, etc. (Li et al., 2023b). Technically, DETR (Carion et al., 2020) presents an alternative approach to solving object detection challenges. DETR conceptualizes object detection as a matching problem based on sets. Leveraging Transformers (Vaswani et al., 2017), initially developed for natural language processing (NLP) functions, the DETR can model the relationships among objects and their broader context within an image using a group of learned object queries. A bipartite match ensures unique predictions from query objects enabled by global optimization techniques. In contrast to conventional object detection techniques, this method does not rely on handcrafted features such as NMS or anchor generation. Despite this, DETR continues to face a variety of challenges that have prevented its widespread acceptance in the research community. An input resolution constraint is imposed by the native Transformer architecture used as the feature encoder in the first case. This limitation arises from the quadratic growth in complexity associated with the self-attention module when dealing with higher input resolutions. As a result, DETR is not entirely compatible with the common feature pyramid approach widely employed in modern object detectors, and it exhibits relatively lower efficiency in detecting

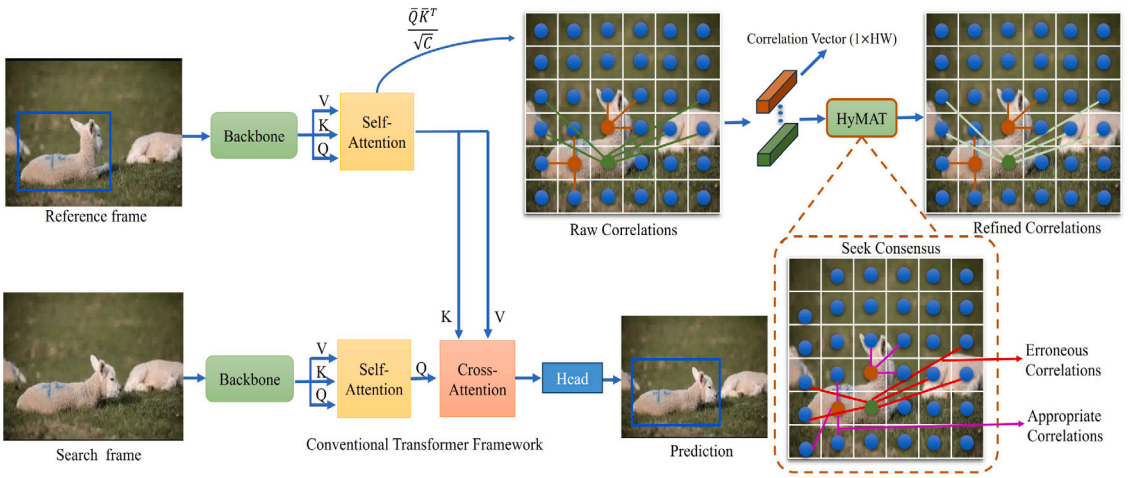


Fig. 1. A description of the method and its motivation. On the left, we see an example of a Transformer Detection framework. On the right, a feature map shows nodes representing features positioned at different points. A self-attention block is made up of these nodes that serve as queries and keys. Correlations between these two are represented by the links between nodes in the attention mechanism.

small objects. Additionally, DETR demands significantly longer training epochs to achieve convergence compared to existing object detectors. Consequently, there is a pressing need for an effective solution to enhance the performance of DETR. Recent efforts, such as Deformable DETR (Zhu et al., 2020), combine sparse spatial sampling through deformable convolution with the relational modeling capabilities of Transformers. This groundbreaking approach is designed to address challenges in DETR, such as sluggish convergence and elevated computational complexity, with the goal of enhancing overall performance. Notably, Deformable DETR has achieved noteworthy improvements in both performance and training efficiency. Thus, the exploration of potential avenues to further enhance the efficiency and performance of DETR remains an intriguing prospect.

In a typical framework for object detection based on Transformers (Dai et al., 2021), the key components revolve around the utilization of attention blocks. In order to predict, among other things, targets within a search frame, a self-attention block can be used to raise the representation of features within the reference frame and the search frame, as shown in Fig. 1. By establishing correlations between them, cross-attention blocks facilitate the establishment of cross-frame correlations. By combining queries with key–value pairs as inputs, the Transformer attention mechanism can produce linear combinations of values based on queries and key–value pairs. It is determined by the correlation between the queries and keys, which determines each combination's weight based on the correlation between them. By scaling the dots between the queries and key values, the correlation map is generated. Although query-key pairs' correlations are computed individually, the inter-dependencies between query-key pairs are ignored. In the absence of contextual awareness, correlations can be inaccurate for a variety of reasons, most importantly when dealing with sub-optimal feature representations or when distracting patches of imagery are present in a cluttered background scene. Consequently, this may result in the generation of attention weights that are noisy and ambiguous.

To tackle the previously mentioned issue, we present a new component known as the hybrid multi-attention (HyMAT) module. This module enhances the traditional attention mechanism first introduced by Vaswani et al. in 2017 (Vaswani et al., 2017). Within HyMAT, the internal attention module aims to achieve consensus across all correlation vectors, enabling a comprehensive refinement of correlations. Fig. 1 illustrates the rationale behind the introduction of the HyMAT module. Whenever one key exhibits a high correlation with a query, the likelihood that adjacent keys will exhibit a high correlation is fairly high. Conversely, weak correlations can be regarded as undesirable noise. Taking advantage of this insight, we integrate the internal attention module to harness these valuable cues. By using the original correlations considering queries, keys, and values, the internal attention module optimizes query-key correlations by amplifying relevant correlations while suppressing erroneous correlations between irrelevant query-key pairs. Additionally, our HyMAT module seamlessly integrates into self-attention blocks along with cross-attention blocks which significantly improve the performance. The significance of these enhancements becomes evident in a Transformer-based video object detection framework, resulting in a substantial improvement in the overall detection performance.

Main contributions of our work:

1. A new hybrid multi-attention (HyMAT) module has been developed to decrease noise and ambiguity in the conventional attention mechanism (Vaswani et al., 2017), resulting in notable enhancements in object detection performance.
2. An elegant framework for video object detection is presented, utilizing the Transformer architecture. In this proposed approach, encoded features are employed more efficiently, and embeddings of target backgrounds are introduced. This enhancement enables a more effective utilization of temporal references in our approach.
3. The short-term branch incorporates a novel Intersection over Union (IoU) prediction head to leverage information from frames in closer proximity to the current frame. By utilizing both long-term and short-term references, there is a substantial improvement in capturing variations in object appearance, leading to enhanced performance.
4. Through rigorous experimentation and in-depth analysis, we provide substantial evidence to confirm the efficacy of the proposed HYMATOD method. Specifically, the developed HYMATOD mechanism consistently attains the best efficiency over two extensively recognized benchmark datasets.

The rest of the paper is organized as follows: Section 2 covers related work, including image and video object detection and Vision Transformers. Section 3 revisits DETR, while Section 4 presents our proposed Hybrid Multi-Attention method and framework. Sections 5 to 7 detail experimental evaluations on the ImageNet VID, UA-DETRAC datasets, and custom videos. Section 8 discusses limitations and future work, and Section 9 concludes the paper.

Abbreviations	
VOD	Video object detection
HyMAT	Hybrid multi-attention
HyMATOD	Hybrid multi-attention object detection
mAP	Mean average precision
R-CNN	Region-based convolutional neural network
YOLO	You only look once
FCOS	Fully convolutional one-stage object detector
NMS	Non-maximum suppression
DETR	Detection transformer
R-FCNN	Region-based fully convolutional networks
NLP	Natural language processing
IoU	Intersection over union
CNN	Convolutional neural network
NMS	Non-maximum suppression
DFF	Deep feature flow
FGPA	Flow-guided feature aggregation
THP	Towards high performance
MaNET	Motion-aware network
SELSA	Sequence level semantics aggregation
TCENet	Temporal context enhanced network
RDN	Relation distillation network
OGEMN	Object guided external memory network
MEGA	Memory enhanced global-local aggregation
VOT	Visual object tracking
HiFT	Hierarchical feature transformer
TQE	Temporal query encoder
TDTD	Temporal deformable transformer decoder
PTSEFormer	Progressive temporal-spatial enhanced transformer
FFN	Feed forward network
GIoU	Generalized IoU
RPN	Region proposal network
RoI	Region of interest

2. Related work

VOD has been a major area of research in computer vision for decades. VOD models have received significant attention in recent years from both researchers and academics, and there has been an extensive amount of research on this topic in the literature, with references such as Jin et al. (2024) and Han et al. (2022). Throughout this part, we will discuss the relevant object detection methods.

2.1. Image object detection

Over the years, object detection methods in the image domain have made significant strides in their development (Ren et al., 2015; Liu et al., 2016; Dai et al., 2016; Qiao et al., 2024; Yang et al., 2023). Image object detection necessitates precise prediction of object locations and categories within the input image. Two-stage object detectors, pioneered by the RCNN family (Dai et al., 2016; Ren et al., 2015), have laid the groundwork for this field. They involve initially predicting rough object proposals, followed by refinement for improved accuracy. Among these, Faster RCNN (Ren et al., 2015) stands out as a widely adopted approach. In Faster RCNN, a convolutional neural network (CNN) extracts image features, which are then used in a region-proposal phase to generate numerous region proposals. Subsequently, a detection stage classifies and refines these proposals. Both stages require non-maximum suppression to eliminate redundancy. Approaches like Deformable Convolution (Dai et al., 2017) and Relation Networks (Hu et al., 2018) have been pivotal in enhancing object detection performance. Deformable convolution, for instance, samples features from dynamic locations to improve receptive field alignment, while Relation Networks employ self-attention (Vaswani et al., 2017) among region proposals to enrich their features with contextual information. However, all these models typically rely on preprocessing steps like anchor design (Liu et al., 2016; Redmon et al., 2016) or post-processing techniques like NMS (Liu et al., 2019). Different from the above, we place a primary focus on Transformer-based image object detectors and design modules to extend their capabilities to tasks within the video domain.

