# ACENet: Adaptive Context Enhancement Network for RGB-T Video Object Detection *

Zhengzheng Tu[1], Le Gu[1], Danying Lin[1], and Zhicheng Zhao[1]✉

[1] Anhui University, Hefei 230601, China
zhengzhengahu@163.com, {awngule, danying_lin}@foxmail.com,
zhaozhicheng@ahu.edu.cn

**Abstract.** RGB-thermal (RGB-T) video object detection (VOD) aims to leverage the complementary advantages of visible and thermal infrared sensors to achieve robust performance under various challenging conditions, such as low illumination and extreme illumination changes. However, existing multimodal VOD approaches face two critical challenges: accurate detection of objects at different scales and efficient fusion of temporal information from multimodal data. To address these issues, we propose an Adaptive Context Enhancement Network (ACENet) for RGB-T VOD. Firstly, we design an Adaptive Context Enhancement Module (ACEM) to adaptively enhance multi-scale context information. We introduce ACEM in the FPN section, where it can adaptively extract context information and incorporate it into the high-level feature maps. Secondly, we design a Multimodal Temporal Fusion Module (MTFM) to perform temporal and modal fusion using coordinate attention with atrous convolution at the early stage, significantly reducing the complexity of fusing temporal information from RGB and thermal data. Experimental results on the VT-VOD50 dataset show that our ACENet significantly outperforms other mainstream VOD methods. Our code will be available at: https://github.com/bscs12/ACENet.

**Keywords:** Video object detection · RGB-T · Multimodal fusion · Adaptive context enhancement.

## 1 Introduction

Video object detection has received increasing attention in recent years, aiming to localize and classify all objects in a given video clip. Compared to image object detection [1][3][24], video object detection enhances performance by utilizing temporal information and offers a robust solution to challenges such as motion blur, local occlusion, and rapidly changing appearances [29][22][10][23]. Video object detection has a wide range of applications in surveillance systems, autonomous driving, video editing, and intelligent traffic management [14][40][32].

* This work was supported by the University Synergy Innovation Program of Anhui Province under Grant GXXT-2022-014, in part by the National Natural Science Foundation of China under Grant 62306005 and 62376005.

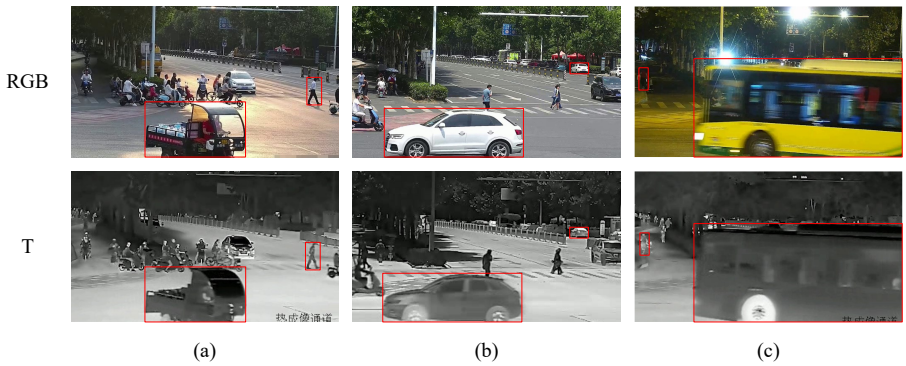


**Fig. 1.** The challenge of coexistence of objects with different scales in the VT-VOD50 [25] dataset. In the visible video, most methods encounter detection difficulties when both large-scale and small-scale objects are present, as illustrated in (a) and (b). In night scenes, the detection difficulty is further exacerbated by weak illumination or local strong light interference, as shown in (c).

However, existing video object detection methods struggle in complex scenes, particularly under conditions such as nighttime and adverse weather, where performance notably decreases. To address this, Tu et al. [25] introduce the RGB-T video object detection task, which enhances the capabilities of the traditional RGB modality by incorporating a thermal modality. Since the thermal modality is less sensitive to light variations yet sensitive to temperature differences, it outperforms the RGB modality in low-light and high-illumination environments. Nevertheless, challenges arise in bright conditions or when objects are distant, as the thermal modality may fail to detect them. Thus, integrating and optimizing the use of both modalities to leverage their respective strengths remains a critical challenge in RGB-T video object detection.

