Information Fusion 114 (2025) 102665

Contents lists available at ScienceDirect

Information Fusion

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

High performance RGB-Thermal Video Object Detection via hybrid fusion with progressive interaction and temporal-modal difference

Qishun Wang [a], Zhengzheng Tu [a,*], Chenglong Li [b,*], Jin Tang [a]

[a] Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei, 230601, China
[b] Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Artificial Intelligence, Anhui University, Hefei, 230601, China

ARTICLE INFO

Keywords:
Video object detection
Multi-modal fusion
RGB-thermal
Temporal difference
Hybrid strategy

ABSTRACT

RGB-Thermal Video Object Detection (RGBT VOD) is to localize and classify the predefined objects in visible and thermal spectrum videos. The key issue in RGBT VOD lies in integrating multi-modal information effectively to improve detection performance. Current multi-modal fusion methods predominantly employ middle fusion strategies, but the inherent modal difference directly influences the effect of multi-modal fusion. Although the early fusion strategy reduces the modality gap in the middle stage of the network, achieving in-depth feature interaction between different modalities remains challenging. In this work, we propose a novel hybrid fusion network called PTMNet, which effectively combines the early fusion strategy with the progressive interaction and the middle fusion strategy with the temporal-modal difference, for high performance RGBT VOD. In particular, we take each modality as a master modality to achieve an early fusion with other modalities as auxiliary information by progressive interaction. Such a design not only alleviates the modality gap but facilitates middle fusion. The temporal-modal difference models temporal information through spatial offsets and utilizes feature erasure between modalities to motivate the network to focus on shared objects in both modalities. The hybrid fusion can achieve high detection accuracy only using three input frames, which makes our PTMNet achieve a high inference speed. Experimental results show that our approach achieves state-of-the-art performance on the VT-VOD50 dataset and also operates at over 70 FPS. The code will be freely released at https://github.com/tzz-ahu for academic purposes.

1. Introduction

Recently, the task of RGB-Thermal Video Object Detection (RGBT VOD) [1] has been proposed to enhance the robustness and efficiency of video object detection. As an RGBT fusion task, effectively leveraging the complementary information provided by both modalities is essential to ensure the model's robustness and accuracy. Existing fusion methods are commonly categorized into data-level fusion (early fusion), feature-level fusion (middle fusion), and decision-level fusion (late fusion) [2,3].

Most existing fusion methods solely concentrate on feature-level fusion (middle fusion) as illustrated in Fig. 1 (b) [4,5]. While middle fusion methods effectively integrate the high-level semantic information from both modalities, they overlook the influence of existing modal differences on the fusion process. Conversely, another commonly employed strategy is data-level fusion (early fusion) as demonstrated in Fig. 1 (a) [6,7]. The two modalities share raw information with each other. However, as illustrated in the study [3], conventional early fusion often leads to a network's multiple modal branches being merged into a single branch prematurely. This approach typically falls short in achieving deep fusion at the feature level. In order to show the impact of different strategies on the distribution of multimodal features, we choose 20 pairs of RGBT images to analyze feature distribution after different fusion strategies. Specifically, we apply Principal Component Analysis (PCA) [8] to their feature maps output from the backbone network, for analyzing the feature distribution. To enhance clarity, we specifically choose the first 5 channels from each feature map to plot five feature points and utilize a standardized coordinate system to present the feature distribution. As depicted in Fig. 2, employing the middle fusion strategy alone results in a scattered distribution of multi-modal features with significant differences. In contrast, employing the early fusion method alone results in more clustered feature points across different modalities. Here, the features from both modalities have a more compact distribution separately resulting in a smaller difference between two feature clusters. This observation shows a reduced disparity between modalities. Some methods [9] attempt to integrate both strategies, however, this integration often necessitates

* Corresponding authors.
E-mail addresses: qishunahu@163.com (Q. Wang), zhengzhengahu@163.com (Z. Tu), lcl1314@foxmail.com (C. Li), tangjin@ahu.edu.cn (J. Tang).

https://doi.org/10.1016/j.inffus.2024.102665
Received 3 June 2024; Received in revised form 19 August 2024; Accepted 1 September 2024
Available online 12 September 2024

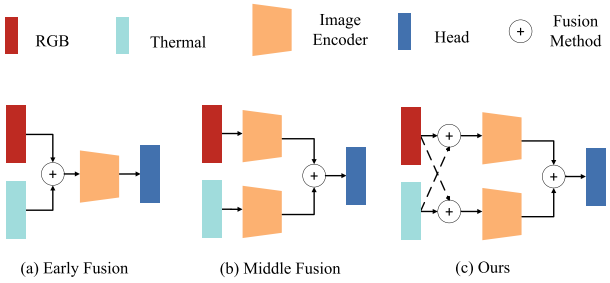

Fig. 1. Two prominent strategies that have attracted significant attention are denoted as (a) for the early fusion strategy and (b) for the middle fusion strategy. Our proposed fusion strategy, indicated as (c), combines the strengths of the above two strategies.

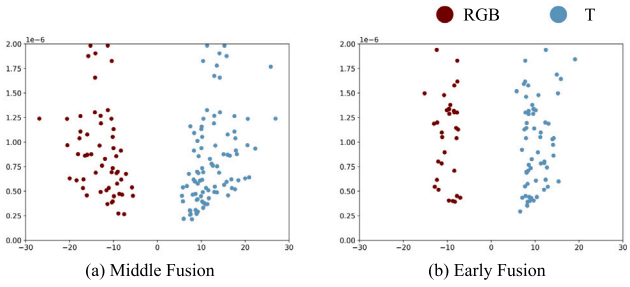

Fig. 2. The distribution of multi-modal features with different fusion strategies.

the introduction of additional branches, which results in an increase in both computational cost and model complexity. Furthermore, these methods disregard the domain dissimilarities between the original RGB branch and the thermal branch. Therefore, we propose to construct a network for early progressive fusion and middle deep fusion. Late fusion typically employs two separate subnetworks to comprehensively learn the features of two modalities and provide two sets of prediction results, which results in significant computational and maintenance costs. We think that the fusion approach is unsuitable for our video-level detection task, hence, it will not be further addressed here. Fig. 1(c) illustrates the general process to realize our idea. It is worth mentioning that, by employing progressive interaction in the early stage, the modal differences are reduced prior to RGBT middle fusion, alleviating inter-modality feature inconsistency during the middle stage.

In addition to improve multi-modal fusion to enhance the precision, efficiency is also crucial for RGBT VOD. Recent RGB-based VOD approaches tend to focus on accuracy rather than speed. This is due to the fact that traditional VOD methods, such as SELSA [10], Temporal RoI Align [11], and BoxMask [12], mostly rely on two-stage detectors [13], which inherently have slower speeds compared with one-stage detectors [14,15]. Additionally, most models select auxiliary frames from the preceding and succeeding frames of the current frame to assist detection [16–18]. However, these methods require the storage and maintenance of these auxiliary features, resulting in computational expenses. YOLOV [19] and EINet [1] are built upon the one-stage detector YOLOX [20], leading to a substantial improvement in inference speed. However, despite having these advancements, they fail to detect objects online, so they cannot meet the demand of real-time applications [21]. Therefore, our method is built based on a one-stage detector for a good inference speed. For the purpose of achieving online detection, we discard subsequent frames and utilize the previous frame and temporal difference to capture spatio-temporal information.