2.2. Video object detection

An object's location in a single frame is determined by the VOD task (Liu et al., 2021; Zhang et al., 2024b), which establishes a connection between objects within multiple frames. Leading-edge methods typically devise intricate pipelines to address this challenge. One prevalent approach (Chen et al., 2020; He et al., 2020) for mitigating this issue is feature aggregation, which enhances frame-level features by amalgamating information from nearby frames. In earlier works, this was accomplished using flow-based warping techniques. FlowNet (Dosovitskiy et al., 2015) is a pioneering approach that introduced the concept of regressing optical flow using an end-to-end deep neural network. It employs an encoder–decoder architecture enhanced with skip connections. Subsequently, deep feature flow (DFF) (Zhu et al., 2017b) integrates the principles of a segmentation network and FlowNet (Dosovitskiy et al., 2015), achieving a 74% reduction in computation time for semantic video segmentation. However, DFF relies on a fixed key frame scheduling policy, which assumes a constant update interval between consecutive key frames. This limitation reduces its flexibility and customizability. For instance, flow-guided feature aggregation (FGFA) (Zhu et al., 2017a) utilizes optical flow to warp the

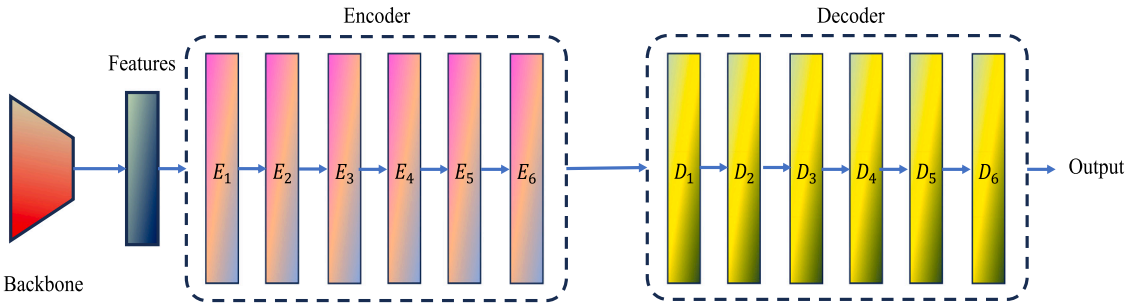


Fig. 2. The DETR architecture.

feature maps of past and future frames, subsequently aggregating these warped feature maps to detect objects in the current frame. On the other hand, towards high performance (THP) (Zhu et al., 2018) extracts optical flow to propagate keyframe features to non-keyframe features. However, the extensive use of optical flow for feature aggregation and warping, typically managed by an auxiliary model, greatly increases both the model size and computational complexity. By combining pixel-level and instance-level calibrations based on motion, motion-aware network (MANet) (Wang et al., 2018) overcomes inaccuracies in pixel-level features caused by flawed flow estimation. Despite these limitations, flow-warping methods are still challenging and expensive to obtain, since substantial flow data is required for model training. In scenarios involving multitask learning, it poses a challenge to seamlessly incorporate both a flow network and a detection network within a single model. To capture long-range dependencies for temporal context, attention-based approaches, such as self-attentions (Vaswani et al., 2017) and non-locals, use attention-based mechanisms. The sequence level semantics aggregation (SELSA) (Wu et al., 2019) advocates aggregating features at the sequence level for videos that are composed of an unordered series of frames.

The temporal context enhanced network (TCENet) (He et al., 2020) proposes using deformable convolution to aggregate temporal contexts. This approach is implemented within a complex framework featuring numerous heuristic designs. The relation distillation network (RDN) (Deng et al., 2019b) is introduced to aggregate object features across multiple video frames, enhancing object detection accuracy. Then, the RDN models object relationships through multi-stage reasoning, progressively refining proposal scores by distilling the connections between objects across different proposals. The Object Guided External Memory Network (OGEMN) (Deng et al., 2019a) introduces a top-down object-guided strategy that computes features with high confidence levels for detected objects and selectively stores the most confident ones. Memory enhanced global-local aggregation (MEGA) (Chen et al., 2020) proposes the consolidation of both global and local information extracted from the video, culminating in the formation of an extensive memory network. Even though they are highly successful, most VOD pipelines are extremely complex. Their performance depends heavily on handcrafted components, such as optic flow models, memory mechanisms, and recurrent neural networks. Seq-NMS (Han et al., 2016) require sophisticated post-processing techniques in order to achieve high performance by creating tubelets based on object continuity across video frames, aggregating classification scores within these tubelets. The proposed HyMATOD method is different from previous video object detection methods in that it uses target background embeddings to distinguish the target and background regions and provide rich contextual cues. This has not been studied in a comprehensive way in previous methods.

2.3. Vision transformers

According to Vaswani et al. (2017), the Transformer model is renowned for its ability to learn long-range dependencies and has achieved impressive results in the field of NLP. As recently as Dosovitskiy et al. (2020), researchers have extended the Transformer's application to computer vision tasks, where Transformer-based approaches have been demonstrated to perform as well as CNN-based approaches in image segmentation, object detection, and visual tracking. The introduction of the DETR series (Carion et al., 2020; Zhu et al., 2020) for image-related tasks sparked interest in harnessing Transformer-based models for object detection.

As an illustration, for the visual object tracking (VOT) tasks (Gao et al., 2022), the authors of Wang et al. (2021) adapted the transformer decoder to manage feature correlations between images, replacing traditional correlation models like depth-wise cross-correlation, which are commonly used in VOT. Additionally, hierarchical feature transformer (HiFT) (Cao et al., 2021) employed the transformer decoder when analyzing images for feature correlation, utilizing hierarchical features extracted from the images through a CNN backbone. Due to the multi-head attention mechanisms within the decoder, tasks involving feature correlation are naturally suited to those mechanisms. In the realm of VOD, the TransVOD series (Zhou et al., 2022) introduced the temporal query encoder (TQE) and the temporal deformable transformer decoder (TDTD) to enhance performance. Similarly, progressive temporal-spatial enhanced transFormer (PTSEFormer) (Wang et al., 2022a) proposed the spatial transition awareness model to fuse temporal and spatial information, resulting in improved predictions. These models mainly depend on aggregating features from adjacent frames to improve the characteristics of the current frame before generating ultimate predictions. In contrast, our paper takes a distinct approach by addressing the challenges of noise and ambiguity within conventional attention mechanisms. We solve these issues by finding agreement between correlations across the entire image, providing a new approach to tackling these problems.

3. Revisiting DETR

A multi-head attention layer is at the core of the detection transformer, which was originally designed for language processing within the Transformer architecture (Vaswani et al., 2017). Fig. 2 illustrates the DETR architecture.

The architecture encompasses a robust combination of components, including a CNN backbone, a transformer model, along with a FFN. To achieve this, the features $f \in \mathscr{R}^{H \times W \times C}$ from the input image $I \in \mathscr{R}^{H_0 \times W_0 \times 3}$ are generated using CNN backbone. These features undergo a meticulous process of channel dimension reduction from C to d before making their way to the encoder. Following this, the features are skillfully flattened into tokens $X \in \mathscr{R}^{HW \times d}$ along the spatial dimension.

The subsequent stage involves the nuanced processing of tokens X in a sequence encoder, resulting in the derivation of encoder features $f^e \in \mathscr{R}^{HW \times d}$. Concurrently, the decoder engages in a series of identical decoder layers to process tokens X and produce decoder features $f^d \in \mathscr{R}^{N \times d}$ in the direction of object queries. The object queries,

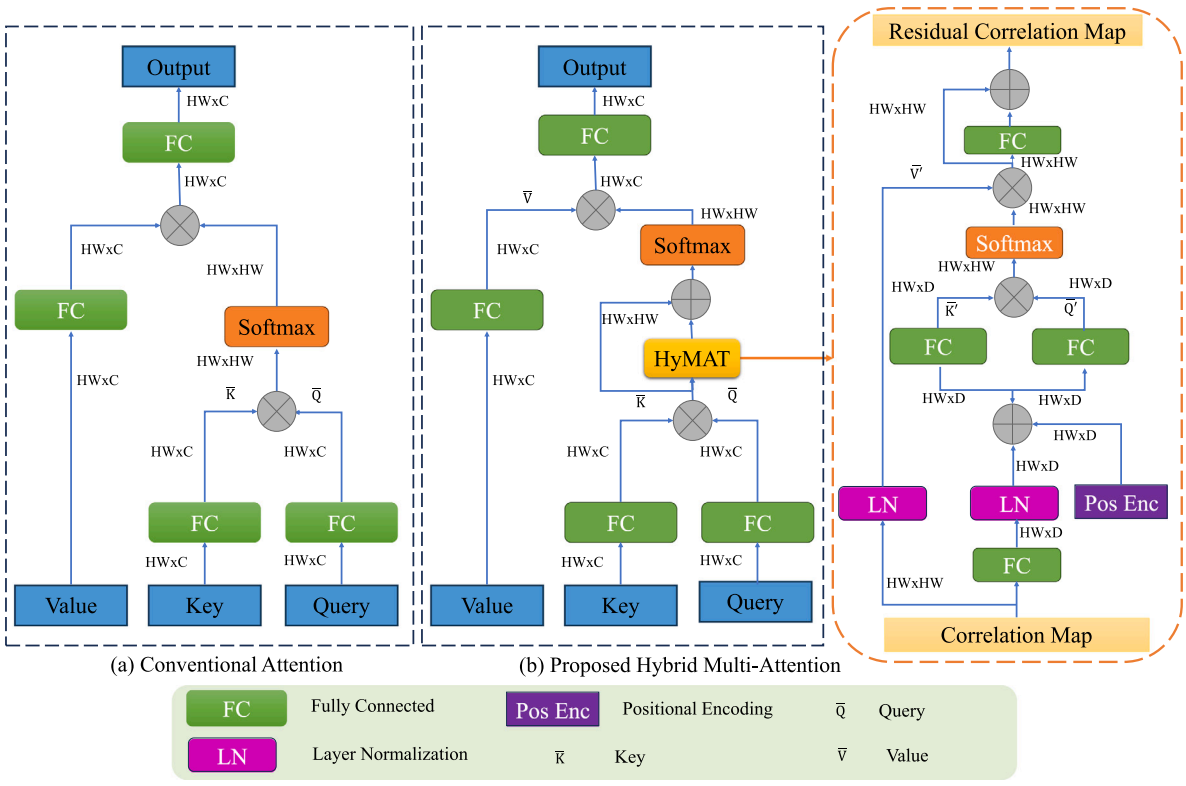


Fig. 3. The diagram compares the conventional attentional structure and the proposed hybrid multi-attentional module (HyMAT). Matrix multiplication is indicated by ⊗, and element-by-element addition by ⊕. The number beside each arrow indicates the feature dimension without considering the batch size.

embodying learnable input embeddings, are shaped by the configurable hyperparameter N in DETR.