Some methods have arisen that attempt to fuse the RGB modality with the thermal modality in different tasks. Tu et al. [25] propose an erasure-based interaction network that eliminates the noise of RGB features with the help of thermal image features. Li et al. [27] propose a multi-task stream-ordering method with cross-modal consistency for the RGB-T salient object detection task. It describes reliability by introducing weights for each modality and integrating them into a graph-based stream-ordering algorithm for adaptive fusion of data from different sources. Zhang et al. [36] stitch the feature maps of the RGB modality with the thermal modality and feed them into a fully connected layer for prediction. Deng et al. [6] propose FEANet for the task of RGB-T semantic segmentation, mining and enhancing multilevel features from channel and spatial views through a feature-enhanced attention module. FEANet preserves spatial information and shifts more attention to fuse features with high resolution from RGB-T images. Liu et al. [18] first extract the RGB and the thermal modal features separately using a dual-stream Swin Transformer encoder, and then perform cross-modal

fusion between layers. Gao et al. [7] propose a multistage multiscale fusion network that performs modal fusion at each stage of the encoder. However, few of these methods perform multimodal fusion at an early stage. In our opinion, early fusion not only reduces the amount of computation but also preserves more original detail information. In addition, most methods often fail to achieve accurate detection of objects at various scales, as shown in Fig. 1.

Some previous approaches address specific problems by introducing contextual information and obtain significant results [31][34][4][37]. Xiao et al. [31] utilize contextual information to improve the detection performance of tiny objects. Yuan et al. [34] propose a contextual aggregation scheme for semantic segmentation tasks, which further demonstrates the importance of contextual information for dense prediction tasks. Inspired by them, in order to achieve accurate detection of objects at different scales, we develop a novel Adaptive Context Enhancement Network (ACENet) for RGB-T VOD. Specifically, we design a Multimodal Temporal Fusion Module (MTFM) and an Adaptive Context Enhancement Module (ACEM). The inputs of MTFM are three adjacent frames of RGB and thermal modal images. MTFM utilizes coordinate attention [13] and atrous convolution [33] for temporal and modal fusion. Additionally, we introduce ACEM in the FPN part, which is able to adaptively extract contextual information and supplement it into the high-level feature map.

Overall, the main contributions of this work include:

- To address the challenge of simultaneously and accurately detecting objects with varying scales in a video, we design a novel Adaptive Context Enhancement Module (ACEM), which employs the dynamic convolution to adaptively extract context information, which is then integrated into the high-level feature map.
- To reduce the complexity of the temporal fusion of RGB and thermal imaging data, we design the Multimodal Temporal Fusion Module (MTFM), which employs the coordinate attention and atrous convolution at an early stage for both temporal and modal fusion. This approach not only significantly reduces the computational cost but also preserves more original details.
- Experimental results on VT-VOD50 [25] dataset indicate that our model significantly outperforms other video object detection models.

## 2 Related Works

### 2.1 Video Object Detection

Zhu et al. [42] introduce DFF, a rapid and precise framework designed for video recognition. This method conducts convolutional operations exclusively on sparse keyframes and extends its deep feature maps to other frames via a flow field propagation. DFF achieves significant speedups by means of this fast streaming computation. FGFA [41] improves per-frame features by aggregating nearby features along the motion path, enhancing the accuracy of video recognition. It utilizes the principle of temporal consistency at the feature level.

Although DFF and FGFA employ optical flow to achieve improvements in speed and accuracy , their detection effectiveness depends on the effectiveness of the optical flow. MEGA [5] mimics human recognition by utilizing both global semantic information and local localization information to recognize objects in videos. MEGA enables keyframes to access more content but requires maintaining a large number of cached features. Gong et al. [9] note that frames in a video contain the same object instances with highly similar features and propose a novel temporal ROI alignment operator. This operator utilizes feature similarity to extract the features suggested by the current frame from the feature maps of other frames. Although good results are achieved, a large number of support frames are still required to assist the current frame. TransVOD [12] treats video object detection as an end-to-end sequence decoding/prediction problem, utilizing temporal converters to link the output of each object query with memory encoding simultaneously. However, TransVOD requires more video memory to achieve higher accuracy. QueryProp [11] investigates object-level feature propagation. A good accuracy-speed tradeoff is achieved by performing query propagation from sparse keyframes to dense non-keyframes and from previous keyframes to the current keyframe.

### 2.2 RGB-T Object Detection

Li et al. [27] propose a multi-task stream-ordering method with cross-modal consistency for the RGB-T salient object detection task. Reliability is enhanced by introducing a weight for each modality and integrating them into a graph-based stream-ordering algorithm for adaptive fusion of data from various sources. Wang et al. [28] introduce CGFNet for RGB-T salient object detection. In the decoding process, the RGB and T modalities serve as the primary and auxiliary guides for full cross-modal fusion, respectively. Zhou et al. [39] propose ECFFNet for RGB-T salient object detection. ECFFNet not only performs cross-modal fusion but also bilaterally integrates foreground and background information to thoroughly delineate the salient object boundaries, thereby achieving superior detection results. Liu et al. [18] introduce SwinNet for RGB-D and RGB-T salient object detection. SwinNet extracts multimodal features using the robust feature extraction capabilities of the dual-stream Swin Transformer encoder and optimizes the intra-layer cross-modal features with spatial alignment and channel recalibration modules. The edge-guided decoder facilitates inter-layer cross-modal fusion, guided by edge features. Tu et al. [25] propose EINet, the first RGB-T video object detection method. EINet greatly improves the efficiency while maintaining performance by eliminating the noise of RGB features with the help of thermal image features.