By combining the aforementioned points, we construct a novel RGBT VOD network via Progressive interaction and Temporal-Modal difference fusion Network, called PTMNet. The initial asymmetric Progressive Interaction Fusion (PIF) module prevents merging multiple inputs of the network prematurely into a single branch and introduces differences between modalities for initial processing. Subsequently, these differences are fed into the robust backbone network, leading to further reduction. In the middle stage, Temporal-modal Difference Fusion (TDF) module engages in deep feature interaction through the creation of modal difference states and temporal difference states. Specifically, it employs modal differences to minimize discrepancies between RGB and T features while capturing complementary information simultaneously. On the other hand, temporal difference is utilized to suppress noise and model temporal and spatial motion relationships.

PTMNet is validated using the VT-VOD50 dataset [1] and compared with other existing methods. The VT-VOD50 dataset is currently the sole existing dataset specifically gathered for RGBT VOD, ensuring a comprehensive evaluation for the method's performance. The results indicate that our method has advantages in terms of performance and efficiency. In summary, the contributions of this study can be summarized as follows:

- Our proposed PIF module not only facilitates interactions between primitive information but also preserves the multi-branch structure of the network, enabling all branches to equally and comprehensively contribute to reducing modal differences.
- Our proposed TDF module utilizes sparse temporal-modal differences to guide the feature fusion process. This approach effectively reduces the influence of noise in the fusion process and accurately models temporal information. Additionally, we incorporate the MLP branch to assist the CNN branch in emphasizing global information throughout the fusion process.
- The proposed PTMNet incorporates both of the aforementioned strategies to ensure the consistency of multi-modal features during the process of network learning. PTMNet achieves state-of-the-art (SOTA) performance on the VT-VOD50 dataset. Additionally, PTMNet can perform online detection while maintaining a real-time detection capability of 70+ FPS.

2. Related work

We explore recently published RGBT fusion methods in visual tasks and VOD methods. The following subsections will provide a detailed introduction and analysis of these.

2.1. RGB-thermal fusion

For combining multi-modal images, three strategies are commonly classified, that are early fusion, middle fusion, and late fusion [3].

For instance, in this study [7], a weighted fusion approach is employed to combine pixel information from input images. However, this approach cannot perform feature-level fusion. Tu et al. [22] introduce a dual-decoder architecture that enables multi-scale interaction between features from two modalities within the middle of the network. Additionally, DCNet [23] leverages the output features of two backbone networks for spatial transformation and semantic interaction. However, the multi-modal features employed by these methods exhibit significant modal disparities. By directly fusing features with dissimilarities, the final result can be adversely impacted by the resulting inconsistencies and weak correlations among the features. ProbEn [24] employs a decision fusion strategy to perform post-processing on the results obtained from each modality. However, this approach heavily depends on the robustness of the initial results. Additionally, the computational cost is relatively higher. Huang et al. [25] present an unsupervised training framework for RGBT image saliency detection, targeting the issue of scarce training samples. Their approach integrates early fusion and middle fusion techniques, introduces multi-graph fusion, and extracts crucial information from diverse data sources to enhance learning outcomes. HMFT [9] attempts to incorporate all three of these strategies. However, it fails in achieving actual early fusion at the input side.

Moreover, the features utilized for fusion in the middle stages still exhibit inconsistency in their expression.

Unlike the aforementioned methods, PTMNet integrates two effective fusion strategies that are early fusion and middle fusion to minimize modality disparities and comprehensively leverage complementary information between modalities. This involves initial coarse-grained interaction at the pixel level followed by fine-grained deep fusion at the feature level.

2.2. Video object detection

The problem of video object detection is very significant due to its practical application, such as in autonomous driving [26,27].

FGFA [28] utilizes neighboring frames to optimize the current frame, resulting in a better detection result, although without the online detection capability. Wang et al. [29] expand the FGFA approach by incorporating feature aggregation across proposals, thus enhancing performance. In contrast to the conventional approach of employing optical flow networks, Jiang et al. [30] employ Learnable Spatio-Temporal Sampling (LSTS) to precisely acquire the semantic aggregation principles between consecutive frames. They also utilize random sampling and iterative algorithms to enhance the quality of the present frame. SELSA [10] conducts semantic matching and feature enhancement by randomly selecting frames globally. However, this temporal-agnostic approach creates new challenges, such as the inability to locate trustworthy semantically similar information. Yao et al. [31] integrate a real-time tracker into the VOD framework to enhance operational efficiency and ensure temporal consistency, successfully enabling real-time processing on the central processing unit (CPU). Temporal ROI Align [11] and BoxMask [12] effectively acquire discriminative information by integrating the instance-level and the pixel-level features. After previous researches [10,28] propose models to depict the relationship between auxiliary frames and the current frame, TF-Blender [32] introduces a novel approach by integrating information exchange among auxiliary frames, leveraging temporal cues effectively. He et al. [33] develop an object query propagation (QueryProp) framework to investigate the propagation of instance-level features. The QueryProp utilizes sparse key frames to transfer features to dense non-key frames, effectively minimizing redundant computation. Yet these methods rely on future frames and are essentially offline detection models. Sun et al. [17] devise a multi-level aggregation framework that is executed using memory banks to integrate instance-level features for precise fine-grained matching and enhancement. EINet [1], designed for the RGBT VOD task, exhibits a high detection efficiency, although more data that are thermal images are introduced and processed. However, its reliance on future video frames hinders the real-time realization for RGBT VOD task.

However, the short temporal window design of our PTMNet requires minimal dependence on auxiliary frames, and its feature-level interaction design prevents ambiguity in handling complex and similar multi-objects, thus achieving an optimal balance between efficiency and performance.

3. Method

We build the framework on the efficient one-stage detector YOLOX [20], which is more suitable for the video task than existing two-stage and transformer-based methods [12,34], as YOLOX is faster and flexible.

3.1. General overview

The overall framework of the network is shown in Fig. 3. Three frames enter PIF in the corresponding branch. After passing through CNNs again, these feature maps are fused in the TDF module for deep interaction. Finally, the inference results are obtained after a multi-scale decoupled detection head.

3.2. Progressive interaction fusion

To minimize the differences between features that need to be fused in the middle stage, we introduce a progressive interaction fusion module before the CNN. This module enables the features of the current branch to receive information from other branches. As the CNN proceeds with feature encoding, the module considerably diminishes domain discrepancies, thereby facilitating more effective fusion during the middle stage.

The flow of PIF is shown in Fig. 3. PIF receives three different inputs. Specifically, the solid line entering represents the input X. X guides the direction of feature encoding as the main input. The dashed lines entering PIF correspond to the inputs Y and Z. They are responsible for introducing information from the other two branches equally. We choose Space-to-Depth (SPD) [35] to process the input three frames so that we can conduct a down-sample while fully preserving the original information. The purpose of this step is to reduce the following calculation. We concatenate X, Y, and Z and perform feature transformation on it by N consecutive multi-layer perceptrons (MLPs). The selection of N will be shown and analyzed in detail in the experiments section. In order to ensure that the main input X has absolute bootstrap for the current branch, we need to keep avoiding excessive feature shifts by reintroducing X during the transformation.

The proposed structure allows for the exchange of information between the features of all three branches, facilitating the initial fusion process while preserving the features of each branch. PIF integrates information across modalities and incorporates both complementary information and domain-specific disparities. Specifically, it tackles challenging domain differences by inputting them into the backbone network, thereby easing the burden of deep feature interaction in the middle stage of the network.

3.3. Temporal-modal difference fusion

Feature-level semantic interaction is critical for the fusion of RGB and thermal inputs. Therefore, we have designed a feature fusion module in the middle of the network that is capable of focusing on both temporal and RGBT information.

The specifics of TDF are illustrated in Fig. 4. The key point lies in TDF utilizing these three inputs to construct a cue tensor, which effectively guides the encoding within the module. The tensor is f_t^{dif} in Fig. 4.