In the final phase, the FFN showcases its predictive prowess, accurately forecasting both the box coordinates together with class labels. This intricate synergy of components ensures a sophisticated and high-performing model in the context of DETR.

Each encoder layer is intricately structured, featuring a self-attention module and a position-wise feedforward network (FFN). To elaborate on the forward propagation of these modules without sacrificing generality, consider the following expressions:

$$Z' = X + LN(MHA(Q, K, V)) \tag{1}$$

$$Z = Z' + LN(FFN(Z')) \tag{2}$$

Here, Q, K, and V denote the query, key, and value, respectively. Specifically, LN and MHA refer to layer normalization and multi-head attention. The input tokens X are divided over h clusters denoted as $X_1, \ldots, X_h$. This division occurs with the channel dimension. Subsequently, this method exploits scaled dot-product attention to each group, followed by a linear projection. This can be expressed as follows:

$$MHA(Q, K, V) = Concat(head_1, \ldots, head_h)W_O, \tag{3}$$

where $head_i = Attention(X_i W_i^Q, X_i W_i^K, X_i W_i^V)$ represents the single head output

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{4}$$

In this context, the term $softmax(\frac{QK^T}{\sqrt{d_k}})$ is commonly defined as the attention map A. We incorporate positional encoding (PE) into the input X to address the permutation invariance of the transformer architecture. This addition ensures the derivation of Q, K, and V results in:

$$\begin{bmatrix} Q \\ K \\ V \end{bmatrix} = \begin{bmatrix} (X + PE)W^Q \\ (X + PE)W^K \\ XW^V \end{bmatrix} \tag{5}$$

Here, $W^Q \in \mathcal{R}^{d_{model} \times d_q}$, $W^K \in \mathcal{R}^{d_{model} \times d_k}$, and $W^V \in \mathcal{R}^{d_{model} \times d_v}$ represent the parameters governing the scaled dot-product attention. Additionally, it is important to note that $d_{model} = \frac{d}{h}$, where d_{model} is the model dimension and h is the number of attention heads.

When adapting the Transformer architecture for object detection, specific adjustments have been introduced to accommodate spatial information. As far as the Transformer's encoder stage is concerned, it exclusively uses the self-attention mechanism, which derives keys and queries from flattened pixels in the feature maps. It is important to note that this approach has a quadratic computational complexity that increases with the spatial dimensions of the input data. A limitation like this significantly restricts the resolution and practicality of obtaining features from a pyramid representation. Transformer's decoder stages incorporate multi-head attention mechanisms in both the self-awareness and cross-awareness mechanisms. In addition to utilizing learnable object embeddings for keys and queries to streamline the self-attention part of the system, the cross-attention part of the system still relies on feature maps for keys. Consequently, this configuration presents challenges in learning to focus a query on a sparsely localized region, especially when starting with an initial uniform attention distribution across the entire feature maps.

4. The proposed method

4.1. Hybrid multi-attention

As depicted in Fig. 3, the traditional attention block takes a query and a set of key–value pairs as input and outputs a weighted sum of

the values, where the weights are determined by the softmax of the scaled dot products between the query and the corresponding keys. The traditional attention mechanism is mathematically defined as follows:

$$Attention(Q, K, V) = (softmax\left(\frac{\bar{Q}\bar{K}^T}{\sqrt{d_k}}\right)\bar{V})W_O, \tag{6}$$

where $\bar{Q} = QW_q$, $\bar{K} = KW_k$, $\bar{V} = VW_v$. Moreover, the weights for Q, K, V and O are represented as W_q, W_k, W_v, and W_o.

The correlation map (M) shows the correlation between each query-key pair, $M = \frac{\bar{Q}\bar{K}^T}{\sqrt{C}} \in \mathscr{R}^{HW\times HW}$ is calculated independently of its potential correlation with other query-key pairs. On the other hand, an isolated approach to correlation computation carries the risk of introducing inaccuracies, especially when dealing with imperfect feature representations or when distracting image elements are present in a cluttered background scene. These inaccuracies can give rise to noisy and uncertain attention patterns. Eventually, Transformer-based detectors perform less optimally as a result of these disruptions which affect feature aggregation in self-attention and information flow in cross-attention.

This paper introduces a novel HyMAT to make the correlation map more interpretable, denoted by M, addressing the challenges mentioned above. As a rule of thumb, when one key displays a strong correlation with a query, the neighboring keys are likely to display similar correlations. In some cases, however, weak correlations indicate the presence of noise. Building on this observation, we present a HyMAT module to capitalize on the informative cues found within the correlations within M. To strengthen valid correlations among relevant query-key pairs while reducing spurious correlations involving irrelevant query-key pairs, the HyMAT module upholds correlation consistency around each key.

Particularly, we incorporate an additional attention component for further developing the refinement of the correlation map, as shown in Fig. 3. As a result of the seamless integration of this attention module into the traditional attention block, it has been referred to as an "internal attention module", creating an attention-within-attention mechanism. It is a modified version of the conventional attention mechanism at its core that makes up the internal attention module. Within this context, we treat the columns in M as vectors, that the internal attention module employs as queries (Q'), keys (K'), and values (V') to generate an augmented correlation map through a residual operation.

Considering the inputs Q', K', and V', our initial step involves the creation of modified queries $\bar{Q}'$, keys $\bar{K}'$, which is shown in Fig. 3. Notably, we begin with a linear transformation aimed at reducing the dimensions of Q' and K' to enhance scalability. Subsequently, normalization techniques in Ba et al. (2016) are exploited, and then, 2-D sinusoidal encoding (Vaswani et al., 2017) is incorporated to impart positional information. Following this, two distinct linear transformations generate variables $\bar{Q}'$ and $\bar{Q}'$. Furthermore, we ensure V' undergoes normalization to produce normalized correlation vectors, denoted as $\bar{V}'$ where $\bar{V}' = LayerNorm(V')$. Finally, the residual correlation map is generated by using $\bar{Q}'$, $\bar{K}'$, and $\bar{V}'$ and the internal attention module as follows:

$$Internal\,Attention(M) = (Softmax\left(\frac{\bar{Q}'\bar{K}'^T}{\sqrt{D}}\right)\bar{V}')(1 + \bar{W}_0'). \tag{7}$$

Here, $\bar{W}_0'$ represents the weight parameters used for linear transformation, serving to adapt and combine the aggregated correlations while maintaining an identical connection.

At its core, within the HyMAT module, each correlation vector in the map M undergoes a process that produces its corresponding residual correlation vector through the aggregation of the raw correlation vectors. Accordingly, a global receptive field is effectively established, at least among the correlations. By using the residual correlation map, we can summarize what our attention block, enhanced by the HyMAT module, looks like:

$$Hybrid\,Attention(Q, K, V) = (Softmax(M + Internal\,Attention(M))\bar{V})W_O. \tag{8}$$

We use the same HyMAT module parameters for all parallel attention heads in a multi-head attention block.

4.2. Proposed framework

We developed a robust Transformer framework for detecting video objects utilizing the HyMAT module. This framework consists of a backbone, and a Transformer architecture followed by two heads which are used for the purpose of target prediction as shown in Fig. 4. Our approach takes a search frame as input and uses a group of middle frames and a persistent long-term reference (the initial frame) to extract features. The network backbone also plays an important role in determining what features to extract from the search frame itself. A Transformer encoder is used to meticulously refine and enhance these extracted features. In addition, learnable embeddings of target and background are introduced to help distinguish target regions from the background. Subsequently, Transform encodes reference features and embeddings so that they can be disseminated across search frames. As a result of the model's output, a target prediction head is meant for localization, and an IoU prediction head is used for short-term reference updates. A VOD approach based on this approach results in remarkable results.

4.2.1. Transformer architecture

We use a Transformer encoder to enhance the features extracted from the convolutional backbone. To do this, we flatten the search frame features into a sequence of feature vectors. To preserve spatial information, we add sinusoidal positional encoding, following a similar approach to the one described in Carion et al. (2020). This series of characteristic vectors is then provided to the Transformer encoder, which consists of multiple stacked layers. Transformer encoder layers are made up of two key components: MHA blocks and FFN. The SA block is essential for capturing relationships between whole feature vectors, thereby enriching the actual components. Notably, we incorporate our proposed HyMAT module to further enhance its capabilities.

On the one hand, the Transformer decoder conveys essential source details upon each long and short-term template. This allows us to remove the self-attention block from the conventional Transformer decoder (Vaswani et al., 2017). Rather, we introduce a dual-branch cross-attention mechanism to efficiently extract insights about long and short-term sources about a target and backgrounds. As a result of extracting source information from the initial frame, the long-term branch can compute reliable target annotations. On the other hand, the reference information from the long-term branch needs to be updated as the target's appearance and background change throughout the video. This has the potential to induce tracker drift in specific scenarios. With the short-term branch, we can address this issue by using information from frames closer to the current frame. Two branches of cross-attentional blocks employ similar structures, Vaswani et al. (2017), in which the search frame features are used to determine queries, and the reference frame features are used to determine keys. A target-background embedding map is used to create values by combining reference features with reference features. Notably, our HyMAT module enhances the cross-attention mechanism, thereby improving the propagation of reference information. This streamlined approach strikes a balance between efficiency and effectiveness, enhancing the robustness of video object detection.