## 3 Methodology

In this section, we detail our ACENet for RGB-T video object detection (VOD). ACENet enhances RGB-T VOD accuracy by utilizing contextual information.

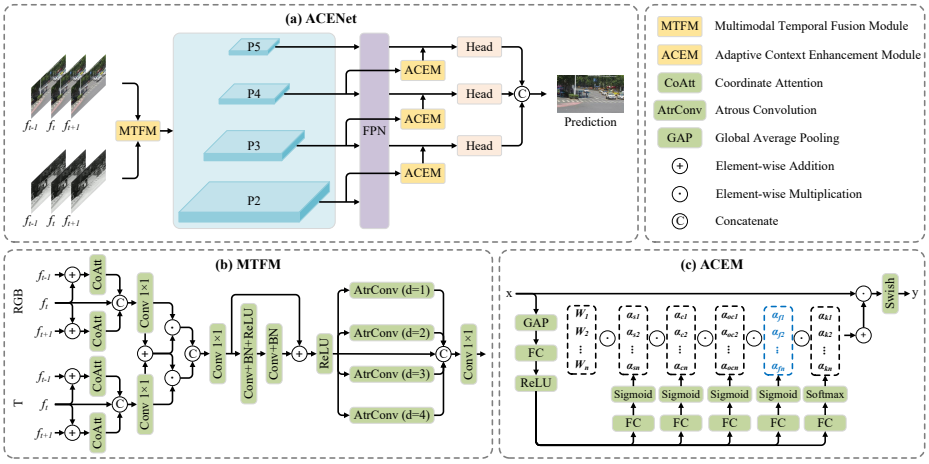


**Fig. 2.** Pipeline of our ACENet. Our ACENet primarily comprises the Multimodal Temporal Fusion Module (MTFM) and the Adaptive Context Enhancement Module (ACEM).

The designed Multimodal Temporal Fusion Module (MTFM) and Adaptive Context Enhancement Module (ACEM) not only extract contextual information from different receptive fields but also adaptively enhance these contextual features. ACENet achieves proficient detection results for objects of varying scales within the video. We provide an overview of ACENet in Section 3.1. The Multimodal Temporal Fusion Module (MTFM) is introduced in Section 3.2, and the Adaptive Context Enhancement Module (ACEM) is presented in Section 3.3.

### 3.1 Overview

We use the YOLOX [8] detector as the baseline for the proposed ACENet. YOLOX employs Darknet53 [20] as its backbone, which boasts powerful feature encoding capabilities. Compared to YOLOv5 [15] and YOLOv7 [26], YOLOX features a more flexible structure. Being a one-stage detector, YOLOX is more suitable for VOD tasks than two-stage detectors such as Faster R-CNN [21].

The general framework of ACENet is illustrated in Fig. 2(a). We select three adjacent video frames each from the RGB and thermal modalities as input. These six frames are then processed by MTFM for temporal and modal fusion. This early fusion approach not only significantly reduces computation and video memory consumption but also preserves more of the original pixel information. The fused features then proceed through the backbone, FPN, and head, which follow the original design of YOLOX, and we will not repeat them here. Additionally, between the FPN and head, we incorporate ACEM for adaptive context enhancement.

### 3.2 Multimodal Temporal Fusion Module

In the baseline method YOLOX, the Focus operation is employed prior to feature extraction by the backbone. Drawing inspiration from Zhang et al. [35], we posit that the Focus operation disrupts the correlation between pixels and the contextual information, adversely affecting the detection of certain objects, particularly smaller ones. Consequently, in ACENet, the Focus operation is omitted.

As illustrated in Fig. 2(b), we designate $f_{t-1}^{RGB}$, $f_t^{RGB}$, and $f_{t+1}^{RGB}$ as the three video frames adjacent to the RGB modality, and $f_{t-1}^{T}$, $f_t^{T}$, and $f_{t+1}^{T}$ as those adjacent to the T modality. For temporal fusion, $f_t^i (i \in \{RGB, T\}$ is initially augmented with $f_{t-1}^i$ and $f_{t+1}^i$. We allow $f_{t-1}^i$ and $f_{t+1}^i$ to sum with $f_t^i$, respectively, proceeding to process these sums to yield $\{f_t^i\}'$ and $\{f_t^i\}''$ via coordinate attention [13]. Then, $\{f_t^i\}'$, $f_t^i$, and $\{f_t^i\}''$ are concatenated and subsequently processed through a $1 \times 1$ convolutional layer to temporally fuse the images of the two modalities. The process is described as follows,