After getting the temporal-modal difference tensor, we process each of the four tensors by extracting the maximum and average values across the channels. This procedure yields eight tensors that represent the respective spatial distributions of these inputs. Initially, we concatenate these eight tensors along the channel axis to produce an output, denoted as $I_1 \in \mathbb{R}^{B\times 8\times H\times W}$, where B represents the input batch size, H and W represent the height and width of the feature map. Conversely, we derive another output, $I_2 \in \mathbb{R}^{B\times 4\times H\times W}$, by discarding them along the channel axis with a stride of 1 from I_1. This approach serves to differentiate between the two and mitigate the network's tendency to learn redundant information. Then I_1 and I_2 are distributed into two separate parallel branches. The first branch employs MLP for feature transformation, while the second branch utilizes the dilated encoder proposed in YOLOF [36]. MLP is capable of modeling long-distance dependencies among spatial pixels, however, it does not take into account the spatial positional structure. Conversely, CNN can capture a limited range of spatial structures but fails to focus on distant spatial locations.

By combining MLP and CNN in parallel, we effectively leverage the strengths of both structures to learn the mapping relationship encompassing the spatial distribution of the input tensor. Following the aforementioned branch into two sub-branches, we obtain two tensors with a channel length of 1. These tensors undergo multiplication and

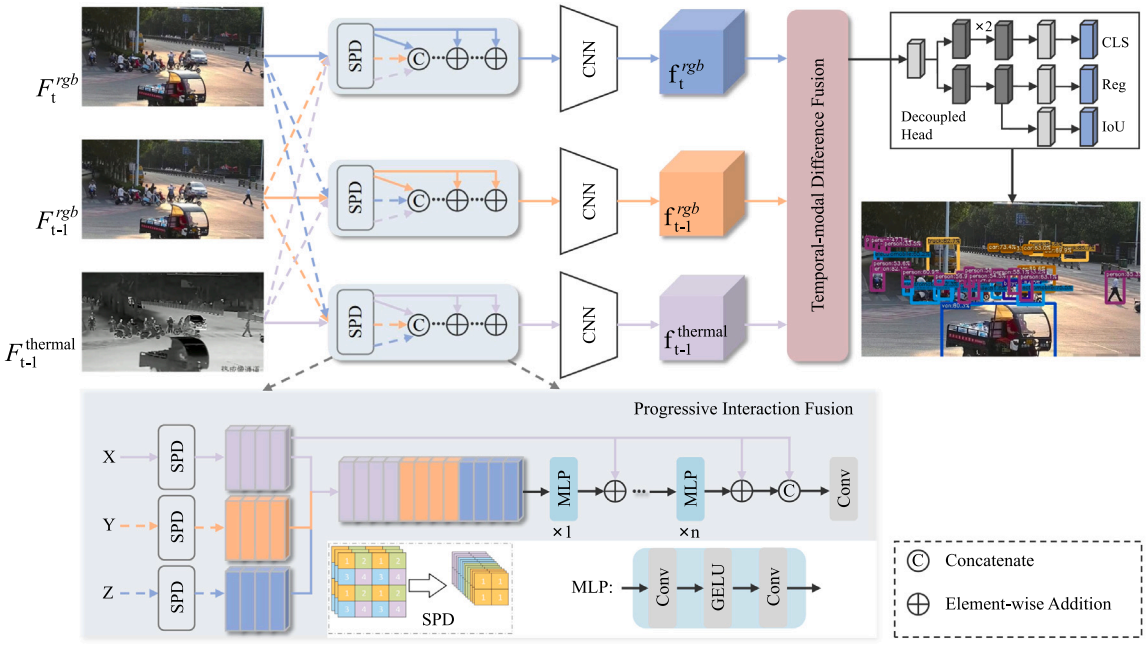

Fig. 3. The framework diagram of the proposed PTMNet.

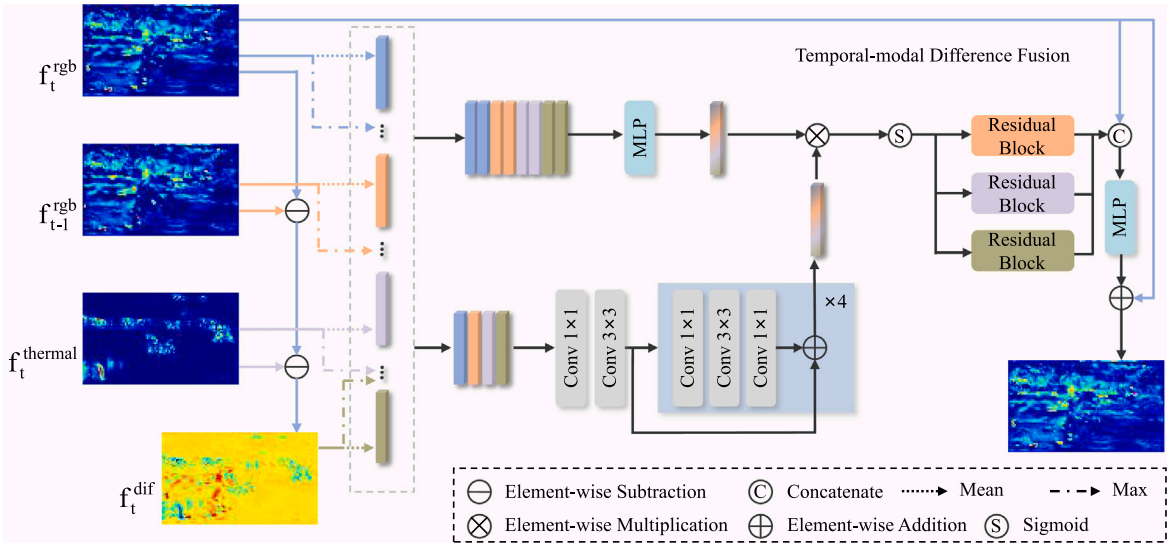

Fig. 4. The implementation of the middle fusion strategy. After the extraction of the spatial distribution from all four inputs, parallel learning takes place on two branches. The top branch utilizes MLP to establish mappings between spatial distributions, while the lower branch employs convolutional networks to capture local spatial relations. By taking advantage of the strengths of both branches, the model comprehensively captures local and global structural relationships within the input features.

then a sigmoid activation function [37], resulting in forming a fine-tuning weight. This weight is utilized to adjust f_{t-1}^{rgb}, $f_t^{thermal}$ and f_t^{dif} through a residual structure, with the objective of emphasizing specific features that are close to f_t^{rgb}. The process can be represented by the equation below:

$$W = \sigma(W_1 * W_2), \tag{1}$$

$$f^* = W * f + f. \tag{2}$$

In this context, σ represents the sigmoid activation function. W_1 and W_2 represent the outcomes obtained from MLP and dilated encoder, respectively. And "*" represents element-wise multiplication. Where f can stand for f_{t-1}^{rgb}, $f_t^{thermal}$ and f_t^{dif}. These results are caused by the process outlined in Eq. (1) to generate refined weights W, which are subsequently utilized in the residual learning process described in Eq. (2) to derive the updated features. The reason for fine-tuning all feature maps using the same weight W is that they should be highly consistent in content relative to the labeled unique ground truth to ensure that the features of the object region are enhanced as they should be. The process of Eq. (2) is the Residual Block in Fig. 4. Next, a tensor is formed by concatenating the feature maps f_t^{rgb}, f_{t-1}^{rgb}, $f_t^{thermal}$ and f_t^{dif}. This tensor undergoes the processing of M consecutive MLPs, sequentially transforming the channels to match the size of f_t^{rgb} to align with the original design.