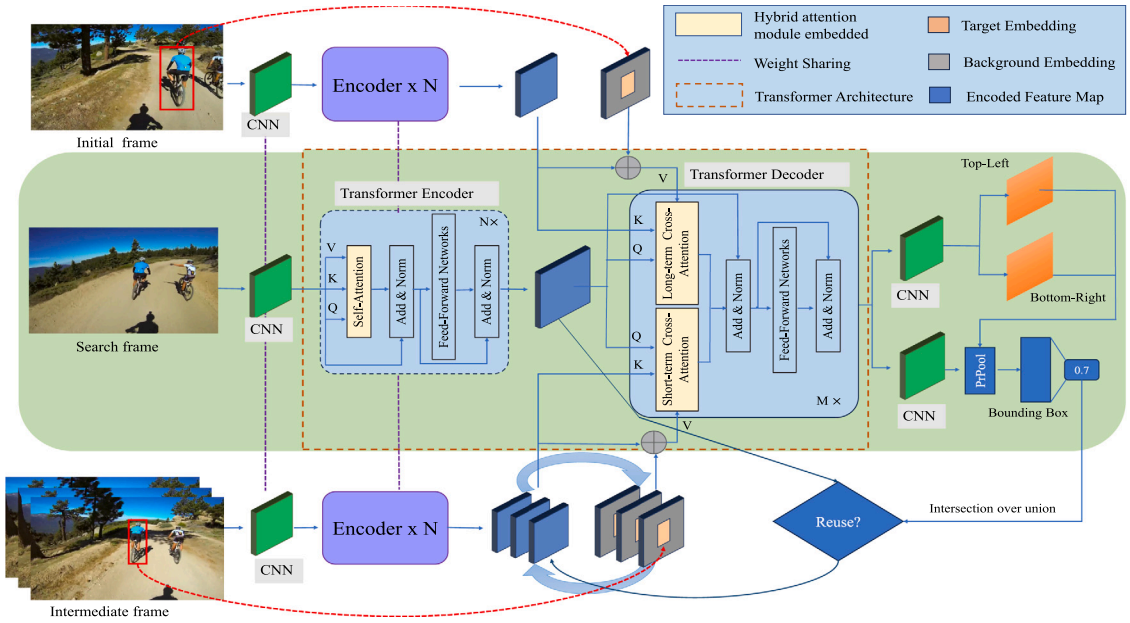


Fig. 4. This figure presents the proposed Transformer Detection framework, which integrates the HyMAT module within self and cross-attention blocks. It illustrates the data flow, including the reuse of encoded features for efficient model updates based on IoU scores. Specifically, the 'Reuse?' block determines whether to reuse previously computed features, enhancing the efficiency of the update procedure. If the estimated IoU score of the predicted bounding box exceeds a predefined threshold, the target-background embedding map for the current search frame is generated and stored in a memory cache along with its encoded features.

4.2.2. Target-background embeddings

To effectively preserve the contextual information, the proposed method utilizes two learnable embeddings: a target and background embedding denoted as $\epsilon^{tgt} \in \mathcal{R}^C$ and $\epsilon^{bg} \in \mathcal{R}^C$ respectively. These embeddings facilitate the creation of target-background embedding maps, represented as $\epsilon^{tgt} \in \mathcal{R}^{HW \times C}$. For a specific position p, the assignment of embeddings is defined as follows:

$$\epsilon(p) = \begin{cases} \epsilon^{tgt}, & \text{if p in the target region,} \\ \epsilon^{bg}, & \text{otherwise.} \end{cases} \tag{9}$$

Subsequently, the reference features exploit the target background embedding maps. By incorporating embedded target-background maps, valuable contextual information is provided in order to enhance the reutilized appearance features.

4.2.3. Prediction heads

In our detection system, we have developed two prediction heads. The first of these heads is based on Yan et al. (2021). The decoded features are then fed into an FCN with two branches, which predicts the target bounding box. With the result of these probability distributions, the box coordinates are calculated. To compensate for changes in the target's appearance during detection, it must maintain short-term references that contain the target at all times. Furthermore, when analyzing the embedding assignment mechanism Eq. (9), the bounding box accuracy of the particular frame is of utmost importance.

Taking inspiration from IoU-Net (Jiang et al., 2018), we introduce an IoU prediction head for each predicted bounding box. With the help of features positioned within the predicted bounding box, this head estimates the IoU through the ground truth. An IoU prediction is generated by analyzing these features in a Processed RoI Pooling layer, followed by a fully connected network in which these features are processed. To decide whether to add a search frame, we predict the IoU score between the bounding box in the search frame and the target bounding box in the current frame. We train two prediction heads for the bounding box and the IoU score. The target prediction loss is a combination of the Generalized IoU (GIoU) loss and the L1 loss between the predicted bounding box and the ground truth bounding box. The GIoU loss (Rezatofighi et al., 2019) is a loss function that is designed to improve the localization accuracy of object detectors by penalizing bounding boxes that are too large or too small. The L1 loss is a loss function that is designed to minimize the absolute difference between two values. To train the IoU prediction head, we sample bounding boxes around the ground truth bounding boxes and define the loss using the mean squared error.

4.2.4. Target prediction

In order to streamline the detection process seamlessly, eliminating the need for cumbersome post-processing steps, the developed algorithm incorporates the anchor-free prediction approach introduced in Yan et al. (2021). This approach yields probability maps $P_{tl}(x, y)$ and $P_{br}(x, y)$ for the top-left and bottom-right corners of the object bounding box. Furthermore, the predicted box coordinates $\hat{x}_{tl}$, $\hat{y}_{tl}$, $\hat{x}_{br}$, $\hat{y}_{br}$ are subsequently derived by

$$\hat{x}_{tl} = \sum_{y=0}^{H} \sum_{x=0}^{W} x \cdot P_{tl}(x, y), \hat{y}_{tl} = \sum_{y=0}^{H} \sum_{x=0}^{W} y \cdot P_{tl}(x, y) \tag{10}$$

$$\hat{x}_{br} = \sum_{y=0}^{H} \sum_{x=0}^{W} x \cdot P_{br}(x, y), \hat{y}_{br} = \sum_{y=0}^{H} \sum_{x=0}^{W} y \cdot P_{br}(x, y) \tag{11}$$

4.2.5. Training and inference

In line with Carion et al. (2020), we employ the Hungarian algorithm (Kuhn, 1955) to associate predictions with ground truth, maintaining consistency with the training methodology of HyMAT Transformer, which remains identical to the original DETR. The loss function is defined as the matching cost, following the conventions outlined in Carion et al. (2020). The loss function is as follows:

$$\mathcal{L} = \lambda_{cls} \cdot \mathcal{L}_{cls} + \lambda_{L1} \cdot \mathcal{L}_{L1} + \lambda_{giou} \cdot \mathcal{L}_{giou} \tag{12}$$

where $\mathcal{L}_{cls}$ represents the focal loss. $\mathcal{L}_{L1}$ and $\mathcal{L}_{giou}$ represent $L1$ loss and generalized IoU loss. λ_{cls}, λ_{L1} and $\lambda giou$ are coefficients.

4.2.6. VOD with HyMAT method

Commencing with the initial frame containing ground truth annotation, we initialize the detector by extracting both long and short-term

Table 2
Experimental environment configuration.

Parameter	Configuration
CPU	Intel Core i9 13th gen, CPU 3.00 GHz
GPU	NVIDIA GeForce RTX 4090 GPU
Accelerated environment	CUDA 10.1, cuDNN8.0.5
Visual studio system	2019, Pytorch 1.10.1 or below
Operating system	Ubuntu 18.040

references. These references are accompanied by precomputed features and the generation of target-background embedding maps. As we progress to subsequent frames, we employ the target prediction head to determine the IoU score for the predicted bounding box, a critical step in model updating. Importantly, this update procedure offers significant efficiency gains compared to the previous method (Wang et al., 2021), thanks to the direct reuse of the encoded features. For each incoming frame, we maintain the practice of uniformly selecting several short-term reference frames from the memory cache concatenating their features, and embedding maps. This process refreshes and updates the short-term reference ensemble effectively. It is worth noting that we consistently include the most recent reference frame from the memory cache to ensure its relevance to the current search frame.

5. Experiments on the ImageNet VID dataset

5.1. Datasets

We assess our HyMATOD on the ImageNet VID dataset (Russakovsky et al., 2015), which is a widely used benchmark for the task of detecting objects in videos. This dataset encompasses 3862 training videos and 555 validation videos, each annotated with bounding boxes for 30 distinct classes. Given that the official testing set's ground truth is not publicly accessible, we adopted established VOD protocols (Zhu et al., 2017a; Wang et al., 2018; Deng et al., 2019b; Wu et al., 2019). Our model training incorporated a fusion of the ImageNet VID and DET datasets (Russakovsky et al., 2015), and we evaluated performance on the validation set using the mean average precision (mAP@IoU = 0.5) metric. In our experiments, we utilized an experimental environment configuration detailed in Table 2 and the specific parameters employed are outlined in Table 3 to configure the HyMATOD method.

5.2. Network architectures

The transformer architecture is similar to that in DETR (Carion et al., 2020) with 6 encoder layers and 6 decoder layers. In alignment with established research, our detector undergoes pre-training using the COCO dataset (Lin et al., 2014). Conforming to the well-established implementation practices in prior research (Zhu et al., 2017a; Wang et al., 2018; Deng et al., 2019b; Wu et al., 2019), the proposed method exploits ResNet-101 (He et al., 2016) as network backbones. All these backbone architectures are pre-trained using the ImageNet (Deng et al., 2009) dataset. Our Transformer encoder comprises three layers stacked together, while the Transformer decoder is composed of a single layer. Within our approach, the multi-head attention blocks are equipped with four heads, each with a channel width of 256. Moreover, the internal module of HyMAT is responsible for reducing channel dimension to 64. Our feed-forward Neural Network (FFN) blocks boast 1024 hidden units. Moving on to the target prediction head, each branch consists of five Conv-BN-ReLU layers. In contrast, the IoU prediction head is constructed using three Conv-BN-ReLU layers in the PrPool layer.

Table 3
Parameters of the proposed HyMATOD method.

