MSANet: Mamba-Based Multi-Scale Attention for RGBT Tracking

yuhan zhao
Anhui Polytechnic University
yue wu
Anhui Polytechnic University
kehan cao
Wuhu NO.1 High School
jixing zhao
Huaneng Chaohu power generation Co.,Ltd
bingyou liu
lby009@mail.ustc.edu.cn

Anhui Polytechnic University
guoyang wan
Anhui Polytechnic University

Research Article

Keywords: RGBT, Dynamic fusion, Multi-Scale Fusion, Mamba

Posted Date: November 14th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-5359152/v1

Additional Declarations: No competing interests reported.

MSANet: Mamba-Based Multi-Scale Attention for RGBT Tracking

yuhan zhao[1], yue wu[1], kehan cao[2], jixing zhao[3], bingyou liu[1*], guoyang wan[1*]

[1]School Of Electrical Engineering, Anhui Polytechnic University, Wuhu, 241000, Anhui, China.

[2] Wuhu NO.1 High School, Wuhu, 241000, Anhui, China.

[3]Huaneng Chaohu power generation Co., Ltd, Chaohu, 238000, Anhui, China.

*Corresponding author(s). E-mail(s): lby009@mail.ustc.edu.cn; 704610266@qq.com;
Contributing authors: 2230342275@stu.ahpu.edu.cn;

Abstract

RGBT (visible and thermal imaging) tracking offers a robust solution for all-weather target tracking by integrating RGB and thermal imaging data. However, traditional fusion methods often struggle in complex scenes with varying conditions. In this paper, we propose a Visual State-Space Module that leverages Mamba's linear complexity long-range modeling capabilities to significantly enhance the robustness of feature extraction. Our method introduces an innovative Multi-Scale Fusion Mechanism that improves the efficiency and accuracy of feature fusion in RGBT tracking. This mechanism captures multi-scale feature information more effectively by generating comprehensive feature maps through the summation of various convolution results, thereby enhancing the model's overall feature representation and discriminative capabilities. We conducted extensive experiments on five publicly available datasets to assess the performance of our method. Experimental results show that our method has certain advantages over existing methods, especially in challenging scenes with background clutter and illumination variations, resulting in more stable and reliable target tracking. It provides a more efficient and robust solution for complex tracking tasks under different environmental conditions.

Keywords: RGBT, Dynamic fusion, Multi-Scale Fusion, Mamba.

1 Introduction

RGBT tracking leverages the complementary strengths of RGB and thermal data to achieve robust visual tracking[1], providing significant potential for continuous, all-weather operation. This capability is essential in various fields, including video surveillance [2], pedestrian tracking [3], and robotics [4]. Recently, significant research has focused on fusing features from RGB and thermal modalities [5], driving advancements in RGBT tracking.

Dynamic and challenging scenes in RGBT tracking limit the effectiveness of traditional fusion structures. Conventional methods often struggle to efficiently process data from both visible and infrared modalities and face difficulties

Fund:Wuhu City Core Technology Research and Development Project (2022hg11, 2023yf012).

Fund: Huaneng Group science and technology project 'HNKJ23-HF20 technical research' funding

with target variations, illumination changes, and background complexity. Therefore, to overcome these challenges and enhance tracking system performance, it is crucial to introduce an emerging strategy: the Visual State Space Module [6] in the feature extraction section. This module not only leverages Mamba's [7] linear complexity in long-range modeling and globally effective sensory field but also enhances feature extraction efficiency and tracking robustness. Consequently, it effectively manages complex and evolving tracking environments, enabling efficient and accurate target tracking.

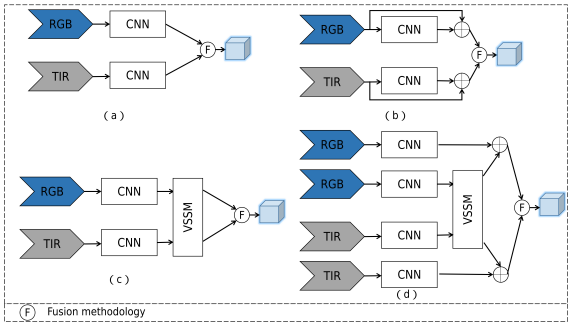


Fig. 1: Comparison of existing RGBT tracking models. (a), (b) and (c) denote the representation model, the residual model and the linear Visual State space Module model.

This paper proposes an innovative multi-scale fusion mechanism. First, we design a multi-scale fusion mechanism, acknowledging that the conventional single-scale convolutional kernel struggles to capture features at varying scales and complexities, limiting its ability to extract comprehensive information. Additionally, feature maps from different modalities may contain unique details, necessitating multi-scale feature extraction to improve feature representation. The module incorporates a multi-scale feature extractor, applying convolution kernels of varying scales to each input feature map. This approach captures feature information at multiple scales, generating multi-scale feature maps by aggregating the results of different convolutions, thereby improving the model's feature representation capability. To further enhance feature processing accuracy and effectiveness, we introduce a finer-grained channel feature processing mechanism that captures complex inter-channel relationships through global average pooling, max pooling, and statistical pooling, integrated with multi-scale convolution operations. Simultaneously, the spatial feature processing mechanism employs average pooling, max pooling, and multi-directional pooling, strengthening the models ability to perceive multi-scale and complex features.Next,we introduce a dynamic fusion network [8], which dynamically adjusts the fusion strategy based on the characteristics of different modalities, effectively overcoming the limitations of traditional fixed-feature fusion modules.

- We introduce a Visual State-Space Module for target tracking that leverages Mamba's long-range modeling capabilities with linear complexity, enhancing both the effectiveness and robustness of feature extraction.
- We propose a novel multi-scale fusion mechanism to address the limitations of fusion structures for robust tracking in dynamically challenging scenarios in RGBT tracking.
- We introduce a dynamic fusion attention network that provides dynamic fusion structures to explore effective fusion schemes for different modalities.

2 RELATED WORK

In this section, we provide a brief overview of the relevant research, focusing on three areas: RGBT tracking, fusion mechanisms, and attention mechanisms.

2.1 RGBT Tracking

Recent advancements in RGBT tracking have led to the development of various innovative algorithms. Existing studies generally fall into two categories. The first focuses on designing distinct modal representations to fully utilize information from each modality. For instance, Peng et al. [9] developed a dynamic fusion-based network that extracts both shared and distinct features while adaptively calculating their contributions. Similarly, Xiao et al. [10] proposed modeling shared and modality-specific representations in challenging scenarios. Feng et al. [11] introduced a sparse hybrid attentional network, based on a hybrid attentional module, which enhances long-range

feature associations and emphasizes original information, notably improving tracking performance in complex conditions like background clutter and camera motion. Although these studies have made notable progress, the lack of carefully crafted fusion strategies hinders the full exploitation of collaborative information between modalities.

The second category is centered on the development of sophisticated fusion modules. For example, Zhang et al. [12] proposed a modality-based cross-weight generator that combines the residuals of generated weights to enhance single-peak feature representation, followed by cascade and convolution operations to produce fused representations. Tang et al. [13] introduced an adaptive fusion method based on decision-level fusion, which extracts and integrates complementary information across modalities. Additionally, Feng et al. [14] proposed a non-local attention-based feature fusion module for more efficient fusion. This method adaptively combines bimodal features by capturing non-local dependencies across channel and spatial dimensions.

However, existing fusion methods are typically fixed and inadequate for managing various complex tracking scenarios simultaneously. In contrast, we propose a dynamic multi-strategy fusion method that meets diverse fusion needs by adjusting the fusion architecture in real time. Furthermore, it proposes a novel multi-scale modal fusion module that integrates multi-scale feature extraction with channel and spatial attention mechanisms, aiming to improve the effectiveness and robustness of feature fusion.

2.2 Fusion Mechanism

In recent years, multi-modal tasks have garnered increasing attention due to their substantial potential in real-world applications. Joint visual target tracking in both RGB and TIR modes has been introduced to provide more robust solutions in practical scenarios. The central challenge in multi-modal tracking lies in the effective fusion of information from these two distinct modalities through a well-designed fusion module. The following sections will present several representative RGBT trackers, classified based on the fusion stage.