It is important to note that the composition of f_t^{dif} significantly influences the encoding direction of the difference fusion module. As demonstrated in Fig. 4, f_t^{dif} is actually determined by the subtraction of f_{t-1}^{rgb} and $f_t^{thermal}$ from f_t^{rgb}. This choice is motivated by the proven significance of temporal differences in video tasks [38–40]. By incorporating spatial variation in a display, the network can model

Table 1
We evaluate PTMNet and the current mainstream detection models simultaneously on VT-VOD50, and we highlight the best results in **bold**.

Methods	Backbone	Type	Fusion stage	AP50 (%)	AP (%)	FPS	Params (M)	FLOPs (G)
YOLOV3 [43]	Darknet53	Image	Early	33.9	17.4	217.6	61.5	193.9
YOLOV5_L [44]	CSPDarknet53	Image	Early	33.2	18	142.9	46.5	109.1
CFT [45]	CFB	Image	Middle	42.5	18.9	222.2	73.7	–
YOLOX [20]	Darknet53	Image	Early	41.2	20.6	263.9	**8.9**	**26.8**
YOLOV6_L [46]	EfficientRep	Image	Early	36.8	19.1	72.5	59.6	150.7
TransVOD++ [34]	SwinB	Image	Early	44.4	23.7	8.5	98.7	736.7
YOLOV7 [47]	CSPDarknet53	Image	Early	37.7	16.5	**294.1**	36.5	103.3
YOLOV9-C [48]	GELAN	Image	Early	49.1	26.9	166.7	25.4	103.2
YOLOV10-M [49]	CSPNet	Image	Early	46.2	25.2	210	16.5	64
TOOD [50]	ResNet-50	Image	Early	36.3	19	25.8	32	199
Deformable DETR [51]	ResNet-50	Image	Early	42.5	23.3	20.7	41.1	197
RT-DETR [52]	ResNet-50	Image	Early	40.2	21.6	–	42.7	130.5
DINO [53]	ResNet-50	Image	Early	47.4	25.9	16.7	47.7	274
DDQ DETR [54]	ResNet-50	Image	Early	48.3	26.5	13	48.3	275
DiffusionDet [55]	ResNet-50	Image	Early	46.9	25.1	–	–	–
DFF [56]	ResNet-50	Video	Early	33.5	14.1	40.4	62.11	24.9
FGFA [28]	ResNet-50	Video	Early	35.1	15.8	9	64.47	41
RDN [57]	ResNet-50	Video	Early	40	–	11.3	–	–
SELSA [10]	ResNet-50	Video	Early	39.4	17.4	10.5	–	–
MEGA [16]	ResNet-50	Video	Early	27.8	–	16.2	–	–
Temporal ROI Align [11]	ResNet-50	Video	Early	38	17	5.1	–	–
CVA-Net [58]	ResNet-50	Video	Early	39.7	19.7	6.9	41.6	548.1
STNet [59]	ResNet-50	Video	Early	38.4	18.4	5	41.6	752.3
EINet [1]	Darknet53	Video	Middle	46.3	24	204.2	11.6	78.2
PTMNet (Ours)	CSPDarknet53	Video	Early&Middle	**51.4**	**27.07**	72.5	11.43	77.67

temporal information effectively. Moreover, extending this idea into RGBT modalities introduces the concept of modal difference, which is the original intention behind constructing f_t^{dif}. The aim has two points: guiding the network to accurately capture temporal information and facilitating the network focus on the similarities and differences between RGB and thermal features. Notably, f_t^{dif} is a sparse tensor that filters out most noise.

3.4. Label assignment and loss function

We adopt the label assignment strategy as same as the baseline method, known as SimOTA [20]. This method represents an enhancement and refinement of OTA [41], promoting faster processing and eliminating the need for additional hyperparameters. SimOTA identifies potential positive sample regions around the object center, and the dynamic allocation head can intelligently determine the optimal number of positive samples allocated. Subsequently, this study calculates the losses on the predicted boxes after allocating both positive and negative samples. The loss calculation involves three specific components denoted as L_{iou}, L_{obj}, and L_{cls}. Components L_{iou} and L_{cls} pertain to the positive sample prediction boxes corresponding to true values, while component L_{obj} pertains to all prediction boxes. Among these, component L_{iou} employs the traditional IoU Loss [42], whereas components L_{obj} and L_{cls} both utilize the BCEWithLogitsLoss. Consequently, the comprehensive loss function is defined as presented in Eq. (3):

$$L = \lambda L_{iou} + L_{obj} + L_{cls}. \tag{3}$$

4. Experiments

In this section, we present a series of comparison and ablation experiments conducted for the PTMNet. Through a detailed analysis and comparison, we ultimately demonstrate the superiority and effectiveness of our proposed method.

4.1. Implementation and metrics

Our network is implemented on two RTX 3090 GPUs using the PyTorch deep learning framework. The batch size for both training and inference is set to 2 per GPU. We use Stochastic Gradient Descent (SGD) with a momentum of 0.9 and weight decay of 0.0005 during training. The dataset used for the experiments is currently the only RGBT VOD dataset available, namely the VT-VOD50 dataset. It is noteworthy that, to uphold the experiment's integrity, we excluded additional data augmentation methods and preserved solely the fundamental mirror augmentation. This approach also guaranteed that the network's ability to perceive the content of RGBT image pairs remained unaffected.

The main metrics utilized to evaluate the methods' performance and efficiency include AP (Average Precision), AP50, and FPS. AP is calculated by averaging the IOU (Intersection over Union) values ranging from 50% to 95% in increments of 5%. AP50 corresponds to an IoU threshold of 0.5. FPS (frames per second) is a crucial metric for assessing the efficiency of a video analysis method.

4.2. Comparative analysis for experimental results

A series of comparative experiments and analyses are conducted to compare PTMNet with existing mainstream detection models. The results are presented in Table 1. It is important to note that all the comparison methods that utilize the early fusion strategy employ the addition of RGB and thermal images to enable the retention of all original information for network learning. Meanwhile, the comparison methods employing the middle fusion strategy utilize multi-modal encoders with different designs. Notably, our PTMNet is capable of implementing both fusion strategies. PTMNet consistently outperforms other methods, demonstrating an improvement of 2.3% in performance compared with the second-best method YOLOV9 [48] on AP50. In terms of inference speed, PTMNet outperforms all two-stage detectors and only behind CFT, EINet and YOLO family [20,43,44,47–49], because they do not contain temporal information in the videos, which also makes their performance far behind PTMNet. Moreover, PTMNet is an online detector, that performs real-time object detection well at standard video frame rates, such as 24fps, 30fps, and 60fps. This advantage makes it significantly superior to other methods.

Some detailed results are also shown in Table 1. For instance, one of DFF [56] and FGFA [28] focuses on speed while the other focuses more on accuracy, which is consistent with the motivations proposed by their designers. Notably, the RDN [57] previously introduced outperforms certain recently published approaches on the VT-VOD50 dataset [1]. This advancement can be attributed to its reliance on supporting frames within a narrow temporal window, contrasting SELSA's [10] emphasis on the overall context. Given the dataset's complexity with numerous

Table 2
Experimental results from integrating PTMNet with six classic backbone networks. Where "♦" signifies the utilization of PTMNet.

Methods	Backbone	AP (%)	Params (M)	FLOPs (G)
PTMNet	ResNet [62]	23.57	5.97	**49.92**
	VGG [63]	23.5	12.04	170.98
	GELAN [48]	22.66	**4.32**	60.45
	Darknet [43]	23.8	8.53	56.15
	EfficientRep [64]	22.73	14.43	89.37
	MobileNetV2 [65]	22.73	13.14	82.69
	♦ CSPDarknet53	**27.07**	11.43	77.67

similar objects, a global matching approach like SELSA may introduce errors. Similarly, MEGA's [16] subpar performance may stem from the instability of its Long Range Memory module when confronted with numerous similar features. Temporal ROI Align [11], CVA-Net [58] and STNet [59] have similar performance and are also relatively close in speed. As image detectors, the speed performance of the YOLO series is consistent with its advantages as one-stage detectors. However, most of them rely on complex data enhancement such as mosaic [60], which is inconsistent with the high consistency of image content required by RGBT fusion. The transformer-based TransVOD++ [34] and DETR series [51–54] detectors exhibit performance at the upper-middle level. However, their efficiencies are not satisfactory.