Parameter	Value	Parameter	Value
Learning rate	0.0001	Weight decay	0.0001
Epoch	500	LR drop epoch	400
Batch size	16	IoU weight	2.0
GIoU weight	2.0	L1 weight	5.0
Optimizer	ADAMW	Epoch interval	20
Backbone multiplier	0.1	Scheduler decay rate	0.1
Scheduler type	Step		

Table 4
Component analysis.

Method	mAP_{50} (%) overall
Base (a)	74.6
(a) + Target Background embeddings (b)	76.4
(a) + (b) + Long-short term branch (c)	79.8
(a) + (b) + (c) + HyMAT in Self-attention (d)	82.1
(a) + (b) + (c) + (d) + HyMAT in Cross-attention (e)	84.7
(a) + (b) + (c) + (d) + (e) + Positional Encoding (f)	86.2
Overall (g)	86.7

5.3. Detection network

We utilize Faster R-CNN (Ren et al., 2015) as our detection network and apply a Region Proposal Network (RPN) to the output from the fourth convolutional layer (conv4). Our anchor design includes three aspect ratios $(1:2, 1:1, 2:1)$ and four scales $(64^2, 128^2, 256^2, 512^2)$, resulting in 12 anchors per spatial location. To reduce redundancy during both the training and inference stages, we employ NMS with an IoU threshold of 0.7, yielding 300 candidate boxes per frame. Subsequently, we use RoIAlign on the output from the fifth convolutional layer (conv5), followed by a fully connected layer to extract the Region of Interest (RoI) features for each box.

5.4. Effectiveness of each component

In order to evaluate the effectiveness of the essential components in our HyMATOD system, we conducted experiments to analyze the influence of each element on the overall performance. The summarized results are presented in Table 4. Method (a) represents the baseline model, Faster R-CNN (Ren et al., 2015), a well-established object detection framework combining a Region Proposal Network (RPN) with a Fast R-CNN detector. The RPN generates region proposals, which are then refined and classified by the Fast R-CNN detector. Using a ResNet-101 backbone without any feature aggregation, this method achieves an overall mAP of 74.6%. Method (b) enhances the baseline method for object detection by introducing target-background embeddings. Then, the Method (c) incorporates both long-term and short-term references. Technically, the performance of the method (c) is significantly increased compared to the previous variant.

Following that, method (d) integrates the developed HyMAT model with the self-attention blocks of the encoder. This significantly improves performance on the ImageNet dataset, demonstrating the HyMAT module's versatility in self-attention blocks. Method (e) incorporates the HyMAT model with the cross-attention blocks of the decoder, further enhancing performance. This demonstrates the HyMAT module's effectiveness in both SA and CA blocks. Method (f) applies the HyMAT module to both SA and CA blocks, resulting in a noteworthy 11.6% improvement across all metrics compared to the baseline framework. In summary, the HyMAT module is a versatile and effective feature aggregation method that can be used to improve the performance of Transformer-based models in a variety of tasks. Additionally, we evaluate the performance of HyMATOD with a baseline method across different epochs, illustrated in Fig. 5. At the initial

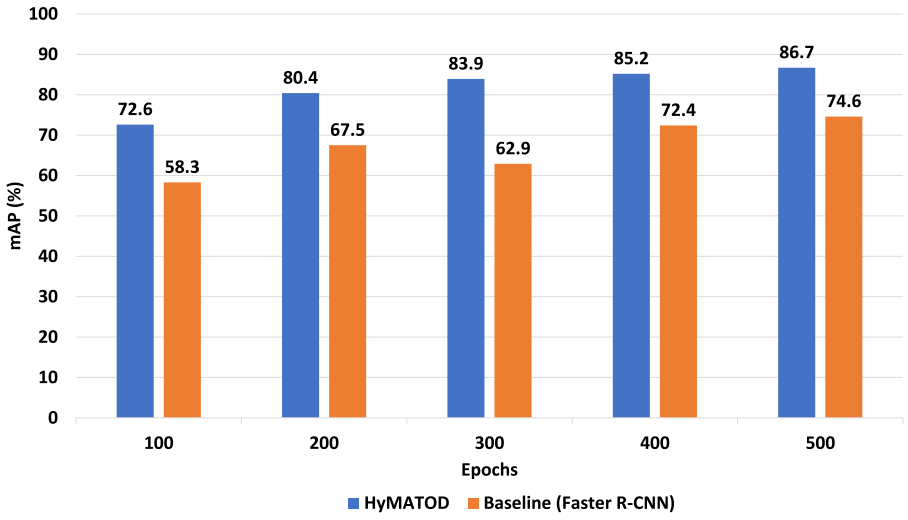


Fig. 5. Performance of HyMATOD with a baseline (Faster R-CNN (Ren et al., 2015)) on different epochs in the ImageNet VID dataset.

epoch 100, HyMATOD achieves an mAP of 72.6%, while the baseline method achieves 58.3%. As the epochs increase, both methods show improvement, but HyMATOD consistently outperforms the baseline. By epoch 200, HyMATOD reaches an mAP of 80.4%, whereas the baseline method achieves 67.5%. At epoch 300, HyMATOD's mAP further increases to 83.9%, while the baseline slightly decreases to 62.9%. When increasing the epoch 400, HyMATOD's performance improves to 85.2%, with the baseline at 72.4%. Finally, at epoch 500, HyMATOD reaches its highest mAP of 86.7%, and the baseline method achieves 74.6%. Specifically, the HyMATOD method demonstrates superior performance compared to the baseline method across all evaluated epochs, showing a significant improvement in mAP from the start to the end of the training process. This suggests that HyMATOD is more effective and efficient in achieving higher precision in comparison to the baseline method throughout the training epochs.

5.5. Statistical analysis

We calculated the 95% confidence intervals for the mean Average Precision (mAP) achieved by our Hybrid Multi-Attention Transformer (HyMAT) model on the ImageNet VID dataset as in Syed and Malathi (2023). The mAP obtained from our experiments was 86.7%, and we computed the confidence interval to assess the statistical significance of this result. The confidence interval for the mAP is calculated as follows:

$$\text{Confidence Interval (CI)} = 86.7\% \pm \text{Margin of Error (ME)}. \tag{13}$$

$$\text{(CI)} = 86.7\% \pm 1.073 \tag{14}$$

$$\text{(CI)} = (85.63\%, 87.77\%) \tag{15}$$

where the Margin of Error (ME) is determined by:

$$\text{ME} = Z \times \frac{\sigma}{\sqrt{n}} \tag{16}$$

$$\text{ME} = 2.262 \times \frac{1.5}{\sqrt{10}} \approx 1.073 \tag{17}$$

where, the standard deviation σ is 1.5%, the sample size $n = 10$, and critical value for 95% CI(Z) $\approx$ 2.262. The resulting confidence interval is (85.63%, 87.77%), indicating that we are 95% confident that the true mean mAP of our model lies within this range. This analysis confirms the statistical significance of the improvements achieved by our HyMAT model, reinforcing the reliability of our findings.

5.6. Effect on parameter analysis

We performed a comprehensive analysis of the hyperparameters to optimize the performance of our HyMATOD method, with a primary focus on the learning rate, which is a key factor in model training. We experimented with various learning rates, ranging from 0.00008 to 0.5, and evaluated the corresponding mean Average Precision (mAP) values on the ImageNet VID dataset. The results, shown in Fig. 6, indicate that a learning rate of 0.0001 achieved the highest mAP of 86.7%, outperforming the other rates tested. This rigorous exploration and fine-tuning of the learning rate were crucial for achieving optimal performance and stability in our model. These findings underscore the significance of carefully selecting hyperparameters to enhance the effectiveness and reliability of the HyMATOD method.

5.7. Comparison with state-of-the-art methods

In particular, our HyMAT model achieves an impressive 86.7% mean Average Precision (mAP) when utilizing ResNet-101, marking a substantial 12.1% absolute enhancement compared to the baseline Faster R-CNN (Ren et al., 2015) as shown in Table 5. Moreover, our HyMAT substantially exceeds FGFA (Zhu et al., 2017a) and MANet (Wang et al., 2018) which are implemented using optical flow-based methods, by 10.4% and 8.6% mAP, respectively. In comparison to certain proposal-level relation-based methods for example RDN (Deng et al., 2019b), LRTR (Shvets et al., 2019), and SELSA (Wu et al., 2019), our HyMAT demonstrates remarkable performance, outperforming them by 4.9%, 6.1%, and 2.4% mAPs, respectively.

The superiority of our approach lies in its ability to effectively explore structural relations, thereby enhancing feature aggregation quality. When compared to frame-level relation-based techniques like TCENet (He et al., 2020), our HyMAT shows a significant improvement of 6.4% in mAP. Furthermore, when evaluated against memory-guided approaches such as MEGA (Chen et al., 2020), OGEMN (Deng et al., 2019a), and MAMBA (Sun et al., 2021), our HyMATOD consistently outperforms them in detection performance. Notably, the proposed HyMAT achieves a substantially higher mAP, surpassing the newly introduced transformer-based method TransVOD (Zhou et al., 2022) by 4.8%. In the context of HVRNet (Han et al., 2020), our HyMATOD outperforms it by 3.5% mAP in object detection. This distinction arises from HVRNet's exclusive focus on feature aggregation at the proposal level, neglecting frame-level feature aggregation, which can hinder its ability to accurately identify small-scale objects.

In a direct comparison with the robust competitor MAMBA, our HyMAT demonstrates a 2.1% higher mean Average Precision (mAP).