$$\{f_t^i\}' = CoAtt(f_t^i + f_{t-1}^i), i \in \{RGB, T\} \tag{1}$$

$$\{f_t^i\}'' = CoAtt(f_t^i + f_{t+1}^i), i \in \{RGB, T\} \tag{2}$$

$$F^i = Conv_{1\times1}(\{f_t^i\}' \parallel f_t^i \parallel \{f_t^i\}''), i \in \{RGB, T\} \tag{3}$$

where $CoAtt(\cdot)$ denotes coordinate attention. $\parallel$ refers to the concatenation operation along the channel dimension. $Conv_{1\times1}(\cdot)$ represents a convolutional layer with a $1 \times 1$ kernel size.

After obtaining the images of the two modalities after temporal fusion, we perform modal fusion. We first add $F^{RGB}$ and $F^T$ to obtain $\tilde{F}$. Then, we multiply $\tilde{F}$ with $F^{RGB}$ and $F^T$ to obtain $\tilde{F}^{RGB}$ and $\tilde{F}^T$, respectively. Next, $\tilde{F}^{RGB}$ is concatenated with $\tilde{F}^T$, and a $1 \times 1$ convolutional layer is applied to produce the modally fused image $\hat{F}$. Subsequently, two layers of convolutional blocks are used for processing $\hat{F}$, incorporating residual connections. The described procedure can be represented by the following equations,

$$\tilde{F} = F^{RGB} + F^T \tag{4}$$

$$\tilde{F}^i = \tilde{F} \times F^i, i \in \{RGB, T\} \tag{5}$$

$$\hat{F} = Conv_{1\times1}(\tilde{F}^{RGB} \parallel \tilde{F}^T) \tag{6}$$

$$\check{F} = f_{\theta_1}(\hat{F}) \tag{7}$$

$$\bar{F} = ReLU(f_{\theta_2}(\check{F}) + \hat{F}) \tag{8}$$

where $\parallel$ refers to the concatenation operation along the channel dimension. $Conv_{1\times1}(\cdot)$ denotes a convolutional layer with a $1 \times 1$ kernel size. $f_{\theta_1}(\cdot)$ denotes a set of operations, consisting of a convolutional layer with a $3 \times 3$ kernel size, a batch normalization layer and a ReLU activation function. $f_{\theta_2}(\cdot)$ denotes a set of operations, consisting of a convolutional layer with a $3 \times 3$ kernel size and a batch normalization layer. $ReLU(\cdot)$ denotes ReLU activation function.

After obtaining $\bar{F}$, in order to aggregate the contextual information, we apply atrous convolution [33] with different dilations to contextually extract features from $\bar{F}$, resulting in $\{A_i\}_{i=1}^{4}$. We concatenate $A_1$ through $A_4$ and then pass them through a $1 \times 1$ convolutional layer to obtain the final modal feature $A'$. $A'$ represents an early feature map after temporal and modal fusion, augmented with contextual information.

### 3.3 Adaptive Context Enhancement Module

To further highlight the contextual information, we design the Adaptive Context Enhancement Module, as depicted in Fig. 2(c). Inspired by Li et al. [16], we develop our ACEM based on ODConv, which dynamically adjusts the shape and size of the convolution kernel based on the features of the input data during the convolution process. Specifically, ODConv applies attention weighting across four dimensions: convolution kernel size, input channel, output channel, and number of convolution kernels. We denote the input by $X$ and the process of ODConv can be described by the following equations,

$$\alpha_i = g_{\phi_2}(g_{\phi_1}(X)), i \in \{s, c, oc\} \tag{9}$$

$$\alpha_k = g_{\phi_3}(g_{\phi_1}(X)) \tag{10}$$

where $g_{\phi_1}(\cdot)$ denotes a set of operations, consisting of a global average pooling layer, a fully-connected layer and a ReLU activation function. $g_{\phi_2}(\cdot)$ denotes a set of operations, consisting of a fully-connected layer and a sigmoid activation function. $g_{\phi_3}(\cdot)$ denotes a set of operations, consisting of a fully-connected layer and a softmax activation function. $\alpha_s$ denotes the weight of the spatial dimension, $\alpha_c$ denotes the weight of the channel dimension, $\alpha_{oc}$ denotes the weight of the output channel dimension, and $\alpha_k$ denotes the weight of the kernel dimension.

In order to adaptively enhance the context, we introduce the receptive field dimension to ODConv, enabling ODConv to adaptively enhance the context information according to the size of the receptive field. This process can be expressed by the following formula,

$$\alpha_f = g_{\phi_2}(g_{\phi_1}(X)) \tag{11}$$

The entirety of ACEM can be represented by the following equation,

$$Y = \sigma((W \times (\alpha_s + \alpha_c + \alpha_{oc} + \alpha_f + \alpha_k)) \times X) \tag{12}$$

where $X$ denotes the input of ACEM. $W$ stands for convolution kernels. $\sigma$ refers to swish activation function. $Y$ represents the output.