Pixel-Level Fusion: Pixel-level fusion is primarily achieved through superimposition, using averaging techniques to achieve satisfactory overlays. However, due to the misalignment between RGB and TIR images, pixel-level fusion is rarely employed in recent trackers. Among the few exceptions, mfDiMP [15] utilizes a simple cascade operator for fusion, although its coarse fusion mechanism assigns equal importance to all positions. In contrast, the MDLatLRR mechanism, based on image fusion, separately merges the high-frequency (detail) and low-frequency (fundamental) components. As previously mentioned, RGB data is more effective in capturing detail, while TIR data offers a more robust target representation via its base component. Consequently, MDLatLRR [16] yields more interpretable and superior results relative to traditional pixel-level fusion methods.

Pixel-level image fusion directly synthesizes fusion information at the pixel level of each image. A major limitation is that the large size of the original image data leads to time-consuming algorithmic implementation, and unprocessed data may cause the strengths and weaknesses of the original sensor information to be superimposed, ultimately compromising fusion outcomes. Additionally, as pixel-level fusion relies on pixel computations, it is highly susceptible to noise and other interferences, resulting in unstable performance.

Feature-Level Fusion: Feature-level fusion is the predominant fusion approach. Based on the function of the fusion blocks, fusion techniques at this level are categorized into two types. The first category includes methods like FANet[17], DAFNet[18], SiamFT[19], and DSiamMFT[20], that output fusion weights. In these approaches, multi-modal features are initially summed or concatenated, then processed through multiple convolutional layers and Softmax operations. As a result, modality-specific weights are learned through this processing, indicating the reliability of each modality, but often neglecting local features. In contrast, the second category eschews explicit fusion models and directly derives fusion features. Notable examples are DAPNet[21], TFNet[22], MANET[23], CAT[24], and mfDiMP[15]. These pioneering deep-learning-based RGBT trackers utilize lightweight subnetworks to minimize redundancy during the fusion of multiple modalities. DAPNet[21], for instance, segregates the modal fusion process and integrates convolutional neural network-based

fusion blocks, applying them at various layers. Building on this idea, TFNet[22] transforms the network into a trident architecture, retaining both the fusion function and the independent RGB and TIR functions, thereby preserving modal features more effectively. Unlike the aforementioned methods, MANET[23] and CAT[24] utilize a multi-branch architecture to acquire fine-grained feature representations that address both channel-sharing challenges (e.g., scale variations) and channel-specific challenges (e.g., illumination variations).

Decision-Level Fusion: In this approach, KL divergence is employed to fuse the response maps of two tracking modules: the correlation filter-based module, which considers both RGB and TIR modalities, and the histogram-based module, which uses only TIR data. Additionally, multi-modal information is integrated based on modal reliability, also computed using KL divergence. The JMMAC [25] section focuses on the fusion sub-network, where a fusion matrix is learned by passing RGB and TIR image patches through the network. Notably, DFAT[26], the winner of the VOT-RGBT2020 challenge, achieves cross-modal fusion at the decision level, addressing biases arising from data discrepancies. Decision-level fusion involves each sensor independently performing classification tasks, after which recognition results from multiple sensors are combined to form a globally optimal decision. This process synthesizes extracted features and recognition results from source images, based on specific rules, to produce a fused image. The input for decision-making originates from the target recognition framework, and the fusion result is obtained through optimal decision processes.

Decision-level fusion offers several advantages, including real-time performance, adaptability, low data requirements, robust anti-interference capabilities, efficient integration of multi-sensor environmental data, and effective error correction. Proper fusion can mitigate errors from individual sensors, ensuring accurate outcomes. However, multi-sensor data also introduces additional risks, as errors from individual sensors can propagate to the decision layer, and the error tolerance of the decision function directly affects the performance of fusion-based classification.

Dynamic fusion: Dynamic networks have become a popular area of research in recent years. Unlike static inference neural network structure search, these networks generate real-time execution paths tailored to the input samples. Currently, their efficiency has led to widespread application across various fields. In the domain of multi-peak analysis, Tsai et al.[27] employed multi-modal routing for multi-modal language analysis, dynamically adjusting the weights between input modes and output representations for each sample. This approach effectively identifies the relative importance of features across multiple modalities. Zeng et al.[28] proposed a channel interaction module for multi-channel sentiment analysis for learning intra and inter-channel interactions. In the field of RGBT, Lu et al.[29] proposed a fusion module based on four different fusion units, the first two of which aim to enhance the discriminative cues within each channel, while the last two are dedicated to extracting cross-channel collaborative information. Despite these advances, the potential of dynamic strategies for RGBT tracking is still largely unexplored.

2.3 Attention Mechanism

Numerous studies have developed attention mechanisms to assess the relative importance of different regions or modalities, addressing various challenges in the process. Attention mechanisms have been widely adopted across multiple applications, enhancing networks' ability to extract robust and distinctive features. In 2014, Google DeepMind introduced the attention mechanism for image classification tasks, presenting a novel recurrent neural network model capable of adaptively selecting and analyzing specific high-resolution regions in images or videos. Hu et al. (2018) [30] introduced the Squeeze-and-Excitation (SE) block, which emphasizes channel interdependencies by learning the relationships between channels and adaptively recalibrating channel-wise feature responses to enhance representational capacity. However, the SE block focuses solely on the channel contributions of feature maps, neglecting the spatial positioning of objects within images, a crucial factor in object detection. Wang et al. (2020) introduced ECA[31], an enhanced version of SENet that eliminates the fully connected layer of the original SENet, replacing it with a 1*1 convolutional kernel, thereby reducing model parameters and making it more lightweight. Li et al.

(2023)[32] further advanced these concepts with additional modifications. They employ spatial and channel attention mechanisms to reduce redundancy in both spatial and channel dimensions. This block is referred to as Spatial and Channel Reconstruction Convolution (SCConv). While the VGG-M-based RGBT tracker shows high accuracy in the unobscured case, it is still not as good as other state-of-the-art trackers when it encounters situations such as occlusion and thermal crossover, partly due to the lack of information for comparing and analyzing the differentiation between the two modalities. Consequently, there is still potential for improving RGBT tracking using VGG-M. In this study, we propose a multi-scale fusion method for RGBT tracking based on attention mechanisms, aimed at minimizing the adverse effects of modality-specific distinguishing features and maximizing complementary cross-channel collaborative information.

3 MSANet

3.1 Overview

In this section, we present the proposed Mamba-Based Multi-Scale Attention Network (MSANet), detailing both the network architecture and the associated learning algorithm. The detailed structure is illustrated in **Fig.2**.

3.2 Vision State-Space Module

To enhance efficiency, transformer-based restoration networks typically partition the input into small chunks[33] or employ moving windows[34], which restrict the ability to interact across the entire image range. Inspired by Mamba's success in modeling long-range dependencies with linear complexity, we introduce a Visual State Space Module (VSSM) for object tracking. This module addresses target loss caused by noise interference during tracking. It converts visual information into a richer, more abstract representation of features. This processing captures the complex structures and relationships within the image, thereby enhancing the model's ability to comprehend the visual data. It enables effective feature extraction and learning from visual data, improving the model?s performance. We integrate the VSSM into the model for feature extraction. To ensure compatibility and optimize performance, we conducted appropriate parameter tuning and architectural optimization of the VSSM. It is demonstrated that the introduction of the Visual State Space Module at each layer significantly enhances the performance and robustness of the system across different datasets, verifying its effectiveness and adaptability in feature extraction and data processing. After introducing the VSSM, experimental results indicate that the module significantly enhances system accuracy. Compared to the baseline model that does not incorporate the VSSM, our enhanced model exhibits higher accuracy and faster convergence across multiple datasets.