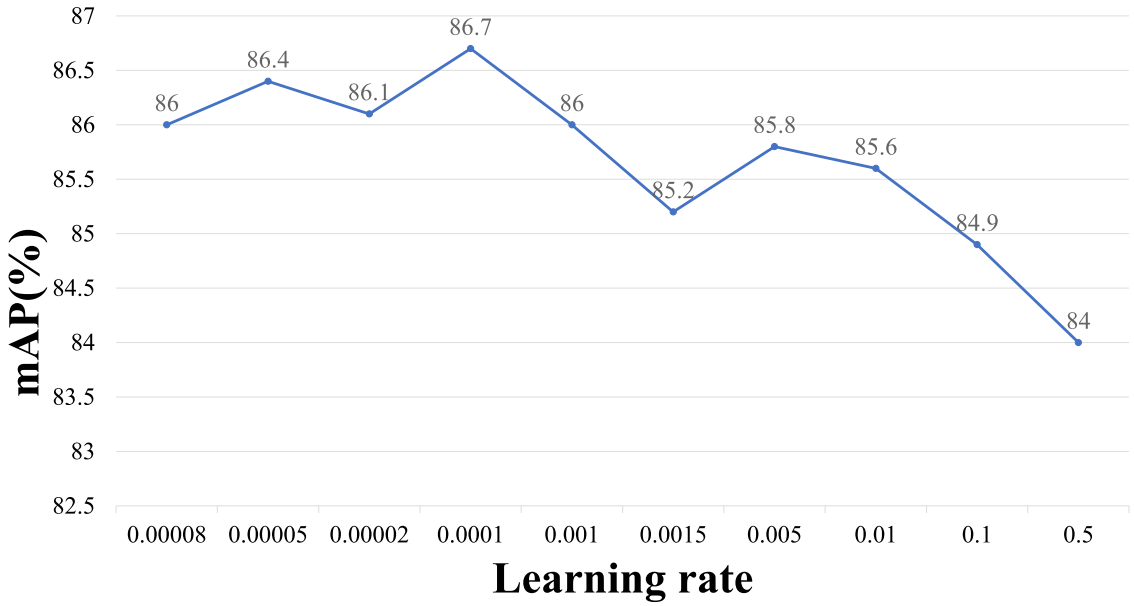


Fig. 6. Performance of HyMATOD with different learning rates on the ImageNet VID dataset.

Table 5
Comparison with state-of-the-art methods on ImageNet VID with Res101 backbone.

Methods	Base detector	Backbone	mAP_{50} (%)
Faster R-CNN (Ren et al., 2015)	–	ResNet-101	74.6
FGFA (Zhu et al., 2017a)	R-FCN	ResNet-101	76.3
CHP (Xu et al., 2020)	CenterNet	ResNet-101	76.7
MANet (Wang et al., 2018)	FPN	ResNet-101	78.1
OGEMN (Deng et al., 2019a)	R-FCN	ResNet-101	79.3
MINet (Deng et al., 2021)	Faster-RCNN	ResNet-101	80.2
TCENet (He et al., 2020)	R-FCN	ResNet-101	80.3
LRTR (Shvets et al., 2019)	FPN	ResNet-101	80.6
RDN (Deng et al., 2019b)	FPN	ResNet-101	81.1
TransVOD (Zhou et al., 2022)	DETR	ResNet-101	81.9
TSFA (He et al., 2022)	Faster-RCNN	ResNet-101	82.5
SELSA (Wu et al., 2019)	Faster-RCNN	ResNet-101	82.7
MEGA (Chen et al., 2020)	R-FCN	ResNet-101	82.9
HVRNet (Han et al., 2020)	Faster-RCNN	ResNet-101	83.2
MAMBA (Sun et al., 2021)	Faster-RCNN	ResNet-101	84.6
HyMATOD	DETR	ResNet-101	86.7

This advantage arises from MAMBA's exclusive dependence on semantic knowledge within aggregated features for object class distinction, without considering the inherent semantic information in class labels. This limitation hinders MAMBA's ability to accurately identify objects with degraded appearances. It is important to highlight that achieving performance improvements on the complex ImageNet VID dataset, particularly at higher performance levels, is a challenging task. Therefore, the 2.1% mAP improvement achieved by our HyMAT over MAMBA represents a significant and noteworthy accomplishment, given the dataset's complexity and performance ceiling.

5.7.1. Qualitative comparisons

Fig. 7 shows some example detection results produced by our HyMATOD method. As can be seen, our HyMATOD method consistently outperforms other methods in terms of detection accuracy. Fig. 7 showcases detection results from the baseline single-frame model (Faster R-CNN) and our HyMATOD under various appearance deterioration conditions. We explore three scenarios: rare poses (top two rows), object occlusion (third and fourth rows), and motion blur (last four rows). HyMATOD consistently outperforms the baseline, as evident in several examples. In the top rows, the baseline misidentifies the lion as a bear due to its unusual posture, a frequent occurrence in videos. This pose makes the lion resemble a bear, confusing the baseline and leading to incorrect recognition. HyMATOD not only corrects this error but also boosts confidence scores for accurate predictions.

Furthermore, the baseline struggles with object occlusion, failing to detect objects in the third row. This stems from its frame-by-frame processing, which cannot utilize temporal information across frames to enhance detection under challenging visibility conditions. In contrast, HyMATOD not only corrects missed detections but also strengthens reliability scores for accurate projections, as illustrated in the fourth row.

In the third frame of row five, the Faster R-CNN incorrectly identifies the squirrel as a domestic cat. Motion blur blurs the squirrel's appearance, causing severe confusion for the baseline and leading to this error. HyMATOD, empowered by its specially designed modules, effectively rectifies this misidentification, as shown in the third frame of row six. These qualitative visualizations underscore HyMATOD's efficacy in tackling demanding video situations.

6. Evaluation on UA-DETRAC dataset

Many prevailing VOD approaches undergo evaluation solely on the ImageNet VID dataset, given that it stands as the primary large-scale benchmark accessible to the public for this task. Nevertheless, the ImageNet VID dataset lacks adequacy in terms of object diversity and

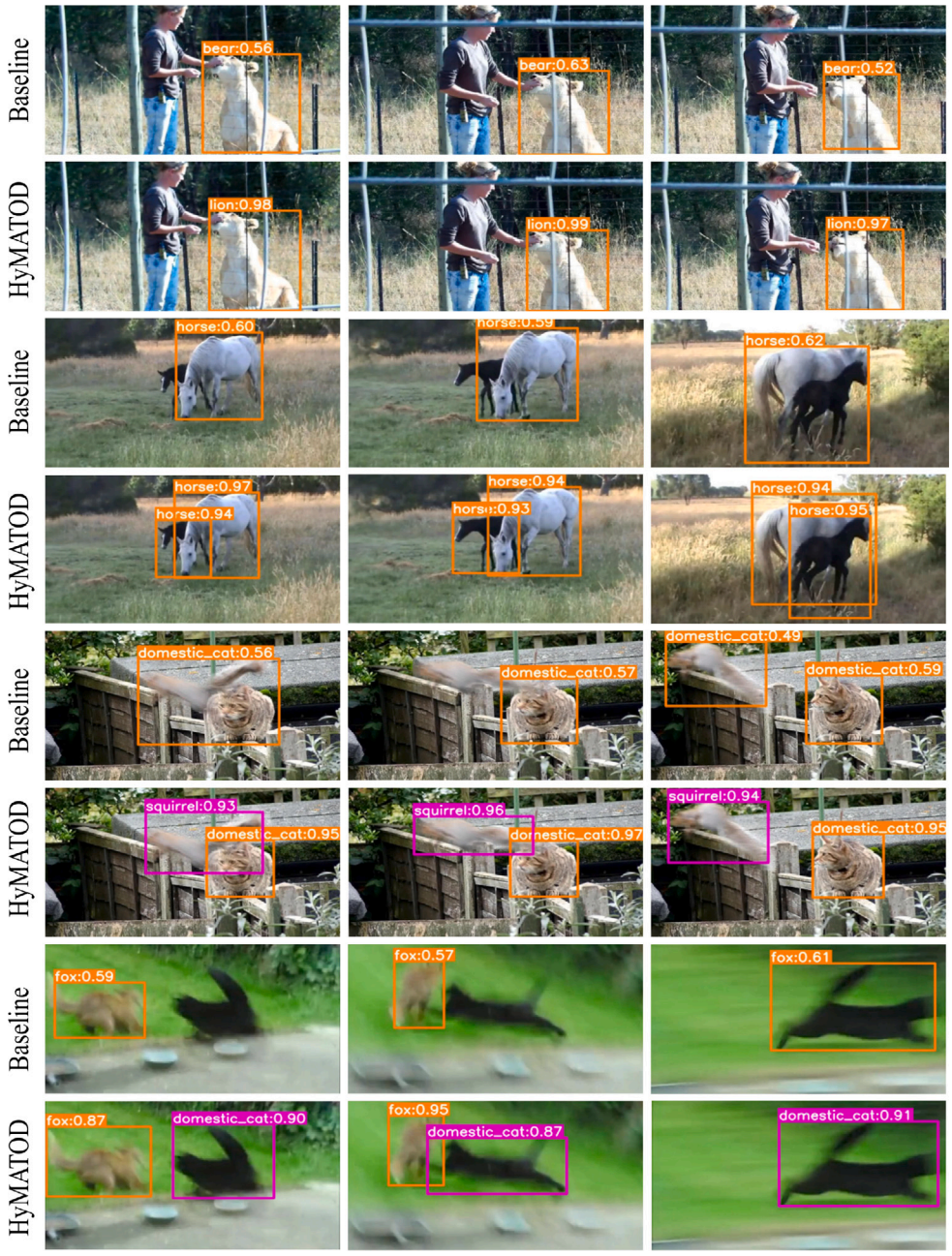


Fig. 7. Visual comparison of detection results between the baseline model and our proposed HyMATOD. Each row represents a distinct video sequence, with the baseline's results displayed in the first row and those of HyMATOD in the second. Detected objects are marked with colored bounding boxes, accompanied by their class names and corresponding confidence scores above each box. A comprehensive examination reveals HyMATOD's consistent superiority in handling challenging visual scenarios, including rare poses (first row), object occlusions (middle two rows), and motion blur (last four rows).

density. Consequently, we extend our evaluation by conducting additional experiments on the widely-used UA-DETRAC (Wen et al., 2020) dataset for the advantage of affirming the significance and inference capability of the proposed HyMATOD.

6.1. Dataset

The UA-DETRAC dataset, introduced by Wen et al. (2020), encompasses 10 h of video footage captured by a camera. This dataset represents diverse traffic patterns and conditions, encompassing urban highways, traffic crossings, and T-junctions. All videos maintain a recording rate of 25 fps. In addition, the dataset encompasses 100 difficult videos, with sixty designated for training and forty for testing. These videos boast over 140,000 manually annotated frames.