We use $\{P_i\}_{i=1}^{5}$ to denote the feature maps extracted by the backbone. To reduce computation, we only utilize $\{P_i\}_{i=2}^{5}$. ACEM is employed for $\{P_i\}_{i=2}^{4}$ to extract the contextual information $\{C_i\}_{i=2}^{4}$. The output of the FPN is represented by $\{\tilde{P}_i\}_{i=2}^{5}$. We add $\{C_i\}_{i=2}^{4}$ to $\{\tilde{P}_i\}_{i=3}^{5}$ to enhance the contextual features of $\{\tilde{P}_i\}_{i=3}^{5}$. The processed features are denoted by $\{\hat{P}_i\}_{i=3}^{5}$.

## 4 Experiments

In this section, we verify the validity and superiority of our model through extensive experiments. In Section 4.1, we first introduce the dataset we use. In Section 4.2, we describe our experimental setup. In Section 4.3, we compare our method experimentally with other methods. In Section 4.4, we conduct ablation experiments on our designed modules. Finally, in Section 4.5, we present visualizations of our method in comparison with other methods.

### 4.1 Datasets

We utilize the VT-VOD50 [25] dataset to validate the effectiveness of ACENet. The VT-VOD50 dataset comprises 50 pairs of RGB-T video sequences, totaling 9,449 pairs of RGB-T images. These videos are collected under realistic traffic conditions. The VT-VOD50 dataset features seven common road objects: cars, vans, electric vehicles, people, buses, trucks, and bicycles. The scenarios vary from cool to hot weather and cover different times of the day from daytime to dusk to nighttime. The dataset includes various resolutions such as $640 \times 368$, $680 \times 404$, $720 \times 576$, $720 \times 404$, and $1920 \times 1080$. Compared to the widely used ImageNet VID dataset, VT-VOD50 presents greater challenges with its broader array of object classes and a larger number of training samples.

### 4.2 Experiments Settings

We train and test ACENet on the VT-VOD50 dataset and compare it to other leading methods. We deploy the model on two NVIDIA GeForce RTX 3090 GPUs and conduct all experiments using the publicly accessible Pytorch 1.8.0 platform and Python 3.8 environment. During network training, each image is first resized to $640 \times 640$, and then mosaic, mix up, and other data augmentation methods are used to enhance the images. The batch size is set to 8 and the maximum number of epochs is set to 100. We use a SGD optimizer [2] with weight decay 0.0005 and momentum 0.9 to train our network. The initial learning rate is set to 0.05. Warm-up and linear decay strategies are employed for learning rate adjustment.

### 4.3 Quantitative Evaluation

We compare the proposed ACENet with several leading detection methods, including DFF [42], FGFA [41], SELSA [30], YOLOv5 [15], Temporal ROI Align [9], YOLOX [8], TransVOD [12], TransVOD++ [38], STNet [19], CVA-Net [17], and YOLOv7 [26]. For a fair comparison, all these models are retrained on the VT-VOD50 dataset [25]. Table 1 presents the quantitative results of our method alongside eleven other methods on the VT-VOD50 dataset. Our method outperforms the others in both AP50 and AP metrics, achieving scores of 47.7% and 23.9%, respectively. Additionally, the FPS of ACENet is significantly higher than

**Table 1.** Quantitative evaluation results on the VT-VOD50 dataset. *ETD* indicates that extra training data (MS COCO) are used. *MD* indicates that multimodal data are used. The best results with multimodal data are highlighted in red, and the best results without using multimodal data are highlighted in blue.