2D Selective Scan Module: The Standard Mamba employs causal processing of input data, thereby enabling the capture of information exclusively from the scanned portions [7]. Although this approach is well-suited to sequential tasks in natural language processing, it presents significant challenges when applied to non-causal data, such as images. To utilize two-dimensional spatial information effectively, we adopt an approach from the literature and introduce the Two-Dimensional Selective Scan Module (2D-SSM). In this approach, two-dimensional image features are transformed into a one-dimensional sequence and scanned in four distinct directions: top-left to bottom-right, bottom-right to top-left, top-right to bottom-left, and bottom-left to top-right. The capture of long-range dependencies within each sequence is achieved through the utilisation of discrete state-space equations. Subsequently, the sequences are combined through summation, and the original two-dimensional structure is reconstructed by reshaping the data, as illustrated in **Fig. 3**.

$$X_1 = LN(2D{-}SSM(SiLU(DConv(Linear(X))))) \tag{1}$$

$$X_2 = SiLU(Linear(X)) \tag{2}$$

$$X_{out} = Linear(X_1 \odot X_2) \tag{3}$$

where DWConv represents depth-wise convolution, and $\odot$ denotes the Hadamard product.

3.3 Multi-Scale Fusion Mechanish

We propose a Multi-Scale Fusion Module (MSFM, details of which are shown in **Fig. 4**) that integrates multi-scale feature extraction with channel

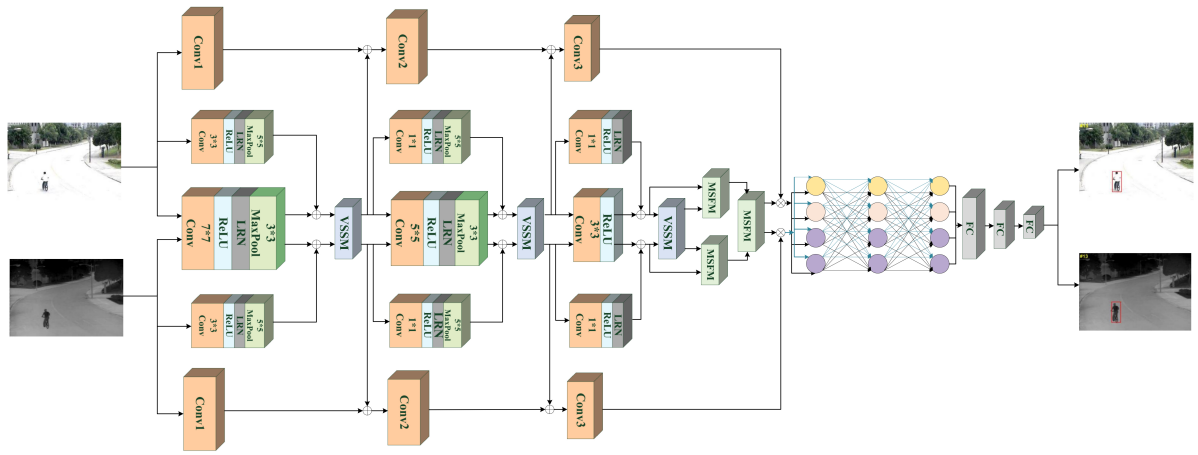


Fig. 2: Overall network architecture of MSANet.In our proposed MSFM module, RGB and TIR are input to a single modality after learning common features between the two modalities through multi-scale convolution to enhance the features of RGB and TIR. Finally, the RGB and TIR features are input to the dynamic fusion module and fed into the prediction head to predict the current state of the target.

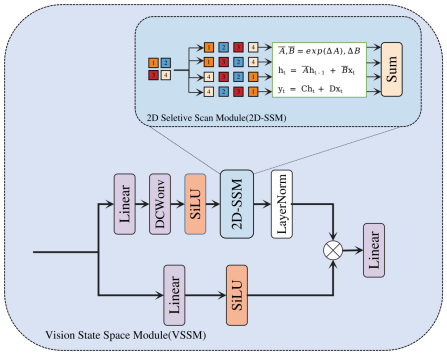


Fig. 3: Vision State-Space Module. 2D-SSM represents 2D Selective Scan Module, DWConv represents depth-wise convolution.

and spatial attention mechanisms. The aim of this module is to enhance the effectiveness and robustness of feature fusion. The module captures features at multiple scales using multi-scale convolution and dynamically adjusts feature importance with an attention mechanism, thereby ensuring that key features receive higher weights during the fusion process. The limitations of traditional single-scale convolutional kernels in capturing features of varying scales and complexities within an image may result in the incomplete extraction of comprehensive information. Moreover, the information present in different modal feature maps may vary in detail, necessitating the use of multi-scale feature extraction to enhance the representation of the features in question. The module includes a multi-scale feature extractor that applies 3x3, 5x5, and 7x7 convolution kernels to each input feature map. The application of multi-scale convolution enables the capture of feature information at varying scales, with the generation of multi-scale feature maps achieved through the summation of results obtained from distinct convolutions. This approach serves to enhance the model's capacity for feature representation. The efficacy of the modules and the positioning of the fusion presence were validated in subsequent ablation experiments, which demonstrated that fusion after conv3 is a more effective method for learning the common features between modalities. Moreover, premature fusion may result in the interference of individual modalities being introduced into the fused features, which could lead to the superposition of interference and, consequently, have an adverse effect on the tracking results.

$$\begin{aligned} f_1 &= ReLU[conv_{3*3}(f^R)] \\ f_2 &= ReLU[conv_{5*5}(f^R)] \\ f_3 &= ReLU[conv_{7*7}(f^R)] \\ f^{R_*} &= f_1 + f_2 + f_3 \end{aligned} \tag{4}$$

The application of simple pooling operations may prove insufficient for the extraction and distinction of complex relationships between channels, particularly when dealing with high-dimensional data. This approach is likely to result in the omission of detailed channel information. We present a more sophisticated channel feature processing mechanism. Subsequently, global average pooling and maximum pooling are performed on the multi-scale feature map. Thereafter, additional statistical information and transformation operations, such as standard deviation pooling and minimum pooling, are integrated to combine these results and further extract features through convolution operations. This approach facilitates the more accurate and effective capture of complex relationships between channels, thereby enhancing the precision and efficacy of feature processing.

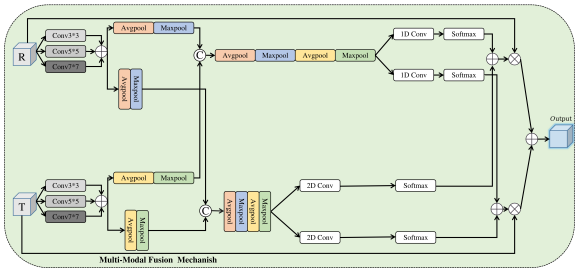


Fig. 4: Multi-Scale Fusion Module.The combination of the two modal spatial weights and channel weights results in the extraction of a robust common feature that is characteristic of both modes.

$$f_{channel}^{R/T} = Cat[Avg_{spatial}(f^{R_*/T_*}), Max_{spatial}(f^{R_*/T_*})] \tag{5}$$

$$f_{channel} = Cat[f_{channel}^{R}, f_{channel}^{T}] \tag{6}$$

$$f_{channel}^{1/2} = Softmax[Conv_{1D}(f_{channel})] \tag{7}$$

A single-scale convolution kernel may inadequately capture the multi-scale information in the feature map, thereby limiting the model's performance in processing features of varying scales and complexities. In the spatial feature processing mechanism, we perform not only average pooling and maximum pooling along spatial dimensions, but also introduce pooling operations in various directions, such as diagonal pooling, to capture spatial information across multiple dimensions. After these pooling results are combined, the spatial features are further extracted using multi-scale convolution operations, thereby enhancing the model's ability to perceive complex scenes and diverse features.

$$f_{spatial}^{R/T} = Cat[Avg_{channel}(f^{R_*/T_*}), Max_{channel}(f^{R_*/T_*})] \tag{8}$$

$$f_{spatial} = Cat[f_{spatial}^{R}, f_{spatial}^{T}] \tag{9}$$

$$f_{spatial}^{1/2} = Softmax[Conv_{2D}(f_{spatial})] \tag{10}$$

Finally, the final output feature is obtained by weighted multiplication with the input map.

$$\begin{aligned} f_{out1/2} &= f_{channel}^{1/2} + f_{spatial}^{1/2} \\ f_{out} &= f_{out1} \otimes f^{R} \oplus f_{out2} \otimes f^{T} \end{aligned} \tag{11}$$