In multiple sets of experiments, TransVOD++ [34] utilized the pre-trained weights from ImageNet dataset [61], while the other methods are trained from scratch. Except for EINet and PTMNet, the remaining methods integrate thermal images with RGB images at the network entrance by pixel-level summation. In this scenario, methods that lack the RGBT fusion strategy experience a decline in performance when incorporating thermal images. This basic fact exemplifies the essential need and efficiency of our multi-stage fusion strategy proposal.

In summary, PTMNet surpasses all two-stage detectors in speed and outperforms all comparison methods in accuracy. It has established a new state-of-the-art (SOTA) record on the VT-VOD50 dataset and possesses real-time online detection capability.

4.3. Impact of different backbones on PTMNet

To investigate the influence of various feature extraction backbones on PTMNet's overall performance, we respectively integrate six traditional backbone networks within PTMNet, instead of CSPDarknet53. The results are presented in Table 2.

Upon substituting PTMNet's backbone with classic image encoders like ResNet [62] and MobileNet [65], we observe the varied reductions in overall performance. We attribute the advantage of CSPDarknet53 to its long-distance residual structure, which facilitates repeated correction and rectification of features after PIF, thereby enhancing middle feature interaction. Networks such as ResNet and Darknet [43] employ solely local residual connections, whereas VGG [63] and GELAN [48] lack entire residual connections.

4.4. Impact of different inputs on PTMNet

Regarding the problem of why PTMNet does not employ $f_{t-1}^{thermal}$ as an input to the network to ensure full symmetry between the two modal inputs, we conduct an experiment using a symmetric architecture to process input from two modalities, consisting of four frames of images. We subsequently adapt the PIF and TDF. The experimental results are presented in Table 3.

Table 3 clearly illustrates that the addition of $f_{t-1}^{thermal}$ to the input not only results in a nearly 50% reduction in the network's speed but also diminishes its precision. We conduct a thorough analysis for the reason. RGBT VOD employs RGB-based annotation of the current frame as the ground truth, emphasizing PTMNet's role in enhancing feature representation of f_t^{rgb}. f_{t-1}^{rgb} and $f_t^{thermal}$ maintain semantic and spatial consistency with f_t^{rgb}, respectively, whereas $f_{t-1}^{thermal}$ diverges from f_t^{rgb} in both aspects. Consequently, the introduction of $f_{t-1}^{thermal}$ not only fails to enhance network precision but also disrupts the feature learning of f_t^{rgb}. Additionally, the increase in computational cost further reduces network efficiency.

Table 3
Experimental results from integrating PTMNet with six classic backbone networks. Where "♦" signifies the utilization of PTMNet.

Methods	Input	AP (%)	AP (%)	FPS
PTMNet	$f_{t-1}^{thermal}$, $f_t^{thermal}$, f_{t-1}^{rgb}, f_t^{rgb}	47.47	24.58	39.1
	$f_t^{thermal}$, f_{t-1}^{rgb}, f_t^{rgb}	**51.4**	**27.07**	**72.5**

Table 4
These experiments aim to demonstrate the effectiveness and performance of the individual components.

Groups	PIF	TDF	Params (M)	AP (%)	AP50 (%)	FPS
(a)			**6.11**	23.9	46.5	**277.8**
(b)	✓		6.12	24.93	48.52	179.9
(c)		✓	11.42	24.95	48.17	110.9
(d)	✓	✓	11.43	**27.07**	**51.4**	72.5

Table 5
A set of experiments is conducted on the number N of MLPs used in the PIF module. Group "♦" signifies the utilization of PTMNet.

N	Gflops	AP50 (%)	AP(%)
0	36.71	47.57	24.5
1	**34.48**	47.43	24.5
♦ 2	34.74	**48.52**	**24.93**
3	34.51	47.44	24.44
4	34.55	46.48	24.23
5	34.60	45.42	23.39

4.5. Ablation studies

The rationality of our framework and the effectiveness of the modules have been further demonstrated through several sets of ablation experiments.

4.5.1. Fusion modules in PTMNet

Table 4 exhibits a sequential illustration of the network's development from the baseline. The result presented as (a) represents our baseline test using RGB data from the VT-VOD50 dataset. (b) is the outcome of incorporating PIF onto (a) and it is noteworthy that PIF incurs a minimal computational cost in return for a substantial improvement in performance. The experimental setting for the (c) group is to add TDF on (a). The improvements in both the AP and AP50 indicators prove the effectiveness of TDF. Lastly, the simultaneous integration of PIF and TDF into the network forms our PTMNet and its performance is demonstrated in group (d).

Table 4 demonstrates that (b) and (c) show an average improvement of 1.85% compared to the baseline (a), validating the effectiveness of the PIF and TDF modules, respectively. When used in conjunction, group (d)'s performance is enhanced by an average of 3.06% in comparison to (b) and (c). It is widely understood that simply stacking two modules does not guarantee mutual enhancement, they can even impede each other's progress. The ablation results of PTMNet confirm the promising potential of our proposed strategy involving the mutual enhancement of early fusion and middle fusion.

4.5.2. Selection of N in PIF

In the PIF, N consecutive MLPs are utilized for feature transformation. Our experiment involves multiple sets to determine the ideal value of N, as detailed in Table 5. Notably, the optimum performance is achieved with N set to 2. For N values of 0, 1, and 3, PIF demonstrates

Table 6
A set of experiments is conducted on the number M of MLPs used in the TDF module. Group "♦" signifies the utilization of PTMNet.

M	Gflops	AP50 (%)	AP (%)
1	**76.44**	46.94	24.37
2	76.93	46.33	24.00
♦ 3	77.67	**51.4**	**27.07**
4	77.91	44.47	23.04
5	78.16	46.16	23.77

Table 7
A series of ablation experiments are conducted on feature fusion structures in TDF, where "x" denotes f_{t-1}^{rgb}, $f_t^{thermal}$ and f_t^{dif} in Section 3.3, and "w" corresponds to the weight W in Section 3.3. Group "♦" signifies the utilization of PTMNet.

Structure	Gflops	AP50 (%)	AP (%)
x	77.67	45.50	23.5
wx	77.67	46.02	23.85
(wx+x)†	77.67	46.45	24.05
♦ (wx+x)	**77.67**	**51.4**	**27.07**

comparable performance. However, accuracy starts to decline with the usage of 4 and 5 MLPs. This suggests that an excessive number of mapping operations will hinder network learning.

4.5.3. Selection of M in TDF

To assess the efficacy and rationale of utilizing M consecutive MLPs in the final stage of TDF, a range of ablation experiments were performed across various M values, as detailed in Table 6. PTMNet selects an M value of 3. Deviating from this value, whether higher or lower, will lead to the decline of network performance. In addition, using MLP can align diverse feature spaces towards a shared encoding direction, ensuring effectiveness only under moderate constraints or supervision by the object's features.

4.5.4. Necessity of residual blocks in TDF

In the TDF module, we investigate the rationale behind incorporating residual structures along with temporal modal difference techniques for enhanced learning outcomes. Through a series of ablation experiments analyzing the feature fine-tuning structure depicted in Eq. (2) based on PTMNet, some critical observations are made.