| Model | Venue | Backbone | ETD | MD | AP50 (%) | AP (%) | FPS |
|---|---|---|---|---|---|---|---|
| **DFF** [42] | CVPR'2017 | ResNet-50 | | | 40.2 | 17.8 | 40.4 |
| | | ResNet-101 | | | 39.5 | 17.6 | 40.9 |
| | | ResNet-X101 | | | 34.4 | 13.7 | 36.5 |
| | | ResNet-50 | | ✓ | 33.5 | 14.1 | 43.3 |
| **FGFA** [41] | ICCV'2017 | ResNet-50 | | | 40.5 | 17.6 | 9.0 |
| | | ResNet-101 | | | 43.6 | 20.1 | 7.5 |
| | | ResNet-X101 | | | 41.0 | 17.9 | 7.0 |
| | | ResNet-50 | | ✓ | 35.1 | 15.8 | 9.2 |
| **SELSA** [30] | ICCV'2019 | ResNet-50 | | | 43.5 | 21.2 | 10.5 |
| | | ResNet-101 | | | 43.9 | 21.2 | 9.8 |
| | | ResNet-X101 | | | 43.1 | 19.6 | 7.8 |
| | | ResNet-50 | | ✓ | 39.4 | 17.4 | 10.6 |
| **YOLOv5** [15] | -'2020 | CSPDarknet53 | | | 36.0 | 20.2 | 140.5 |
| | | CSPDarknet53 | | ✓ | 33.2 | 18.0 | 142.9 |
| **Temporal ROI Align** [9] | AAAI'2021 | ResNet-50 | | | 41.8 | 19.9 | 5.1 |
| | | ResNet-101 | | | 43.0 | 20.8 | 5.0 |
| | | ResNet-X101 | | | 40.4 | 18.3 | 4.4 |
| | | ResNet-50 | | ✓ | 38.0 | 17.0 | 5.2 |
| **YOLOX** [8] | arXiv'2021 | Darknet53 | | | 42.8 | 21.4 | 260.1 |
| | | Darknet53 | | ✓ | 41.2 | 20.6 | 263.9 |
| **TransVOD** [12] | ACM MM'2021 | ResNet-50 | ✓ | | 40.9 | 21.5 | 28.9 |
| | | ResNet-50 | ✓ | ✓ | 36.7 | 20.4 | 23.5 |
| **TransVOD++** [38] | TPAMI'2022 | Swin-B | ✓ | | 46.0 | 25.0 | 8.5 |
| | | Swin-B | ✓ | ✓ | 44.4 | 23.7 | 8.5 |
| **STNet** [19] | MICCAI'2023 | ResNet-50 | | | 40.6 | 19.5 | 5.9 |
| | | ResNet-50 | | ✓ | 38.4 | 18.4 | 5.0 |
| **CVA-Net** [17] | MICCAI'2023 | ResNet-50 | | | 41.9 | 21.3 | 7.2 |
| | | ResNet-50 | | ✓ | 39.7 | 19.7 | 6.9 |
| **YOLOv7** [26] | CVPR'2023 | CSPDarknet53 | | | 40.8 | 20.5 | 416.0 |
| | | CSPDarknet53 | | ✓ | 37.7 | 16.5 | 416.7 |
| **ACENet (Ours)** | - | Darknet53 | | | 43.6 | 21.8 | 149.6 |
| | | Darknet53 | | ✓ | 47.7 | 23.9 | 140.1 |

those of other methods except YOLOX and YOLOv7. This indicates that our MTFM module can dramatically accelerate the inference process. In addition, it can be observed that the performance of other methods decreases instead after using multimodal data. This is because we simply add the RGB video frames with the T video frames, which creates interference between the two modalities. This also proves that our MTFM is able to overcome the modal mutual exclusion problem and fuse RGB and T modalities well. Notably, our method surpasses

**Table 2.** Results of the ablation study on the VT-VOD50 dataset for the proposed modules in ACENet. *MTFM* denotes the Multimodal Temporal Fusion Module, and *ACEM* represents the Adaptive Context Enhancement Module. The optimal results are emphasized in **bold** type.

| MTFM | ACEM | Params (M) | GFLOPs | AP50 (%) | AP (%) | FPS |
|---|---|---|---|---|---|---|
| | | 7.32 | 259.47 | 42.31 | 21.15 | 36.7 |
| ✓ | | **7.01** | **63.42** | 44.49 | 22.24 | **144.2** |
| | ✓ | 7.34 | 261.19 | 45.41 | 22.70 | 34.9 |
| ✓ | ✓ | 7.03 | 65.14 | **47.74** | **23.87** | 140.1 |

TransVOD and TransVOD++, both of which utilize additional training data. This suggests that our method does not require extra data to achieve superior performance. We attribute this advantage to the full utilization of contextual information by the ACEM module.

### 4.4 Ablation Study

We verify the effectiveness of our proposed modules through ablation experiments on the VT-VOD50 dataset. As shown in Table 2, the first line represents the result of the baseline, which is YOLOX with simple modifications to make it suitable for the RGB-T VOD task. Initially, we add the MTFM to the baseline. We observe that AP50 and AP receive a slight boost, while GFLOPs is reduced by 196.05 and FPS is improved by 107.5. This demonstrates that MTFM can markedly decrease the number of parameters and computational demands, and accelerate the inference process. Subsequently, we incorporate only the ACEM module into the baseline, without the MTFM module, and note further improvements in AP50 and AP. This confirms that our ACEM module enhances detection effectiveness. Lastly, we integrate both MTFM and ACEM modules to strike a balance between speed and accuracy, illustrating that the MTFM and ACEM modules complement each other effectively.