3.4 Dynamic Attention Network

Channel Enhancement Unit. We adopt a method for input modal feature enhancement from the channel perspective. To keep the parameters efficient, we introduce an efficient channel attention architecture to compute the attention weights. Specifically, for a given modal intermediate feature F, we first obtain the aggregated feature $F_g \in R^C$ by a spatial pooling method. Then, we consider only the interactions between each channel $F_i$and its K neighbors in order to compute the weights of the channel F_i. As illustrated in **Fig.5(a)**.

$$\begin{aligned} F_g &= Avgpool(Conv_{1D}(F_t)) \\ F_t^* &= Sigmoid(F_g) \\ F_{CE} &= F_t^* \otimes F_t \end{aligned} \tag{12}$$

Spatial Enhancement Unit. In simple scenes, humans can quickly recognize key objects, suggesting that unimodal information is usually sufficient for visual tasks. Furthermore, cross-modal interactions may lead to feature distortion when there is a severe imbalance in the quality of information across modalities. Therefore, we believe that inter-modal interactions are not always necessary, especially in simple scenes or when the information quality of a particular modality is poor. Based on this observation, we design single-modal spatial enhancement units that utilize the modality's own contextual information to enhance the feature representation of

a specific region in order to improve the discriminative power of the task, as illustrated in **Fig.5(b)**.

$$\begin{aligned} F_t^* &= Avgpool(F_t) \\ F_T &= Sigmoid[Norm(F_t^* \otimes F_t)] \\ F_{SE} &= F_T \otimes F_t \end{aligned} \tag{13}$$

Cross-Modal Enhancement Fusion Unit. The fusion of RGB and thermal modal features has been a key issue in RGBT tracking. Inspired by the excellent performance of the Transformer network in various modal fusion tasks, we employ the efficient mutual attention technique [35] to construct cross-modal enhanced fusion units that perform unidirectional feature fusion from RGB to thermal and from thermal to RGB, respectively. Unlike the self-attention module, where the query (Q), key (K), and value (V) are derived from the same feature, during RGB to thermal unidirectional fusion, the unit uses thermal features as the query and RGB features as the key and value. For thermal to RGB unidirectional fusion, the unit uses RGB features as the query and thermal features as the key and value. The input features of both modalities undergo processing steps, which include a normalization layer and a 1*1 convolutional layer with c output channels to obtain a vectorized mapping. Subsequently, cross-modal attention is applied between the vectorized features of both modalities. Based on this attention mechanism, thermal features can be enhanced by extracting relevant features from the key and value modes and fused with query modes through a simple addition operation. The parameter c is chosen to be much smaller than H - W, where H and W denote the height and width of the input feature map. This design effectively reduces the complexity of the attention operation, as illustrated in **Fig.5(c)**.

$$\begin{aligned} Att_{r2t} &= Softmax(\frac{Q_t ir \otimes K_{rgb}}{\sqrt{d_k}}) \\ F_{r2t} &= F_{tir} + Att_{r2t} \otimes V_{rgb} \end{aligned} \tag{14}$$

Eq.14 is expressed in terms of enhancing the TIR image using the RGB image, and enhancing the RGB image using the TIR image is similar to this equation. Where, $F_{tir}$and F_{r2t} represent the original thermal modal signature and the enhanced fused thermal modal signature, respectively.

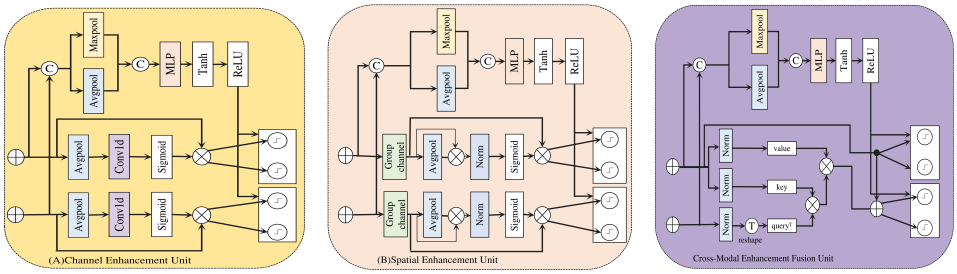
3.5 Discussion

We propose an innovative multi-scale fusion mechanism. The proposed mechanism dynamically adjusts the fusion strategy based on various modal features, effectively overcoming the limitations of fixed feature fusion modules. Subsequently, we design a multi-scale fusion mechanism. The traditional single-scale convolutional kernel is limited in its ability to capture features at various scales and complexities in images, rendering it inadequate for extracting comprehensive information. Through multi-scale convolution, features at different scales are captured, resulting in multi-scale feature maps that enhance feature representation. By leveraging global average pooling, maximum pooling, and statistical information pooling, complex relationships between channels are efficiently captured and combined with multi-scale convolution operations. The accuracy and effectiveness of feature processing are enhanced through adaptive multi-scale fusion mechanisms, thereby surpassing the constraints of fixed feature fusion paradigms. The PR/SR scores on the RGBT234 and LasHeR datasets reach 85.9%/61.6% and 57.2%/43.9%, respectively.

4 EXPERIMENTS

4.1 Implementation Details

We use VGG-M as the base tracker, utilizing three convolutional blocks from VGG-M as the feature extractor. The feature extractor's parameters were initialized using the pre-trained model provided by VGG-M, while the remaining network parameters were randomly initialized. We set the initial learning rate to 0.0001, decay rate to 0.0005, and momentum to 0.9. The entire model is trained using stochastic gradient descent (SGD) to minimize classification and regression loss functions. We trained the complete tracking network end-to-end using the LasHeR training set, evaluating it on the GTOT [36], RGBT210[37], RGBT234[38], and LasHeR test sets[39]. For evaluation on VTUAV[40], the VTUAV training set was used as the training data. MSANet was implemented on the PyTorch platform, running on a



(a) (b) (c)

Fig. 5: (a)Channel Enhancement Unit, (b)Spatial Enhancement Unit, (c)Cross-Modal Enhancement Unit. Ⓒ indicates the concatenation operation across the feature channel dimension.

single Nvidia RTX4070Ti GPU with 12GB of RAM.

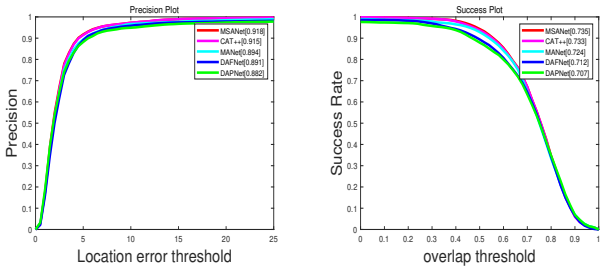
4.2 Quantitative Comparison

We evaluate our algorithm on five widely used RGBT tracking benchmarks, comparing its performance with current state-of-the-art trackers. **Table 1** presents the effectiveness of our proposed method and provides a summary of the comparative results.

Evaluation on GTOT dataset. Table 1 illustrates the GTOT dataset comparison results. In the GTOT dataset, our method surpasses the state-of-the-art trackers by 0.3%/0.2%, 2.6%/0.2%, and 3.7%/1.6% compared to CAT++, CMD, and DFNet, respectively, in terms of PR/SR. Additionally, we compare our method with CMPP, the current top performer on the GTOT dataset. Our PR is 0.8% lower than that of CMPP; this lower PR can be attributed to the prevalence of small objects in the GTOT dataset. The feature pyramid strategy and the pre-frame information base of CMPP enhance feature representation and current frame depiction, respectively.

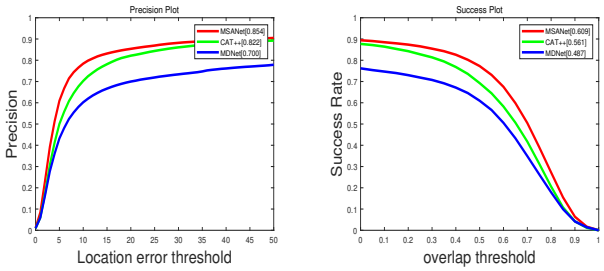
Evaluation on RGBT210 dataset. As shown in **Fig.7**. The PR/SR of MSANet on the RGBT210 dataset is 85.4%/60.9%, respectively. Compared with mfDiMP, the winner of VOT2019-RGBT, the PR/SR of MSANet is significantly improved by 6.8%/5.4%. In addition, compared with DMCNet, our method exhibits a 5.7%/5.4% performance advantage in terms of PR/SR.