The results, displayed in the first row of Table 7, reveal that without utilizing W for feature fine-tuning, the performance of PTMNet is suboptimal. This deficiency primarily stems from the lack of interaction among the three feature maps during fusion. The second row shows that even with incorporating W for feature fine-tuning, the network's performance shows only a small enhancement. This minor improvement suggests that the learning process of W is unsatisfactory. The results presented in the third row are from our experimentation with the residual structure outlined in Eq. (2) applied to four sets of middle features (f_t^{rgb}, f_{t-1}^{rgb}, $f_t^{thermal}$ and f_t^{dif}), and they indicate that the outcomes are still suboptimal. The final row of Table 7 shows the design implemented by PTMNet, which incorporates a residual structure for features f_{t-1}^{rgb}, $f_t^{thermal}$ and f_t^{dif}, without adjusting f_t^{rgb}. This approach prevents significant deviations among features from different modalities during deep fusion, thereby optimizing results.

4.5.5. PIF's contribution to middle fusion

To further investigate the improvement in coherence between RGB and thermal features using PIF, we conduct experiments with a single-frame detector EINet. EINet only incorporates a single feature-level fusion strategy, so we incorporate the PIF at the network's entrance, allowing each of its six branches to receive complementary information from other frames. The results of the experiments are displayed in Table 8. The experiments demonstrate that this strategy of PIF can effectively enhance the middle fusion and unlock the greater potential of the fusion network.

Table 8
The experimental results demonstrate the impact of incorporating our proposed PIF module into the EINet that initially only employed a middle fusion strategy.

Methods	Params (M)	Gflops	AP (%)	AP50 (%)	FPS
EINet (single-frame)	23.65	**60.18**	22.41	43.97	**142.2**
EINet (single-frame)+PIF	**23.65**	61.73	**22.70**	**45.31**	131.9

Table 9
Experiments were conducted to investigate the construction of the tensor f_t^{dif}. Group "♦" signifies the utilization of PTMNet.

f_t^{dif}	AP (%)	AP50 (%)
♦ f_t^{rgb}-f_{t-1}^{rgb}-$f_t^{thermal}$	**27.07**	**51.4**
–	24	46.77 (↓4.63)
f_t^{rgb}-f_{t-1}^{rgb}	24.36	48.42 (↓2.98)
f_t^{rgb}-$f_t^{thermal}$	25.16	49.77 (↓1.63)
f_{t-1}^{rgb}-$f_t^{thermal}$	24.98	49.30 (↓2.10)
f_t^{rgb}*f_{t-1}^{rgb}*$f_t^{thermal}$	24.26	47.66 (↓3.74)
f_t^{rgb}+ f_{t-1}^{rgb}−$f_t^{thermal}$	24.72	47.66 (↓3.74)
f_t^{rgb}−f_{t-1}^{rgb}+ $f_t^{thermal}$	24.63	47.31 (↓4.09)

Table 10
The table shows the experimental results of extending the PIF and the TDF to another task with a different framework.

Methods	F_beta(val) ↑	MAE ↓
MobileSal	0.8944	0.055
MobileSal+PIF+TDF	**0.8984**	**0.048**

4.5.6. Temporal-modal difference tensor in TDF

The construction of f_t^{dif} plays a crucial role in the TDF as it determines the direction of the fusion module. To investigate this, we conduct a series of experiments, as presented in Table 9. The utilization of the temporal-modal difference tensor as a guidance for the feature encoder results in the best experimental outcome. The other groups, however, exhibit varying degrees of decline. The use of solely the temporal difference in the second line results in the smallest drop, proving its effectiveness in modeling temporal information.

4.5.7. Scalability of the proposed PIF and TDF

In order to investigate the generalizability of the fusion strategy employed by PTMNet, we extend it to another model that is MobileSal [66], which is an algorithm for handling the RGB-Depth salient object detection task. The experimental results are shown in Table 10. We find that the PIF still aids the network in better feature fusion at the middle stage, and results in an overall performance improvement for MobileSal in conjunction with the TDF.

4.6. Feature visualization

To justify the fusion of combining the two fusion strategies, we visualize and select representative feature maps generated during the training stage, as depicted in Fig. 5.

Fig. 5 illustrates the RGB image, the thermal image and the feature map of the middle stage in the network. Red rectangles highlight the areas for comparison. We can see that the baseline neglects the foreground object and erroneously enhances responses in background noise areas, showing in the highlighted ground regions of both images. Introducing PIF enhances attention towards the object area due to complementary thermal data. Consequently, the network focuses more on the foreground, accurately capturing some discriminative features. Subsequent addition of TDF facilitates semantic fine-grained fusion of modalities, enabling comprehensive learning of effective object information, including the obscured and the partially occluded objects.

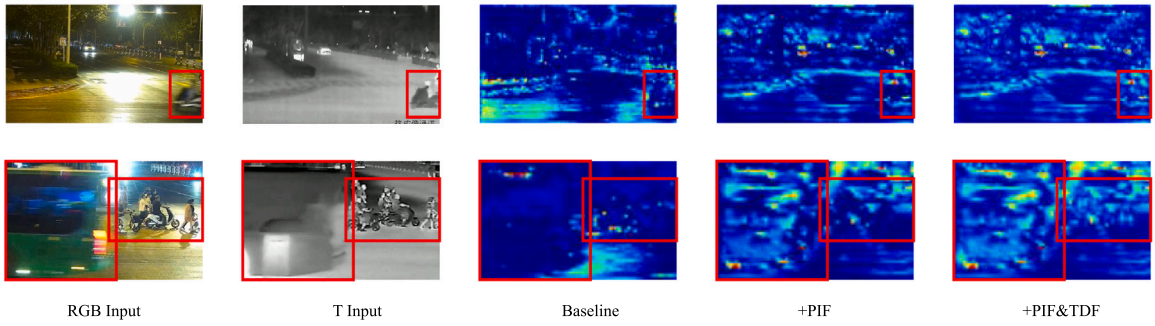

Fig. 5. We visualize the feature maps of the two sets of example inputs, with PIF and TDF incrementally added to the baseline for comparison.

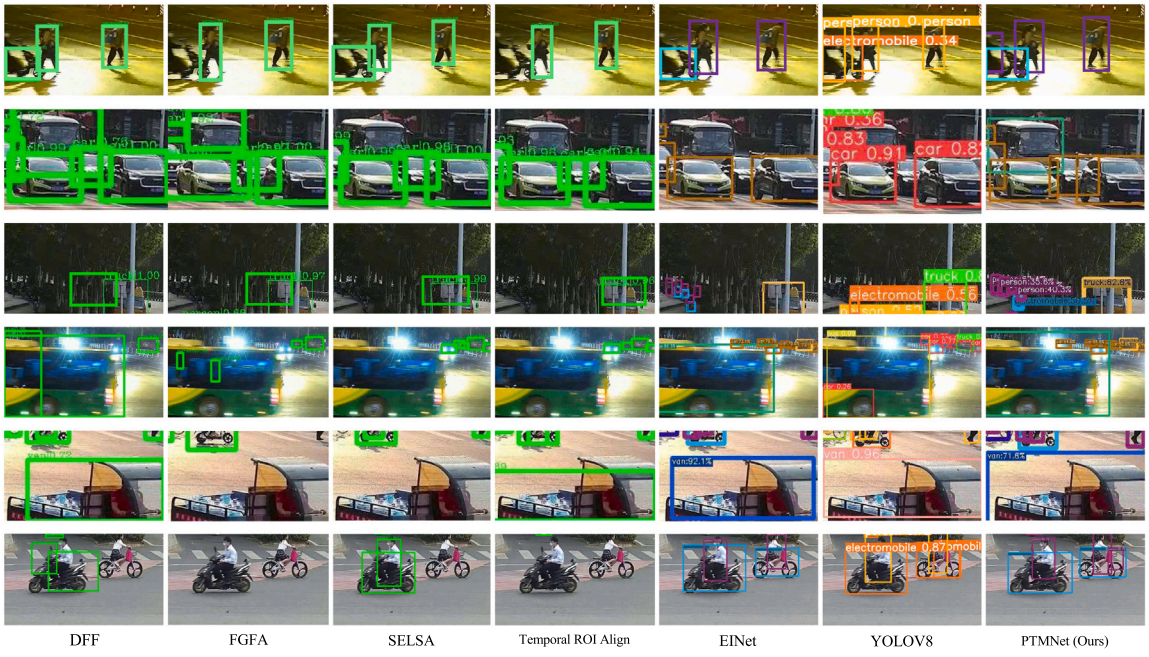

Fig. 6. To visualize the prediction results, various classic models were applied to representative scenes in VT-VOD50. These scenes sequentially include scenarios such as low light, object overlap, shadows, motion blur, moving off edges, and high resolution.