### 4.5 Qualitative Evaluation

For a more intuitive comparison, we compare our model with other methods for visualization of detection, and Fig. 3 presents a visual comparison between our ACENet and other baseline models. The first row shows the ground truth, and it is evident that our ACENet detection results are the closest to the ground truth. To further validate the effectiveness of our model, we visualize and analyze a variety of challenging scenarios, including the simultaneous presence of large-scale and small-scale objects (from the second to the fourth column), nighttime scenarios (from the third to the fourth columns), fast-moving objects (from the third to the fourth columns), and bright light interference in the fourth column. It can be observed that our model can still accurately detect all objects in


**Fig. 3.** Visual comparison of our ACENet with other baseline methods. The first and second rows display the original RGB and T modal video frames, respectively. In the RGB video frames, the blue boxes represent the ground truth for the objects. For clarity, we display only a limited number of detection results. It should be noted that the leftmost object in column (a) is solely the red electric bicycle, and the leftmost object in column (b) is solely the person on the electric bicycle. The detection results from different methods are outlined in green boxes, while undetected objects are indicated by red dashed boxes.

these challenging scenarios compared to other baseline methods. Notably, other methods lose detection when confronted with large-scale objects or small objects, especially in night scenes, while our model is able to detect both. The above visualization results also demonstrate that our proposed ACENet can not only address the problem of detecting objects at different scales but can also cope with various challenges.

## 5 Conclusion

In this work, we develop an Adaptive Context Enhancement Network (ACENet) for RGB-thermal (RGB-T) video object detection (VOD). Specifically, we design a Multimodal Temporal Fusion Module (MTFM) and an Adaptive Context Enhancement Module (ACEM). The MTFM takes three neighboring frames of the RGB and thermal modal videos as input, and then performs temporal and modal fusion using coordinate attention and atrous convolution. Subsequently, we introduce the ACEM in the FPN component, which adaptively extracts contextual information and supplements it to the high-level feature maps. Experimental results on the VT-VOD50 dataset demonstrate that our ACENet significantly outperforms other mainstream VOD methods. In future work, we will continue to explore the application of the state-of-the-art YOLO method to the RGB-T VOD task.

## References

1. Beery, S., Wu, G., Rathod, V., Votel, R., Huang, J.: Context r-cnn: Long term temporal context for per-camera object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13075–13085 (2020)
2. Bottou, L.: Stochastic gradient descent tricks. In: Neural Networks: Tricks of the Trade: Second Edition, pp. 421–436. Springer (2012)
3. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6154–6162 (2018)
4. Cao, Y., Xu, J., Lin, S., Wei, F., Hu, H.: Global context networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(6), 6881–6895 (2020)
5. Chen, Y., Cao, Y., Hu, H., Wang, L.: 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 (2020)
6. Deng, F., Feng, H., Liang, M., Wang, H., Yang, Y., Gao, Y., Chen, J., Hu, J., Guo, X., Lam, T.L.: Feanet: Feature-enhanced attention network for rgb-thermal real-time semantic segmentation. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4467–4473. IEEE (2021)
7. Gao, W., Liao, G., Ma, S., Li, G., Liang, Y., Lin, W.: Unified information fusion network for multi-modal rgb-d and rgb-t salient object detection. IEEE Transactions on Circuits and Systems for Video Technology **32**(4), 2091–2106 (2021)
8. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021)
9. Gong, T., Chen, K., Wang, X., Chu, Q., Zhu, F., Lin, D., Yu, N., Feng, H.: Temporal roi align for video object recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 1442–1450 (2021)
10. Han, M., Wang, Y., Chang, X., Qiao, Y.: 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. pp. 431–446. Springer (2020)
11. He, F., Gao, N., Jia, J., Zhao, X., Huang, K.: Queryprop: Object query propagation for high-performance video object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 834–842 (2022)