Evaluation on RGBT234 dataset. As illustrated in **Fig. 8**, in the target tracking task on the RGBT234 dataset, MSANet demonstrates



(a) PR (b) SR

Fig. 6: Precision Rate (PR) and Success Rate (SR) for the GTOT dataset.



(a) PR (b) SR

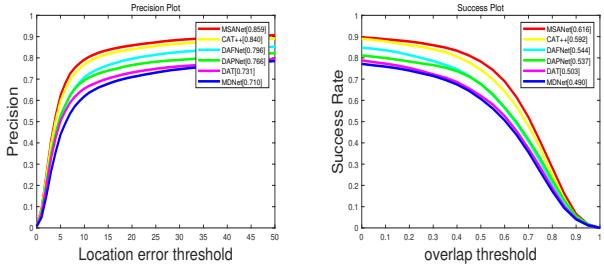
Fig. 7: Precision Rate (PR) and Success Rate (SR) for the RGBT210 dataset.

exceptional performance, achieving an accuracy rate of 85.9% and a success rate of 61.6%, surpassing other methods. In contrast, CAT++ follows closely with an accuracy rate of 84.0% and a success rate of 59.2%. Other methods, such as DFNet and CMD, exhibit accuracy rates of 77.2% and 82.4% and success rates of 51.3% and 58.4%, respectively, further highlighting the

Table 1: THE PR(↑), NPR(↑), AND SR(↑) SCORES (%) OF VARIOUS TRACKERS ON FIVE DATASETS . THE BEST AND SECOND RESULTS ARE IN red AND blue COLORS ,RESPECTIVELY

		GTOT		RGBT210		RGBT234		LasHeR			VTUAV	
MethodS	Backbone	PR	SR	PR	SR	PR	SR	PR	NPR	SR	PR	SR
MANet[23]	VGG-M	89.4	72.4	-	-	77.7	53.9	45.5	38.3	32.6	-	-
DAPNet[21]	VGG-M	88.2	70.7	-	-	76.6	53.7	43.1	38.3	31.4	-	-
mfDiMP[15]	ResNet-50	83.6	69.7	78.6	55.5	-	-	44.7	39.5	34.3	67.3	55.4
CMPP[41]	VGG-M	92.6	73.8	-	-	82.3	57.5	-	-	-	-	-
MaCNet[42]	VGG-M	88.0	71.4	-	-	79	55.4	48.2	42	35	-	-
CAT[24]	VGG-M	88.9	71.7	79.2	53.3	80.4	56.1	45	39.5	31.4	-	-
FANet[17]	VGG-M	89.1	72.8	-	-	78.7	55.3	44.1	38.4	30.9	-	-
ADRNet[43]	VGG-M	90.4	73.9	-	-	80.7	57.0	-	-	-	62.2	46.6
JMMAC[25]	VGG-M	90.2	73.2	-	-	79	57.3	-	-	-	-	-
MANet++[44]	VGG-M	88.2	70.7	-	-	80	55.4	46.7	40.4	31.4	-	-
APFNet[45]	VGG-M	90.5	73.7	-	-	82.7	57.9	50	43.9	36.2	-	-
DMCNet[46]	VGG-M	90.9	73.7	79.7	55.5	83.9	59.3	49	43.1	35.5	-	-
FTNet[47]	VGG-M	91.2	73.6	-	-	83.7	60.1	52.6	-	38.1	-	-
MIRNet[48]	VGG-M	90.9	74.4	-	-	81.6	58.9	-	-	-	-	-
HMFT [49]	ResNet-50	91.2	74.9	78.6	53.7	78.8	56.8	-	-	-	75.8	62.7
MFG [50]	ResNet-18	88.9	70.7	74.9	46.7	75.8	51.5	-	-	-	-	-
DFNet [51]	VGG-M	88.1	71.9	-	-	77.2	51.3	-	-	-	-	-
DRGCNet[52]	VGG-M	90.5	73.5	-	-	82.5	58.1	48.3	42.3	33.8	-	-
CMD[53]	ResNet-50	89.2	73.4	-	-	82.4	58.4	59	54.6	46.4	-	-
CAT++ [54]	VGG-M	91.5	73.3	-	-	84.0	59.2	50.9	44.4	35.6	-	-
Our	VGG-M	91.8	73.5	85.4	60.9	85.9	61.6	57.2	52.6	43.9	82.5	67.9

advantages of MSANet. The reason MSANet stands out among various methods is primarily attributed to its advanced multi-scale modal feature fusion technique, which obtains the common features of the two modalities at the end of the backbone; the acquired raw modal information is subsequently fed into the dynamic modal fusion technique, where the weights of the common features of the two modalities, acquired by MSFM, are associated with the modal differences while retaining the individual features of each modality. This capability enables MSANet to maintain exceptionally high positioning accuracy and tracking stability in complex scenes. Nevertheless, MSANet has potential for further improvement. Future improvements should focus on optimizing the underlying network structure and strengthening multi-scale feature fusion to enhance the model's accuracy and robustness,

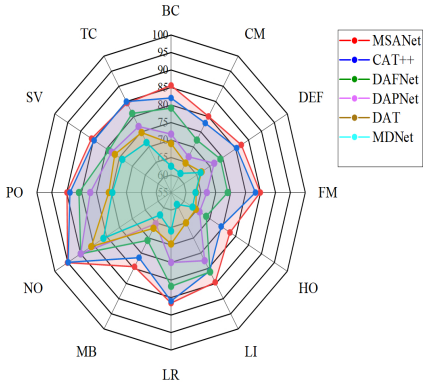


(a) PR (b) SR

Fig. 8: Precision Rate (PR) and Success Rate (SR) for the RGBT234 dataset.

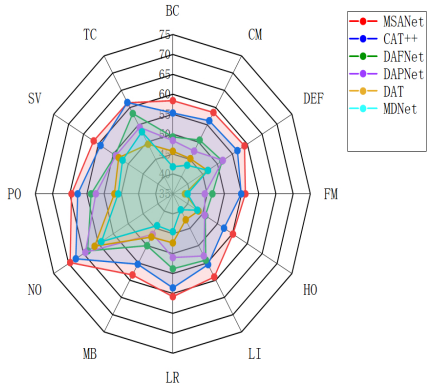
thereby increasing MSANet's competitiveness in practical applications.

As shown in **Fig.9** and **Fig.10**. MSANet excels in most performance metrics, especially in target coverage, centroid metrics and feature fusion,



(a) Attribute-based PR

Fig. 9: Precision Rate (PR) for Different Challenge Attributes on the RGBT234 Dataset.

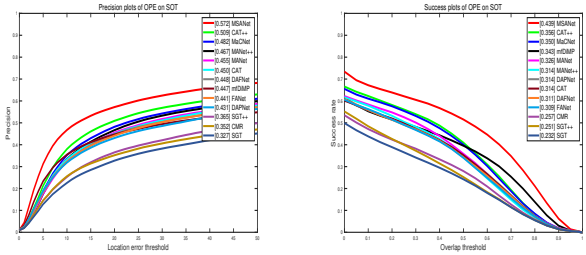


(a) Attribute-based SR

Fig. 10: Success Rate (SR) for Different Challenge Attributes on the RGBT234 Dataset.

demonstrating excellent accuracy and robustness for target tracking needs in a variety of complex scenarios.