4.7. Results visualization

To visually assess the performance of different methods, we choose specific scenes and employ diverse network models to visualize and present their detection results in Fig. 6. PTMNet demonstrates superior robustness and effectiveness compared to other networks when encountering low light, shadows, or instances of object overlapping. In scenarios with motion-blurred objects, PTMNet tends to predict the object's bounding box conservatively, typically yielding a slightly larger box than the ground truth. Even the object will depart from the scene, PTMNet can locate and classify the object. In contrast, when detecting objects in high-resolution images, most two-stage detectors struggle with scale variations and often fail to capture the object adequately. Conversely, one-stage detectors, including PTMNet, exhibit better performance in such scenarios.

4.8. Deficiency and discussion

Although PTMNet has achieved efficient online detection results, we find that sometimes it fails when facing multiple challenges in the inference process. Fig. 7 depicts a challenging scene including low light, bright light sources and motion blur. Currently, while PTMNet successfully detects the blurred electromobile, it fails to recognize more obscured objects. However, we can detect the object occasionally in a long time. Hence, optimizing efficiency while using temporal information effectively remains a significant challenge for RGBT VOD.

However, after PTMNet adapting to the scenes in VT-VOD50, its performance will not meet detection expectations when facing other more different scenes. Therefore, in the future, we plan to construct a more extensive dataset and benchmark for RGBT VOD.

5. Conclusion

This paper proposes a novel RGBT video object detector called PTMNet, which achieves high-performance real-time online detection by utilizing multiple modal data. PTMNet employs the early fusion strategy of progressive interaction to fuse the complementary information from different modalities. This approach introduces the modal differences in order to reduce them during feature extraction in the backbone network. Consequently, it enhances the correlation of inter-modal features, facilitating feature-level fusion. Additionally, PTMNet integrates a feature-level fusion architecture utilizing sparse temporal-modal differences. This approach not only capitalizes on spatial offsets for learning temporal information, but it also employs feature erasure between RGB and thermal to direct the network's attention to shared

Fig. 7. The figure shows a case of failure for PTMNet when encountering the dual challenges of low light and motion blur.

objects, thus enhancing performance. Moreover, the input and the one-stage detection architecture designed for PTMNet significantly ensure the network's inference efficiency.

CRediT authorship contribution statement

Qishun Wang: Writing – original draft. **Zhengzheng Tu:** Supervision, Formal analysis, Data curation. **Chenglong Li:** Visualization, Validation. **Jin Tang:** Resources, Project administration.

Declaration of competing interest

All authors disclosed no relevant relationships.

Data availability

Data will be made available on request.

Acknowledgments

This work is supported by the University Synergy Innovation Program of Anhui Province, China under Grant GXXT-2022-014, National Natural Science Foundation of China under Grant no. 62376005, 62376004; Natural Science Foundation of Anhui Province, China Grant no. 2208085J18

References

[1] Zhengzheng Tu, Qishun Wang, Hongshun Wang, Kunpeng Wang, Chenglong Li, Erasure-based interaction network for RGBT video object detection and a unified benchmark, 2023, arXiv preprint arXiv:2308.01630.

[2] Di Feng, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudius Glaeser, Fabian Timm, Werner Wiesbeck, Klaus Dietmayer, Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges, IEEE Trans. Intell. Transp. Syst. 22 (3) (2020) 1341–1360.

[3] Martin Brenner, Napoleon H. Reyes, Teo Susnjak, Andre L.C. Barczak, Rgb-d and thermal sensor fusion: A systematic literature review, IEEE Access (2023).

[4] Yabin Zhu, Chenglong Li, Bin Luo, Jin Tang, Xiao Wang, Dense feature aggregation and pruning for RGBT tracking, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 465–472.

[5] Andong Lu, Cun Qian, Chenglong Li, Jin Tang, Liang Wang, Duality-gated mutual condition network for RGBT tracking, IEEE Trans. Neural Netw. Learn. Syst. (2022).

[6] Zijian Zhao, Jie Zhang, Shiguang Shan, Noise robust hard example mining for human detection with efficient depth-thermal fusion, in: International Conference on Automatic Face and Gesture Recognition, 2020.

[7] Ahmet Ozcan, Omer Cetin, A novel fusion method with thermal and RGB-D sensor data for human detection, IEEE Access 10 (2022) 66831–66843.

[8] Jonathon Shlens, A tutorial on principal component analysis, 2014, arXiv preprint arXiv:1404.1100.

[9] Pengyu Zhang, Jie Zhao, Dong Wang, Huchuan Lu, Xiang Ruan, Visible-thermal UAV tracking: A large-scale benchmark and new baseline, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 8886–8895.

[10] Haiping Wu, Yuntao Chen, Naiyan Wang, Zhaoxiang Zhang, Sequence level semantics aggregation for video object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9217–9225.

[11] Tao Gong, Kai Chen, Xinjiang Wang, Qi Chu, Feng Zhu, Dahua Lin, Nenghai Yu, Huamin Feng, Temporal ROI align for video object recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, No. 2, 2021, pp. 1442–1450.

[12] Khurram Azeem Hashmi, Alain Pagani, Didier Stricker, Muhammad Zeshan Afzal, BoxMask: Revisiting bounding box supervision for video object detection, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 2030–2040.

[13] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, Adv. Neural Inf. Process. Syst. 28 (2015).

[14] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, Alexander C Berg, 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, 2016, pp. 21–37.

[15] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, Piotr Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.

[16] Yihong Chen, Yue Cao, Han Hu, Liwei Wang, Memory enhanced global-local aggregation for video object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10337–10346.

[17] Guanxiong Sun, Yang Hua, Guosheng Hu, Neil Robertson, Mamba: Multi-level aggregation via memory bank for video object detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, No. 3, 2021, pp. 2620–2627.

[18] Liang Han, Zhaozheng Yin, Global memory and local continuity for video object detection, IEEE Trans. Multimed. (2022).

[19] Yuheng Shi, Naiyan Wang, Xiaojie Guo, Yolov: Making still image object detectors great at video object detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, No. 2, 2023, pp. 2254–2262.

[20] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, Jian Sun, Yolox: Exceeding yolo series in 2021, 2021, arXiv preprint arXiv:2107.08430.

[21] Shih-Chia Huang, Bo-Hao Chen, Highly accurate moving object detection in variable bit rate video-based traffic monitoring systems, IEEE Trans. Neural Netw. Learn. Syst. 24 (12) (2013) 1920–1931.

[22] Zhengzheng Tu, Zhun Li, Chenglong Li, Yang Lang, Jin Tang, Multi-interactive dual-decoder for RGB-thermal salient object detection, IEEE Trans. Image Process. 30 (2021) 5678–5691.