12. He, L., Zhou, Q., Li, X., Niu, L., Cheng, G., Li, X., Liu, W., Tong, Y., Ma, L., Zhang, L.: End-to-end video object detection with spatial-temporal transformers. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 1507–1516 (2021)
13. Hou, Q., Zhou, D., Feng, J.: Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13713–13722 (2021)
14. Jiao, L., Zhang, R., Liu, F., Yang, S., Hou, B., Li, L., Tang, X.: New generation deep learning for video object detection: A survey. IEEE Transactions on Neural Networks and Learning Systems **33**(8), 3195–3215 (2021)
15. Jocher, G.: ultralytics/yolov5. [Online], https://github.com/ultralytics/yolov5
16. Li, C., Zhou, A., Yao, A.: Omni-dimensional dynamic convolution. arXiv preprint arXiv:2209.07947 (2022)
17. Lin, Z., Lin, J., Zhu, L., Fu, H., Qin, J., Wang, L.: A new dataset and a baseline model for breast lesion detection in ultrasound videos. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 614–623. Springer (2022)
18. Liu, Z., Tan, Y., He, Q., Xiao, Y.: Swinnet: Swin transformer drives edge-aware rgb-d and rgb-t salient object detection. IEEE Transactions on Circuits and Systems for Video Technology **32**(7), 4486–4497 (2021)
19. Qin, C., Cao, J., Fu, H., Anwer, R.M., Khan, F.S.: A spatial-temporal deformable attention based framework for breast lesion detection in videos. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 479–488. Springer (2023)
20. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
21. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence **39**(6), 1137–1149 (2016)
22. Shi, Y., Wang, N., Guo, X.: Yolov: Making still image object detectors great at video object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2254–2262 (2023)
23. Sun, G., Hua, Y., Hu, G., Robertson, N.: 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 (2021)
24. Sun, P., Zhang, R., Jiang, Y., Kong, T., Xu, C., Zhan, W., Tomizuka, M., Li, L., Yuan, Z., Wang, C., et al.: Sparse r-cnn: End-to-end object detection with learnable proposals. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14454–14463 (2021)
25. Tu, Z., Wang, Q., Wang, H., Wang, K., Li, C.: Erasure-based interaction network for rgbt video object detection and a unified benchmark. arXiv preprint arXiv:2308.01630 (2023)
26. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7464–7475 (2023)
27. Wang, G., Li, C., Ma, Y., Zheng, A., Tang, J., Luo, B.: Rgb-t saliency detection benchmark: Dataset, baselines, analysis and a novel approach. In: Image and Graphics Technologies and Applications: 13th Conference on Image and Graphics Technologies and Applications, IGTA 2018, Beijing, China, April 8–10, 2018, Revised Selected Papers 13. pp. 359–369. Springer (2018)

28. Wang, J., Song, K., Bao, Y., Huang, L., Yan, Y.: Cgfnet: Cross-guided fusion network for rgb-t salient object detection. IEEE Transactions on Circuits and Systems for Video Technology **32**(5), 2949–2961 (2021)
29. Wang, S., Zhou, Y., Yan, J., Deng, Z.: Fully motion-aware network for video object detection. In: Proceedings of the European conference on computer vision (ECCV). pp. 542–557 (2018)
30. Wu, H., Chen, Y., Wang, N., Zhang, Z.: Sequence level semantics aggregation for video object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9217–9225 (2019)
31. Xiao, J., Zhao, T., Yao, Y., Yu, Q., Chen, Y.: CONTEXT AUGMENTATION AND FEATURE REFINEMENT NETWORK FOR TINY OBJECT DETECTION (2022), https://openreview.net/forum?id=q2ZaVU6bEsT
32. Xu, C., Si, J., Guan, Z., Zhao, W., Wu, Y., Gao, X.: Reliable conflictive multi-view learning. Proceedings of the AAAI Conference on Artificial Intelligence **38**(14), 16129–16137 (Mar 2024). https://doi.org/10.1609/aaai.v38i14.29546, https://ojs.aaai.org/index.php/AAAI/article/view/29546
33. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
34. Yuan, Y., Huang, L., Guo, J., Zhang, C., Chen, X., Wang, J.: Ocnet: Object context for semantic segmentation. International Journal of Computer Vision **129**(8), 2375–2398 (2021)
35. Zhang, J., Lei, J., Xie, W., Fang, Z., Li, Y., Du, Q.: Superyolo: Super resolution assisted object detection in multimodal remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–15 (2023)
36. Zhang, X., Zhang, X., Du, X., Zhou, X., Yin, J.: Learning multi-domain convolutional network for rgb-t visual tracking. In: 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). pp. 1–6. IEEE (2018)
37. Zhao, S., Gong, M., Fu, H., Tao, D.: Adaptive context-aware multi-modal network for depth completion. IEEE Transactions on Image Processing **30**, 5264–5276 (2021)
38. Zhou, Q., Li, X., He, L., Yang, Y., Cheng, G., Tong, Y., Ma, L., Tao, D.: Transvod: end-to-end video object detection with spatial-temporal transformers. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
39. Zhou, W., Guo, Q., Lei, J., Yu, L., Hwang, J.N.: Ecffnet: Effective and consistent feature fusion network for rgb-t salient object detection. IEEE Transactions on Circuits and Systems for Video Technology **32**(3), 1224–1235 (2021)
40. Zhu, H., Wei, H., Li, B., Yuan, X., Kehtarnavaz, N.: A review of video object detection: Datasets, metrics and methods. Applied Sciences **10**(21), 7834 (2020)
41. Zhu, X., Wang, Y., Dai, J., Yuan, L., Wei, Y.: Flow-guided feature aggregation for video object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 408–417 (2017)
42. Zhu, X., Xiong, Y., Dai, J., Yuan, L., Wei, Y.: Deep feature flow for video recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2349–2358 (2017)