Evaluation on LasHeR dataset. As illustrated in **Fig. 11**, our algorithm achieves a precision rate (PR) of 57.2% and a success rate (SR) of 43.9% on the LasHeR dataset, which is outstanding compared to other algorithms. As shown in Table 2, the performance of each algorithm under various challenge attributes (e.g., lighting changes, scale variations, target occlusions, etc.) is presented. The MSANet algorithm demonstrates superior performance on AIV, achieving scores of 34.9%/35.5%/34.9%, which is optimal among all algorithms. This suggests that our algorithm exhibits the highest consistency in tracking when



(a) PR (b) SR

Fig. 11: Precision rate (PR), success rate (SR) of LasHeR dataset.

addressing changes in target shape. The performance in deformation processing is also optimal, achieving scores of 52.7%/50.8%/44.3%, indicating that our algorithm maintains high tracking accuracy despite significant changes in target shape. In terms of scale variations, our algorithm again leads with scores of 47.5%/43.1%/37.5%, demonstrating greater consistency in handling changes in target size. The overall performance of our algorithm on the LasHeR dataset is exceptional, particularly under the challenging attributes of AIV (Aspect Ratio Variation), DEF (Target Deformation), HO (Occlusion), SV (Scale Variation), and PO (Pose Variation), all of which exhibit significant advantages, with scores that are the highest compared to other algorithms.

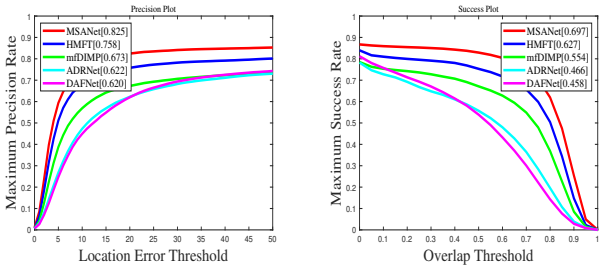
Evaluation on VTUAV dataset. On the VTUAV dataset, MSANet outperforms other algorithms with an accuracy rate of 82.5% and a success rate of 69.7%, demonstrating excellent tracking performance, especially in scenarios with small errors and high overlap requirements The results are shown in **Fig.12**, which proves the excellent performance of MSANet and thus confirms its effectiveness. These scores show that MSANet still exhibits high accuracy and robustness in RGBT tracking from a UAV perspective.

4.3 Ablation Study

In this section, we perform two different ablation studies on the RGBT234 dataset and the LasHeR test set to demonstrate the effectiveness of our proposed method.

	MSANet	CAT++[54]	DAFNet[18]	CAT[24]	FANet[17]	mfDiMP[15]	MANet++[44]	MACNet[42]
AIV	34.9/35.5/34.9	26.7/27.4/25.1	24.4/23.6/19.6	22.6/22.5/19.0	18.8/18.4/18.4	16.6/16.3/16.4	18.8/19.2/15.8	17.3/16.9/15.6
ARC	46.8/43.9/37.4	37.2/33.5/27.8	32.9/29.6/24.7	32.5/29.2/24.4	31.7/27.9/23.9	37.8/33.7/30.9	35.5/32.3/25.7	36.0/32.9/28.5
BC	54.4/50.4/42.5	47.9/42.8/34.4	39.5/37.0/29.7	39.8/36.3/29.8	40.2/36.7/29.5	34.9/30.3/27.0	43.6/39.5/31.4	42.2/37.7/31.9
CM	55.5/50.4/42.6	49.2/43.5/35.1	42.1/37.5/29.8	41.9/37.3/29.4	42.0/37.1/29.3	40.8/34.8/30.6	42.2/37.5/29.4	46.7/41.0/33.9
DEF	52.7/50.8/44.3	43.6/41.3/35.2	40.2/40.3/31.7	38.3/37.1/30.6	33.1/32.9/28.2	40.3/37.8/34.2	39.4/39.3/30.8	41.4/40.8/34.0
FL	45.1/40.1/32.8	38.0/31.5/25.5	33.7/27.7/22.0	38.7/32.8/22.6	35.3/29.8/20.7	32.3/26.7/25.7	37.8/31.8/21.6	34.6/31.1/22.2
FM	52.7/49.1/40.9	45.8/41.0/33.1	39.4/35.3/28.6	39.9/36.3/29.1	38.9/34.6/28.5	41.3/36.5/32.4	41.1/36.4/28.9	43.7/39.2/33.0
HI	57.6/51.2/43.2	50.8/42.1/35.6	47.9/40.2/33.4	52.5/43.1/35.7	52.7/44.0/35.5	46.7/41.4/35.1	53.3/43.4/34.7	52.0/42.3/37.4
HO	32.6/35.1/33.4	25.8/28.5/27.9	14.1/14.9/17.8	22.6/24.9/23.4	16.7/19.8/22.7	19.8/21.1/23.8	24.5/27.7/24.4	28.1/29.6/29.1
LI	45.4/42.6/36.7	38.6/33.7/28.3	35.6/31.1/25.0	31.5/28.2/22.6	33.0/27.9/23.5	29.6/27.2/23.8	35.8/31.5/24.0	36.0/31.0/26.7
LR	51.7/40.9/35.5	49.7/35.8/30.7	43.5/30.9/25.7	42.4/30.8/25.2	43.2/31.8/26.0	40.2/28.7/25.6	47.4/33.7/26.8	43.9/32.5/28.0
MB	49.1/43.9/37.2	44.3/37.0/30.3	38.5/32.8/26.1	39.8/33.3/26.6	40.0/33.0/26.0	37.6/32.4/28.7	39.7/33.1/26.6	40.4/34.5/29.8
NO	77.2/74.7/59.0	71.2/66.8/46.0	70.6/65.2/46.2	65.4/59.7/43.0	59.7/55.8/40.5	76.5/73.2/57.5	63.6/57.7/40.7	74.0/68.4/51.7
OV	41.7/38.2/31.9	27.9/33.0/25.1	25.1/29.3/23.2	26.0/30.2/23.0	24.7/30.4/23.6	40.6/39.2/34.9	28.0/31.0/22.0	34.8/41.8/36.7
PO	54.5/49.6/42.0	48.2/41.6/34.3	41.8/35.9/29.3	41.8/36.4/29.5	41.5/35.5/29.2	39.7/34.3/30.8	44.0/37.9/30.1	44.6/38.6/32.8
SA	47.5/43.1/37.5	44.9/38.0/32.0	40.2/33.8/28.4	37.4/32.0/26.5	39.1/32.6/28.2	37.2/30.6/29.5	41.1/34.2/27.9	40.8/34.2/30.4
SV	56.7/52.2/43.5	49.9/44.0/34.8	44.1/38.6/30.4	44.4/39.2/30.7	44.1/38.7/30.7	45.2/39.9/34.9	46.4/40.6/31.1	48.0/42.6/34.8
TC	48.8/44.1/38.4	44.4/37.9/31.5	38.4/31.5/26.3	37.0/31.4/26.2	37.4/30.7/26.4	38.0/32.6/28.8	40.1/32.6/26.8	39.8/32.7/28.7
TO	47.8/44.5/37.7	41.1/36.3/30.2	34.4/31.4/24.4	36.1/33.0/26.0	34.1/30.8/25.0	32.2/26.4/25.0	35.4/32.3/25.4	38.6/34.1/29.2

Table 2: Attribute-based comparison with seven competitors on the lasher dataset. the best and second results are in red and blue colors ,respectively .



(a) PR (b) SR

Fig. 12: Precision Rate (PR) and Success Rate (SR) for the VTUAV dataset.

Analysis of Vision State-Space Module. We evaluated the VSSMs incorporated into various layers and summarized the results in Table 3. The best performance is observed in both datasets when VSSMs are incorporated into each layer, indicating that the configuration of VSSMs is reasonable. Furthermore, these results demonstrate that the system's performance is significantly improved regardless of the layer to which the VSSM is incorporated. This observation further confirms the reasonableness of the VSSM configuration. The overall performance of the system improves with the introduction of VSSMs at each layer. Specifically, the incorporation of

Table 3: Ablation Studies of Feature Extraction in VSSM.

Layers(VSSM)			RGBT234		LasHeR	
Conv1	Conv2	Conv3	PR	SR	PR	SR
✓	×	×	83.1	59.4	55.1	41.6
✓	✓	×	83.7	59.8	55.8	42.1
✓	✓	✓	85.9	61.6	57.2	43.9

VSSMs at each layer refines data feature extraction and enhances the model's learning capability, resulting in excellent outcomes across various datasets. Additionally, the reasonable configuration of VSSMs is evident in their adaptability to the requirements of various datasets, offering a flexible and efficient solution. Regardless of the datasets' nature, the layered incorporation strategy of VSSMs ensures optimal performance in feature extraction and data processing, demonstrating a high degree of robustness and adaptability. The incorporation and reasonable configuration of VSSMs in each layer not only enhance system performance but also validate the effectiveness and feasibility of this technique in practical applications. This finding establishes a solid theoretical foundation and practical basis for further optimization and promotion of VSSMs, demonstrating their significant potential in enhancing system performance and broad applicability.