6.2. Results and analysis

In order to assess the efficacy of HyMATOD on the UA-DETRAC, we conducted a comparative analysis with four state-of-the-art methods, and the summarized results are outlined in Table 6. The primary objective is to showcase the practical applicability of our method in intelligent transportation systems. Remarkably, HyMATOD outperforms Faster R-CNN by a substantial range of 9.5% precision. This significant improvement is attributed to the detector's oversight of robust temporal dependencies across frames, which proves detrimental, particularly for objects with deteriorated appearances.

Moreover, our method achieves a notably higher mAP (+7.9%) compared to the FGFA. This improvement stems from the inefficiency

Fig. 8. The qualitative visualization of detection results from our HyMATOD on the UA-DETRAC dataset indicates its ability to reliably detect objects in video scenarios characterized by high object density.

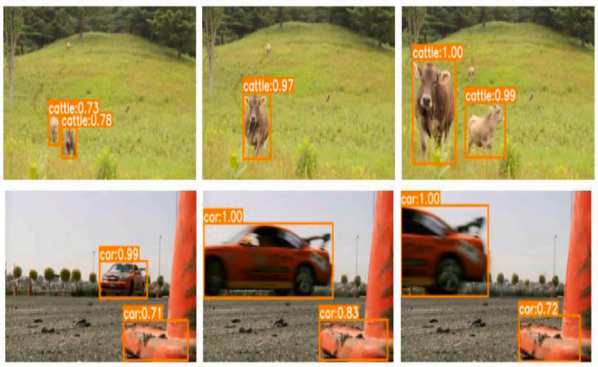


Fig. 9. Examples of failure cases in our detection system.

of optical flow-guided method in modeling long-range temporal dependencies, leading to suboptimal performance. Additionally, HyMATOD surpasses the competing methods SELSA and MEGA by 2.2% and 1.7% mAPs, respectively. HyMATOD also exhibits superior computational efficiency when compared to these methods. Unlike FGFA, which relies on computationally intensive optical flow calculations, our approach reduces overhead through the reuse of attention maps and efficient positional encoding. Furthermore, unlike SELSA and MEGA, which involve complex graph-based models that add to computational complexity, HyMATOD's dual-level graph relation modeling is designed for optimized feature aggregation without excessive computational costs. The incorporation of these elements not only enhances performance but also ensures that our method is more computationally efficient. In addition to the quantitative results, the visualization outcomes in Fig. 8 examine our HyMATOD.

Table 6
Comparison with state-of-the-art methods on UA-DETRAC with ResNet-101 backbone.

Methods	Base detector	Backbone	mAP_{50} (%)
Faster R-CNN	–	ResNet-101	79.0
FGFA	R-FCN	ResNet-101	80.6
SELSA	Faster R-CNN	ResNet-101	86.3
MEGA	Faster R-CNN	ResNet-101	86.8
HyMATOD	DETR	ResNet-101	88.5

6.3. Computational efficiency

HyMATOD demonstrates superior computational efficiency compared to previous approaches, such as Faster R-CNN (Ren et al., 2015) and traditional DETR models. Our method leverages a Hybrid Multi-Attention Mechanism that strategically allocates processing power to the most critical regions within each video frame, significantly reducing unnecessary computations. This targeted approach is further enhanced by the Reuse of Attention Maps across layers, which minimizes redundant processing and reduces the overall computational load. Additionally, our optimized transformer architecture incorporates Efficient Positional Encoding, enabling the model to capture essential spatial and temporal details without excessive computational costs. These innovations allow HyMATOD to achieve state-of-the-art performance while maintaining a lower computational burden, making it more suitable for real-time video object detection tasks compared to existing methods. This efficiency ensures that our approach is not only effective in challenging scenarios but also practical for deployment in resource-constrained environments.

6.4. Failure cases

Fig. 9 illustrates two failure cases of our detection system. The first row highlights classification ambiguity, where the system struggles to

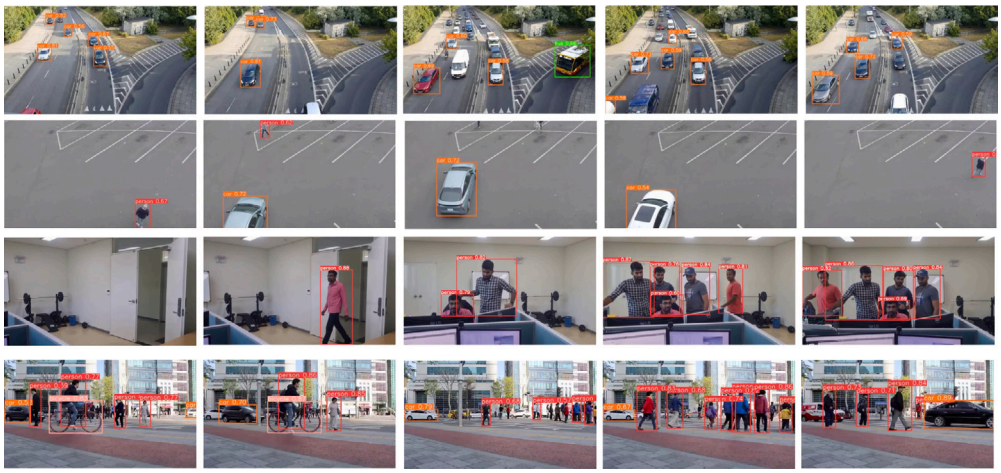


Fig. 10. Performance of the proposed method on various custom videos.

accurately identify cattle due to appearance changes. Enhancing inter-class discrepancy through contrastive learning could mitigate this issue. The second row shows missed detections for small objects, where the system successfully identifies larger objects (cars) but fails to detect smaller ones. Integrating both proposal-level and frame-level feature aggregation could improve detection accuracy for small objects.

7. Evaluation on custom videos

In addition to benchmarking against established datasets like ImageNet VID and UA-DETRAC, evaluating the HyMATOD method on custom videos is crucial for assessing its efficacy in real-world applications. Custom videos present diverse environmental conditions and object appearances that often differ significantly from standardized benchmarks. This evaluation allows for rigorous testing of the method's robustness and generalization across varied scenarios, including diverse lighting conditions, camera viewpoints, and object scales encountered in practical deployments. By conducting evaluations on custom videos representative of specific application domains such as surveillance and autonomous driving, we aim to demonstrate the method's performance under realistic conditions (refer to Fig. 10). This comprehensive assessment on custom videos not only validates the method's effectiveness but also informs further optimizations tailored to real-world challenges, ensuring its reliability and applicability beyond controlled laboratory conditions.

8. Limitations and future work

While the HyMATOD method represents a significant advancement in object detection, several considerations for its further development and practical deployment merit attention. The computational complexity inherent in the HyMAT module, while beneficial for accuracy, poses challenges for real-time applications such as surveillance and autonomous driving. Future research will focus on optimizing the module's architecture and exploring computational-efficient variants without compromising its superior performance in object detection tasks. This includes investigating model quantization techniques to reduce network parameter precision and exploring parallelization strategies leveraging GPUs or specialized accelerators like NPUs and FPGAs to enhance inference speed and responsiveness in real-time systems. Additionally, algorithmic optimizations such as network pruning and knowledge distillation will be explored to streamline inference processes and reduce computational overhead, making the HyMATOD method more feasible for deployment in resource-constrained environments. Specifically, In videos where objects undergo rapid or complex motion, such as abrupt changes in direction or speed, the temporal correlation might be less reliable. Despite our efforts to enhance temporal consistency through the HyMAT module, there could be instances where the motion prediction within distant frames becomes challenging, leading to potential detection inaccuracies. Moreover, the HyMAT module's reliance on correlations across frames may not be sufficient to capture these sudden changes, resulting in potential misalignment or incorrect tracking of objects. While the HyMATOD framework represents a significant step forward in VOD, especially in terms of leveraging temporal references and improving detection accuracy, it is essential to consider these potential limitations. Future work could explore further optimizations, such as adaptive complexity reduction, enhanced motion modeling, and improved generalization across diverse datasets, to address these challenges and ensure more robust performance across a wider range of scenarios.

9. Conclusions

In this study, we have proposed a hybrid multi-attention module (HyMAT) to tackle the inherent difficulty of independently computing correlations in the attention mechanism. The proposed HyMAT module proficiently amplifies relevant correlations and mitigates inaccurate ones by consolidating agreement across all correlation vectors. Additionally, we have introduced a simplified Transformer detection framework named HyMATOD to enhance video object detection performance. This is achieved by incorporating efficient mechanisms for feature reuse and embedding assignment, aiming to fully exploit temporal references. Finally, the extensive experiments demonstrate the effectiveness of the proposed approach. On the challenging ImageNet VID dataset, HyMATOD achieves an impressive 86.7% mAP with ResNet-101, outperforming state-of-the-art methods. Furthermore, the incorporation of both long-term and short-term references enhances the computational efficiency of our approach compared to the methods used for comparison. However, it is important to note that our method introduces additional hyperparameters, which must be meticulously adjusted for optimal performance, posing a limitation. Hence, minimizing the number of these extra hyperparameters is essential to alleviate the tuning complexity. Looking ahead, we anticipate that this research will attract more focus on the structural examination of Transformer-based trackers, object segmentation, and multi-object tracking.

CRediT authorship contribution statement

Sathishkumar Moorthy: Writing – original draft, Validation, Methodology, Conceptualization. **Sachin Sakthi K.S.:** Writing – review & editing, Visualization, Validation, Investigation. **Sathiyamoorthi Arthanari:** Writing – original draft, Investigation, Formal analysis. **Jae Hoon Jeong:** Investigation, Formal analysis. **Young Hoon Joo:** Writing – review & editing, Validation, Supervision, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the Basic Science Research Program under Grant NRF-2016R1A6A1A03013567 and Grant NRF-2021R1A2B5B01001484 and by the framework of the International Cooperation Program under Grant NRF-2022K2A9A2A06045121 through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Republic of Korea.