[23] Zhengzheng Tu, Zhun Li, Chenglong Li, Jin Tang, Weakly alignment-free RGBT salient object detection with deep correlation network, IEEE Trans. Image Process. 31 (2022) 3752–3764.

[24] Yi-Ting Chen, Jinghao Shi, Zelin Ye, Christoph Mertz, Deva Ramanan, Shu Kong, Multimodal object detection via probabilistic ensembling, in: European Conference on Computer Vision, Springer, 2022, pp. 139–158.

[25] Liming Huang, Kechen Song, Jie Wang, Menghui Niu, Yunhui Yan, Multi-graph fusion and learning for RGBT image saliency detection, IEEE Trans. Circuits Syst. Video Technol. 32 (3) (2022) 1366–1377.

[26] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D. Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al., End to end learning for self-driving cars, 2016, arXiv preprint arXiv:1604.07316.

[27] Wei Li, C.W. Pan, Rong Zhang, J.P. Ren, Y.X. Ma, Jin Fang, F.L. Yan, Q.C. Geng, X.Y. Huang, H.J. Gong, et al., AADS: Augmented autonomous driving simulation using data-driven algorithms, Sci. Robotics 4 (28) (2019) eaaw0863.

[28] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, Yichen Wei, Flow-guided feature aggregation for video object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 408–417.

[29] Shiyao Wang, Yucong Zhou, Junjie Yan, Zhidong Deng, Fully motion-aware network for video object detection, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 542–557.

[30] Zhengkai Jiang, Yu Liu, Ceyuan Yang, Jihao Liu, Peng Gao, Qian Zhang, Shiming Xiang, Chunhong Pan, Learning where to focus for efficient video object detection, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16, Springer, 2020, pp. 18–34.

[31] Chun-Han Yao, Chen Fang, Xiaohui Shen, Yangyue Wan, Ming-Hsuan Yang, Video object detection via object-level temporal aggregation, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16, Springer, 2020, pp. 160–177.

[32] Yiming Cui, Liqi Yan, Zhiwen Cao, Dongfang Liu, Tf-blender: Temporal feature blender for video object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 8138–8147.

[33] Fei He, Naiyu Gao, Jian Jia, Xin Zhao, Kaiqi Huang, Queryprop: Object query propagation for high-performance video object detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, No. 1, 2022, pp. 834–842.
[34] Qianyu Zhou, Xiangtai Li, Lu He, Yibo Yang, Guangliang Cheng, Yunhai Tong, Lizhuang Ma, Dacheng Tao, TransVOD: End-to-end video object detection with spatial-temporal transformers, IEEE Trans. Pattern Anal. Mach. Intell. (2022) 1–16.
[35] Raja Sunkara, Tie Luo, No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2022, pp. 443–459.
[36] Qiang Chen, Yingming Wang, Tong Yang, Xiangyu Zhang, Jian Cheng, Jian Sun, You only look one-level feature, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13039–13048.
[37] Warren S. McCulloch, Walter Pitts, A logical calculus of the ideas immanent in nervous activity, Bull. Math. Biophys. 5 (1943) 115–133.
[38] Limin Wang, Zhan Tong, Bin Ji, Gangshan Wu, Tdn: Temporal difference networks for efficient action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1895–1904.
[39] Nitish Srivastava, Elman Mansimov, Ruslan Salakhutdinov, Unsupervised learning of video representations using LSTMs, in: International Conference on Machine Learning, 2015, arXiv:1502.04681.
[40] Karen Simonyan, Andrew Zisserman, Two-stream convolutional networks for action recognition in videos, Adv. Neural Inf. Process. Syst. 27 (2014).
[41] Zheng Ge, Songtao Liu, Zeming Li, Osamu Yoshie, Jian Sun, Ota: Optimal transport assignment for object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 303–312.
[42] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, Thomas Huang, Unitbox: An advanced object detection network, in: Proceedings of the 24th ACM International Conference on Multimedia, 2016, pp. 516–520.
[43] Joseph Redmon, Ali Farhadi, Yolov3: An incremental improvement, 2018, arXiv preprint arXiv:1804.02767.
[44] Glenn Jocher, Ultralytics YOLOv5, 2020.
[45] Fang Qingyun, Han Dapeng, Wang Zhaokui, Cross-modality fusion transformer for multispectral object detection, 2021, arXiv preprint arXiv:2111.00273.
[46] Chuyi Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie, et al., YOLOv6: A single-stage object detection framework for industrial applications, 2022, arXiv preprint arXiv:2209.02976.
[47] Chien-Yao Wang, Alexey Bochkovskiy, Hong-Yuan Mark Liao, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 7464–7475.
[48] Chien-Yao Wang, I-Hau Yeh, Hong-Yuan Mark Liao, YOLOv9: Learning what you want to learn using programmable gradient information, 2024, arXiv preprint arXiv:2402.13616.
[49] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, Guiguang Ding, Yolov10: Real-time end-to-end object detection, 2024, arXiv preprint arXiv:2405.14458.
[50] Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R. Scott, Weilin Huang, Tood: Task-aligned one-stage object detection, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV, IEEE Computer Society, 2021, pp. 3490–3499.
[51] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, Jifeng Dai, Deformable DETR: Deformable transformers for end-to-end object detection, in: International Conference on Learning Representations, 2021.
[52] 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, pp. 16965–16974.
[53] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, Heung-Yeung Shum, Dino: Detr with improved denoising anchor boxes for end-to-end object detection, 2022, arXiv preprint arXiv:2203.03605.
[54] Shilong Zhang, Xinjiang Wang, Jiaqi Wang, Jiangmiao Pang, Chengqi Lyu, Wenwei Zhang, Ping Luo, Kai Chen, Dense distinct query for end-to-end object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 7329–7338.
[55] Shoufa Chen, Peize Sun, Yibing Song, Ping Luo, Diffusiondet: Diffusion model for object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 19830–19843.
[56] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, Yichen Wei, Deep feature flow for video recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2349–2358.
[57] Jiajun Deng, Yingwei Pan, Ting Yao, Wengang Zhou, Houqiang Li, Tao Mei, Relation distillation networks for video object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7023–7032.
[58] Zhi Lin, Junhao Lin, Lei Zhu, Huazhu Fu, Jing Qin, Liansheng Wang, A new dataset and a baseline model for breast lesion detection in ultrasound videos, in: Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, Shuo Li (Eds.), Medical Image Computing and Computer Assisted Intervention, MICCAI 2022, Springer Nature Switzerland, Cham, 2022, pp. 614–623.
[59] Chao Qin, Jiale Cao, Huazhu Fu, Rao Muhammad Anwer, Fahad Shahbaz Khan, A spatial-temporal deformable attention based framework for breast lesion detection in videos, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2023, pp. 479–488.
[60] Alexey Bochkovskiy, Chien-Yao Wang, Hong-Yuan Mark Liao, Yolov4: Optimal speed and accuracy of object detection, 2020, arXiv preprint arXiv:2004.10934.
[61] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. 25 (2012).
[62] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[63] Karen Simonyan, Andrew Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.
[64] Kaiheng Weng, Xiangxiang Chu, Xiaoming Xu, Junshi Huang, Xiaoming Wei, Efficientrep: An efficient repvgg-style convnets with hardware-aware neural network design, 2023, arXiv preprint arXiv:2302.00386.
[65] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, Liang-Chieh Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510–4520.
[66] Yu-Huan Wu, Yun Liu, Jun Xu, Jia-Wang Bian, Yu-Chao Gu, Ming-Ming Cheng, MobileSal: Extremely efficient RGB-D salient object detection, IEEE Trans. Pattern Anal. Mach. Intell. 44 (12) (2022) 10261–10269.