Analysis of Multi-Scale Fusion Mechanish. As illustrated in Table 4, the introduction

Table 4: Ablation Studies of Fusion Units in MSFM

Layers(MSFM)			RGBT234		LasHeR	
Conv1	Conv2	Conv3	PR	SR	PR	SR
✓	✓	✓	80.9	56.9	53.5	40.1
×	✓	✓	83.6	57.2	55.8	42.4
×	×	✓	85.9	61.6	57.2	43.9

of MSFM in Conv2 and Conv3 yields 83.6% PR and 57.2% SR on the RGBT234 dataset, and 55.8% PR and 42.4% SR on the LasHeR dataset. Although some enhancements are observed, the results remain unsatisfactory. The introduction of MSFM in Conv1, Conv2, and Conv3 results in a PR drop to 80.9% and an SR drop to 56.9% on the RGBT234 dataset, and a PR drop to 53.5% and an SR drop to 40.1% on the LasHeR dataset. This suggests that the introduction of MSFM solely at Conv1, Conv2, and Conv3 is less effective than other combinations; therefore, we hypothesize that premature feature fusion may be susceptible to modal contamination. The introduction of MSFM solely at Conv3 yields a PR of 85.9% and an SR of 61.6% on the RGBT234 dataset, as well as a PR of 57.2% and an SR of 43.9% on the LasHeR dataset, representing the highest values. This indicates that the introduction of MSFM at the Conv3 layer has the most substantial impact on the overall performance of the system. Optimal Setup: The introduction of MSFM at the Conv3 layer alone represents the most effective approach for achieving optimal PR and SR results across both datasets. This configuration not only reduces computational resource consumption but also achieves superior performance compared to introducing MSFM across all layers. Layer Optimization: Although the introduction of MSFM in multiple layers can enhance performance, the resulting improvement is not as effective as anticipated. Prioritizing optimization at key layers (e.g., Conv3) may represent a more efficient and resource-friendly solution.

Visualization of tracking results. In order to visualise the effectiveness of the proposed MSANet, we compare some advanced trackers in **Fig.13** A number of state-of-the-art trackers are compared. In particular, the visual tracking results of the visual tracking results for sequences, which are all from the RGBT234 dataset. For clarity, we provide three frame pairs for each sequence.

5 CONCLUSION

In this study, we propose a novel Mamba-based multi-scale attention method (MSANet) for RGBT tracking, marking the first application of Mamba in this field. We introduce an emerging strategy, the Vision State-Space Module, in the feature extraction component, which not only leverages the linear complexity of Mamba in long-range modeling and the global effective receptive field but also enhances the efficacy of feature extraction and the robustness of tracking. We demonstrate the superior performance of MSANet compared to existing RGBT trackers across five prominent RGBT tracking datasets, underscoring its robustness and efficacy. We acknowledge that while the incorporation of the Vision State-Space Module in the backbone ensures the robustness of feature extraction, it also imposes a significant computational burden. In future work, we aim to enhance the efficiency of trunk feature extraction and investigate more effective RGBT tracking methodologies.

References

[1] Wang, X., et al.: Describe and attend to track: Learning natural language guided structural representation and visual attention for object tracking. arXiv **1811.10014** (2018)

[2] Jain, D.K., Zhao, X., Gan, C., Shukla, P.K., Jain, A., Sharma, S.: Fusion-driven deep feature network for enhanced object detection and tracking in video surveillance systems. Information Fusion **102429** (2024)

[3] Zhang, P., Li, Y., Zhuang, Y., Kuang, J., Niu, X., Chen, R.: Multi-level information fusion with motion constraints: Key to achieve high-precision gait analysis using low-cost inertial sensors. Information Fusion **89**, 603–618 (2023)

[4] Lesort, T., Lomonaco, V., Stoian, A., Maltoni, D., Filliat, D., Diaz-Rodriguez, N.: Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges. Information Fusion **58**, 52–68 (2020)

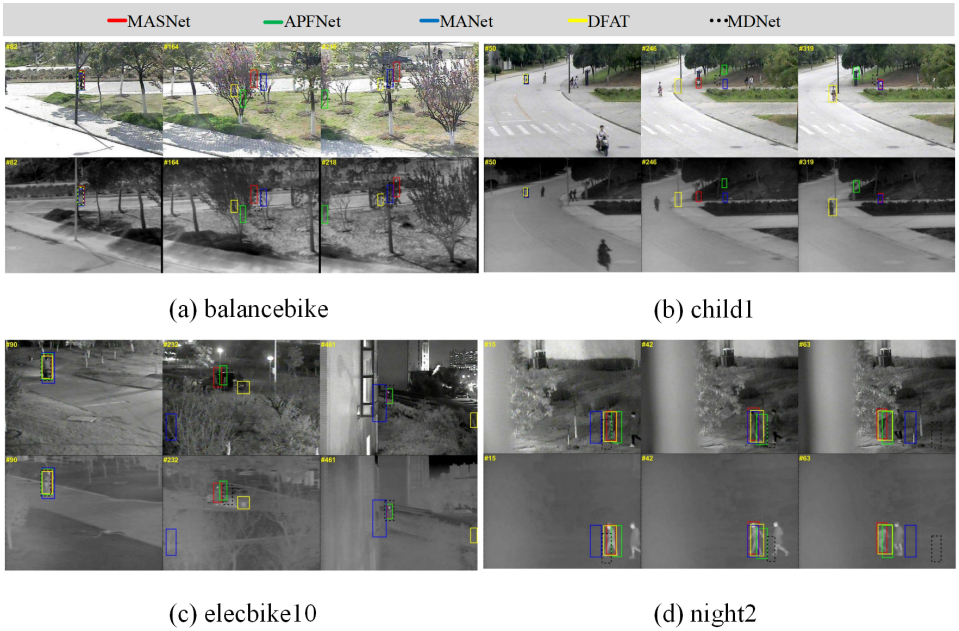


Fig. 13: Qualitative results of MSANet versus other competitors on the RGBT234 dataset

[5] Wang, C., Xu, C., Cui, Z., Zhou, L., Yang, J.: Cross-modal pattern-propagation for rgb-t tracking. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2020)

[6] Guo, H., Li, J., Dai, T., Ouyang, Z., Ren, X., Xia, S.-T.: Mambair: A simple baseline for image restoration with state-space model. arXiv preprint (2024)

[7] Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)

[8] Lu, A., Wang, W., Li, C., Tang, J., Luo, B.: After: Attention-based fusion router for rgbt tracking. arXiv preprint (2024)

[9] Peng, J., Zhao, H., Hu, Z.: Dynamic fusion network for rgbt tracking. IEEE Transactions on Intelligent Transportation Systems **24**(4), 3822–3832 (2023)

[10] Xiao, Y., Yang, M., Li, C., Liu, L., Tang, J.: Attribute-based progressive fusion network for rgbt tracking. AAAI-22 Technical Tracks **3** (2022)

[11] Feng, M., Su, J.: Sparse mixed attention aggregation network for multimodal images fusion tracking. Engineering Applications of Artificial Intelligence **127**, 107273 (2024)

[12] Zhang, T., Liu, X., Zhang, Q., Han, J.: Siamcda: Complementarity- and distractor-aware rgb-t tracking based on siamese network. IEEE Transactions on Circuits and Systems for Video Technology **32**(3), 1403–1417 (2022)

[13] Tang, Z., Xu, T., Li, H., Wu, X.-J., Zhu, X., Kittler, J.: Exploring fusion strategies for accurate rgbt visual object tracking. Information Fusion **99**, 101881 (2023)

[14] Feng, L., Song, K., Wang, J., Yan, Y.: Exploring the potential of siamese network for rgbt object tracking. Journal of Visual Communication and Image Representation **95**, 103882 (2023)

[15] Zhang, L., Danelljan, M., Gonzalez-Garcia, A., Weijer, J., Khan, F.S.: Multi-modal fusion for end-to-end rgb-t tracking. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pp.