Data availability

No data was used for the research described in the article.

References

Ba, J.L., Kiros, J.R., Hinton, G.E., 2016. Layer normalization. arXiv preprint arXiv:1607.06450.

Cai, Z., Vasconcelos, N., 2018. Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6154–6162.

Cao, Z., Fu, C., Ye, J., Li, B., Li, Y., 2021. Hift: Hierarchical feature transformer for aerial tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15457–15466.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers. In: European Conference on Computer Vision. Springer, pp. 213–229.

Chen, Y., Cao, Y., Hu, H., Wang, L., 2020. Memory enhanced global-local aggregation for video object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10337–10346.

Chen, Y., Li, N., Zhu, D., Zhou, C.C., Hu, Z., Bai, Y., Yan, J., 2024. BEVSOC: Self-supervised contrastive learning for calibration-free bev 3d object detection. IEEE Internet Things J..

Dai, X., Chen, Y., Yang, J., Zhang, P., Yuan, L., Zhang, L., 2021. Dynamic detr: End-to-end object detection with dynamic attention. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2988–2997.

Dai, J., Li, Y., He, K., Sun, J., 2016. R-fcn: Object detection via region-based fully convolutional networks. Adv. Neural Inf. Process. Syst. 29.

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y., 2017. Deformable convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 764–773.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Ieee, pp. 248–255.

Deng, H., Hua, Y., Song, T., Zhang, Z., Xue, Z., Ma, R., Robertson, N., Guan, H., 2019a. Object guided external memory network for video object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6678–6687.

Deng, J., Pan, Y., Yao, T., Zhou, W., Li, H., Mei, T., 2019b. Relation distillation networks for video object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7023–7032.

Deng, J., Pan, Y., Yao, T., Zhou, W., Li, H., Mei, T., 2021. MINet: Meta-learning instance identifiers for video object detection. IEEE Trans. Image Process. 30, 6879–6891.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T., 2015. Flownet: Learning optical flow with convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2758–2766.

Gao, S., Zhou, C., Ma, C., Wang, X., Yuan, J., 2022. Aiatrack: Attention in attention for transformer visual tracking. In: European Conference on Computer Vision. Springer, pp. 146–164.

Gu, Y., Hu, Z., Zhao, Y., Liao, J., Zhang, W., 2024. MFGTN: A multi-modal fast gated transformer for identifying single trawl marine fishing vessel. Ocean Eng. 303, 117711.

Guo, C., Fan, B., Zhang, Q., Xiang, S., Pan, C., 2020. Augfpn: Improving multi-scale feature learning for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12595–12604.

Han, W., Khorrami, P., Paine, T.L., Ramachandran, P., Babaeizadeh, M., Shi, H., Li, J., Yan, S., Huang, T.S., 2016. Seq-nms for video object detection. arXiv preprint arXiv:1602.08465.

Han, M., Wang, Y., Chang, X., Qiao, Y., 2020. Mining inter-video proposal relations for video object detection. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16. Springer, pp. 431–446.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., et al., 2022. A survey on vision transformer. IEEE Trans. Pattern Anal. Mach. Intell. 45 (1), 87–110.

He, F., Gao, N., Li, Q., Du, S., Zhao, X., Huang, K., 2020. Temporal context enhanced feature aggregation for video object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34, pp. 10941–10948.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2961–2969.

He, F., Li, Q., Zhao, X., Huang, K., 2022. Temporal-adaptive sparse feature aggregation for video object detection. Pattern Recognit. 127, 108587.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Hu, H., Gu, J., Zhang, Z., Dai, J., Wei, Y., 2018. Relation networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3588–3597.

Jiang, B., Luo, R., Mao, J., Xiao, T., Jiang, Y., 2018. Acquisition of localization confidence for accurate object detection. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 784–799.

Jin, S., Wang, X., Meng, Q., 2024. Spatial memory-augmented visual navigation based on hierarchical deep reinforcement learning in unknown environments. Knowl.-Based Syst. 285, 111358.

Jocher, G., Chaurasia, A., 2023. Ultralytics. Accessed on Jun 9.

Kuhn, H.W., 1955. The hungarian method for the assignment problem. Nav. Res. Logist. Q. 2 (1–2), 83–97.

Li, H., Liu, Y., Liang, X., Yuan, Y., Cheng, Y., Zhang, G., Tamura, S., 2023a. Multi-object tracking via deep feature fusion and association analysis. Eng. Appl. Artif. Intell. 124, 106527.

Li, Y., Miao, N., Ma, L., Shuang, F., Huang, X., 2023b. Transformer for object detection: Review and benchmark. Eng. Appl. Artif. Intell. 126, 107021.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2980–2988.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer, pp. 740–755.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C., 2016. Ssd: Single shot multibox detector. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer, pp. 21–37.

Liu, S., Huang, D., Wang, Y., 2019. Adaptive nms: Refining pedestrian detection in a crowd. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6459–6468.

Liu, Q., Yuan, H., Hamzaoui, R., Su, H., Hou, J., Yang, H., 2021. Reduced reference perceptual quality model with application to rate control for video-based point cloud compression. IEEE Trans. Image Process. 30, 6623–6636.

Lu, X., Li, B., Yue, Y., Li, Q., Yan, J., 2019. Grid r-cnn. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7363–7372.

Pan, S., Xu, G.J., Guo, K., Park, S.H., Ding, H., 2023. Video-based engagement estimation of game streamers: An interpretable multimodal neural network approach. IEEE Trans. Games.

Qi, F., Tan, X., Zhang, Z., Chen, M., Xie, Y., Ma, L., 2024. Glass makes blurs: Learning the visual blurriness for glass surface detection. IEEE Trans. Ind. Inform..

Qiao, M., Xu, M., Jiang, L., Lei, P., Wen, S., Chen, Y., Sigal, L., 2024. HyperSOR: Context-aware graph hypernetwork for salient object ranking. IEEE Trans. Pattern Anal. Mach. Intell..

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 779–788.

Redmon, J., Farhadi, A., 2018. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Adv. Neural Inf. Process. Syst. 28.

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S., 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 658–666.

Roh, B., Shin, J., Shin, W., Kim, S., 2021. Sparse detr: Efficient end-to-end object detection with learnable sparsity. arXiv preprint arXiv:2111.14330.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115, 211–252.

Shvets, M., Liu, W., Berg, A.C., 2019. Leveraging long-range temporal relationships between proposals for video object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9756–9764.

Sun, G., Hua, Y., Hu, G., Robertson, N., 2021. Mamba: Multi-level aggregation via memory bank for video object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35, pp. 2620–2627.

Syed, S., Malathi, K., 2023. Single shot multi-box detector algorithm over fast R-CNN: An ingenious technique for increasing object detection classification accuracy. J. Surv. Fish. Sci. 10 (1S), 2193–2203.

Tian, Z., Shen, C., Chen, H., He, T., 2019. Fcos: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9627–9636.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Wang, H., Tang, J., Liu, X., Guan, S., Xie, R., Song, L., 2022a. Ptseformer: Progressive temporal-spatial enhanced transformer towards video object detection. In: European Conference on Computer Vision. Springer, pp. 732–747.

Wang, Y., Zhang, X., Yang, T., Sun, J., 2022b. Anchor detr: Query design for transformer-based detector. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36, pp. 2567–2575.

Wang, N., Zhou, W., Wang, J., Li, H., 2021. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1571–1580.

Wang, S., Zhou, Y., Yan, J., Deng, Z., 2018. Fully motion-aware network for video object detection. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 542–557.

Wen, L., Du, D., Cai, Z., Lei, Z., Chang, M.-C., Qi, H., Lim, J., Yang, M.-H., Lyu, S., 2020. UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking. Comput. Vis. Image Underst. 193, 102907.

Wu, H., Chen, Y., Wang, N., Zhang, Z., 2019. Sequence level semantics aggregation for video object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9217–9225.

Xu, Z., Hrustic, E., Vivet, D., 2020. Centernet heatmap propagation for real-time video object detection. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16. Springer, pp. 220–234.

Yan, B., Peng, H., Fu, J., Wang, D., Lu, H., 2021. Learning spatio-temporal transformer for visual tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10448–10457.

Yang, M., Cai, C., Wang, D., Wu, Q., Liu, Z., Wang, Y., 2023. Symmetric differential demodulation-based heterodyne laser interferometry used for wide frequency-band vibration calibration. IEEE Trans. Ind. Electron..

Yao, Z., Ai, J., Li, B., Zhang, C., 2021. Efficient detr: improving end-to-end object detector with dense prior. arXiv preprint arXiv:2104.01318.

Zhang, H., Liu, H., Kim, C., 2024a. Semantic and instance segmentation in coastal urban spatial perception: A multi-task learning framework with an attention mechanism. Sustainability 16 (2), 833.

Zhang, R., Tan, J., Cao, Z., Xu, L., Liu, Y., Si, L., Sun, F., 2024b. Part-aware correlation networks for few-shot learning. IEEE Trans. Multimed..

Zheng, D., Dong, W., Hu, H., Chen, X., Wang, Y., 2023. Less is more: Focus attention for efficient detr. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6674–6683.

Zhou, Q., Li, X., He, L., Yang, Y., Cheng, G., Tong, Y., Ma, L., Tao, D., 2022. TransVOD: end-to-end video object detection with spatial-temporal transformers. IEEE Trans. Pattern Anal. Mach. Intell..

Zhou, X., Wang, D., Krähenbühl, P., 2019a. Objects as points. arXiv preprint arXiv:1904.07850.

Zhou, X., Zhuo, J., Krahenbuhl, P., 2019b. Bottom-up object detection by grouping extreme and center points. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 850–859.

Zhu, X., Dai, J., Yuan, L., Wei, Y., 2018. Towards high performance video object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7210–7218.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J., 2020. Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159.

Zhu, X., Wang, Y., Dai, J., Yuan, L., Wei, Y., 2017a. Flow-guided feature aggregation for video object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 408–417.

Zhu, X., Xiong, Y., Dai, J., Yuan, L., Wei, Y., 2017b. Deep feature flow for video recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2349–2358.