2252–2261 (2019)

[16] Li, H., Wu, X., Kittler, J.: Mdlatlrr: A novel decomposition method for infrared and visible image fusion. IEEE Transactions on Image Processing **29**, 4733–4746 (2020)

[17] Zhu, Y., Li, C., Tang, J., Luo, B.: Quality-aware feature aggregation network for robust rgbt tracking. IEEE Transactions on Intelligent Vehicles **6**(1), 121–130 (2021)

[18] Gao, Y., Li, C., Zhu, Y., Tang, J., He, T., Wang, F.: Deep adaptive fusion network for high performance rgbt tracking. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pp. 91–99 (2019)

[19] Zhang, X., Ye, P., Peng, S., Liu, J., Gong, K., Xiao, G.: Siamft: An rgb-infrared fusion tracking method via fully convolutional siamese networks. IEEE Access **7**, 122122–122133 (2019)

[20] Zhang, X., Ye, P., Peng, S., Liu, J., Xiao, G.: Dsiammft: An rgb-t fusion tracking method via dynamic siamese networks using multi-layer feature fusion. Signal Processing: Image Communication **84**, 115756 (2020)

[21] Zhu, Y., Li, C., Luo, B., Tang, J., Wang, X.: Dense feature aggregation and pruning for rgbt tracking. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 465–472 (2019)

[22] Zhu, Y., Li, C., Tang, J., Luo, B., Wang, L.: Rgbt tracking by trident fusion network. IEEE Transactions on Circuits and Systems for Video Technology (2021)

[23] Li, C.L., Lu, A., Zheng, A.H., Tu, Z., Tang, J.: Multi-adapter rgbt tracking. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pp. 2262–2270 (2019)

[24] Li, C., Liu, L., Lu, A., Ji, Q., Tang, J.: Challenge-aware rgbt tracking. In: European Conference on Computer Vision, pp. 222–237 (2020)

[25] Zhang, P., Zhao, J., Bo, C., Wang, D., Lu, H., Yang, X.: Jointly modeling motion and appearance cues for robust rgb-t tracking. IEEE Transactions on Image Processing **30**, 3335–3347 (2021)

[26] Tang, Z., Xu, T., Li, H., Wu, X.-J., Zhu, X.-F., Kittler, J.: Exploring fusion strategies for accurate rgbt visual object tracking. Information Fusion **99**, 101881 (2023)

[27] Tsai, Y.-H.H., Ma, M.Q., Yang, M., Salakhutdinov, R., Morency, L.-P.: Multimodal routing: Improving local and global interpretability of multimodal language analysis. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), p. 1823 (2020)

[28] Zeng, Y., Li, Z., Chen, Z., Ma, H.: A feature-based restoration dynamic interaction network for multimodal sentiment analysis. Engineering Applications of Artificial Intelligence **127**, 107335 (2024)

[29] Lu, A., Wang, W., Li, C., Tang, J., Luo, B.: After: Attention-based fusion router for rgbt tracking. (2024)

[30] Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.: Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(8), 2011–2023 (2020)

[31] Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: Eca-net: Efficient channel attention for deep convolutional neural networks. arXiv preprint arXiv:2004.01467 (2020)

[32] Li, J., Wen, Y., He, L.: Scconv: Spatial and channel reconstruction convolution for feature redundancy. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)

[33] Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Gao, W.: Pre-trained image processing transformer. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 12299–12310 (2021)

[34] Liang, J., Cao, J., Sun, G., Zhang, K., Gool, L.V., Timofte, R.: Swinir: Image restoration using swin transformer. Proceedings of the IEEE/CVF International Conference on Computer Vision, 1833–1844 (2021)

[35] Sun, L., Sakaridis, C., Liang, J., Jiang, Q., Yang, K., Sun, P., Ye, Y., Wang, K., Gool, L.V.: Event-based fusion for motion deblurring with cross-modal attention. arXiv preprint arXiv:2304.12345 (2023)

[36] Li, C., Cheng, H., Hu, S., Liu, X., Tang, J., Lin, L.: Learning collaborative sparse representation for grayscale-thermal tracking. IEEE Transactions on Image Processing **25**(12), 5743–5756 (2016)

[37] Li, C., Zhao, N., Lu, Y., Zhu, C., Tang, J.: Weighted sparse representation regularized graph learning for rgb-t object tracking. In: Proceedings of ACM International Conference on Multimedia (2017)

[38] Li, C., Liang, X., Lu, Y., Zhao, N., Tang, J.: Rgb-t object tracking: benchmark and baseline. Pattern Recognition **96**, 106977 (2019)

[39] Li, C., Xue, W., Jia, Y., Qu, Z., Luo, B., Tang, J., Sun, D.: Lasher: A large-scale high-diversity benchmark for rgbt tracking. IEEE Transactions on Image Processing **31**, 392–404 (2021)

[40] Pengyu, Z., Zhao, J., Wang, D., Lu, H., Ruan, X.: Visible-thermal uav tracking: A large-scale benchmark and new baseline. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2022)

[41] Wang, C., Xu, C., Cui, Z., Zhou, L., Yang, J.: Cross-modal pattern propagation for rgb-t tracking. IEEE Conference on Computer Vision and Pattern Recognition (2020)

[42] Zhang, H., Zhang, L., Zhuo, L., Zhang, J.: Object tracking in rgb-t videos using modal-aware attention network and competitive learning. Sensors (2020)

[43] Zhang, P., Wang, D., Lu, H., Yang, X.: Learning adaptive attribute driven representation for real-time rgb-t tracking. International Journal of Computer Vision **129**, 2714–2729 (2021)

[44] Lu, A., Li, C., Yan, Y., Tang, J., Luo, B.: Rgbt tracking via multi-adapter network with hierarchical divergence loss. IEEE Transactions on Image Processing **30**, 5613–5625 (2021)

[45] Xiao, Y., Yang, M., Li, C., Liu, L., Tang, J.: Attribute-based progressive fusion network for rgbt tracking. Proceedings of the AAAI Conference on Artificial Intelligence, 2831–2838 (2022)

[46] Lu, A., Qian, C., Li, C., Tang, J., Wang, L.: Duality-gated mutual condition network for rgbt tracking. IEEE Transactions on Neural Networks and Learning Systems (2022)

[47] Cheng, A., Lu, A., Zhang, Z., Li, C., Wang, L.: Fusion tree network for rgbt tracking. IEEE International Conference on Advanced Video and Signal Based Surveillance, 1–8 (2022)

[48] Hou, T., Ren, T., Wu, G.: Mirnet: A robust rgbt tracking jointly with multi-modal interaction and refinement. 2022 IEEE International Conference on Multimedia and Expo (ICME), 1–6 (2022)

[49] Pengyu, P., Zhao, J., Wang, D., Lu, H., Ruan, X.: Visible-thermal uav tracking: A large-scale benchmark and new baseline. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2022)

[50] Wang, X., Shu, X., Zhang, S., Jiang, B., Wang, Y., Tian, Y., Wu, F.: Mfgnet: Dynamic modality-aware filter generation for rgb-t tracking. IEEE Transactions on Multimedia (2022)

[51] Peng, H., Zhao, H., Hu, Z.: Dynamic fusion network for rgbt tracking. IEEE Transactions on Intelligent Transportation Systems **24**(4), 3822–3832 (2022)

[52] Mei, D., Zhou, D., Cao, J., Nie, R., He,

K.: Differential reinforcement and global collaboration network for rgbt tracking. IEEE Sensors Journal **23**(7), 7301–7311 (2023)

[53] Zhang, T., Guo, H., Jiao, Q., Zhang, Q., Han, J.: Efficient rgb-t tracking via cross-modality distillation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5404–5413 (2023)

[54] Liu, Y., Li, C., Xiao, Y., Ruan, R., Fan, M.: Rgbt tracking via challenge-based appearance disentanglement and interaction. IEEE Transactions on Image Processing (2024)