Multispectral State-Space Feature Fusion: Bridging Shared and Cross-Parametric Interactions for Object Detection

Jifeng Shen, Haibo Zhan, Shaohua Dong, Xin Zuo, Wankou Yang, Haibin Ling

Paper presentation by Emanuel de la Cruz Espinosa

7/14/2026

Introduction

[a]School of Electrical and Information Engineering, Jiangsu University, Zhenjiang, 212013, China
[b]Department of Computer Science and Engineering, University of North Texas, Denton, TX 76207, USA
[c]School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang, 212003, China
[d]School of Automation, Southeast University, Nanjing, 210096, China
[e]Bodhi Intelligence Lab, Department of Artificial Intelligence, Westlake University, Hangzhou, Zhejiang 310030, China

Information Fusion

Date: March 2026

Article: 103895

Volume: Volume 127, Part C

26.7	17.4
CiteScore	Impact Factor

The source code is available at https://github.com/61s61min/MS2Fusion.git.

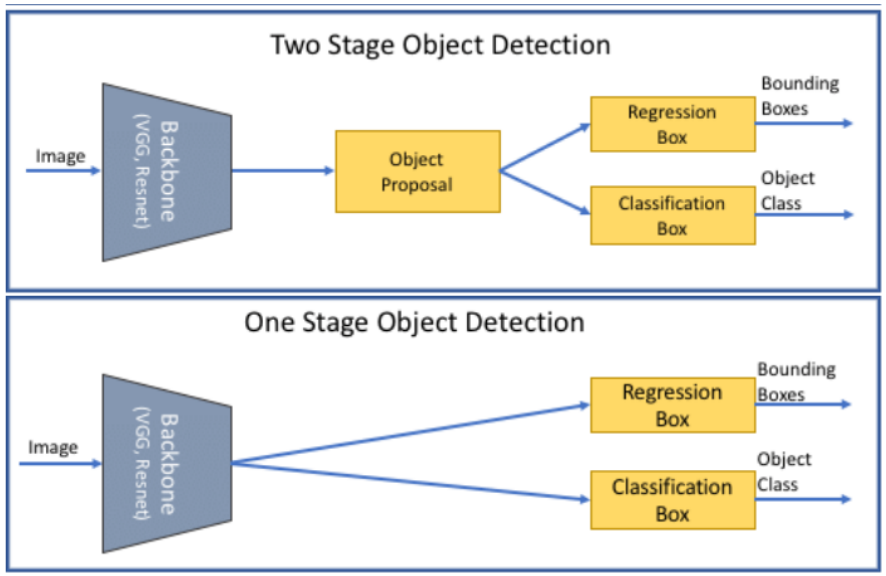
Background

In object detection, **RGB** images are typically used for unimodal detection.

There are two main approaches:

- two-stage detectors, which first generate **region proposals** and then perform **classification** and **bounding box regression**. (R-CNN, Fast R-CNN, Faster R-CNN, Mask R-CNN)

- Single-stage detectors perform object detection directly on the image without generating region proposals, resulting in faster detection speeds. (YOLO, SSD, RetinaNet)



Source publication:Semantic Image Cropping

Anchor-based methods: utilize **predefined** anchor points, each representing specific sizes and aspect ratios, to detect objects **through regression adjustments**.

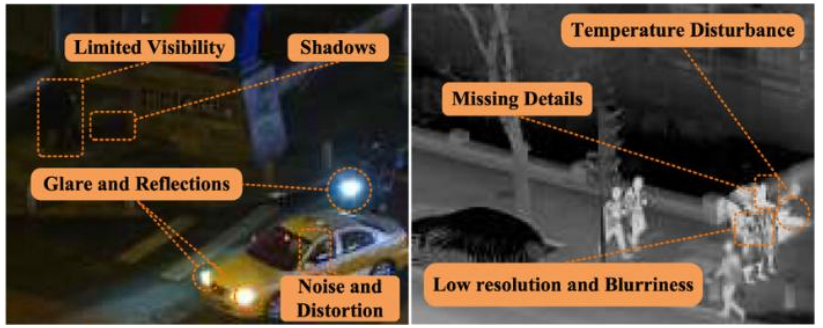
Anchor-free methods: predict object boundaries or centers directly without anchors, simplifying the detection process with **improved efficiency and often higher accuracy**. (CornerNet, FCOS, and CenterNet)

Background

More recently, DETR (DEtection TRansformer) introduced a fully end-to-end approach by leveraging Transformer to eliminate the need for hand-designed components like anchors or non-maximum suppression (NMS).

While achieving competitive accuracy, its **computational cost** and **slow convergence** remain challenges, prompting follow-up improvements like Deformable DETR.

Current single-modal generic object detection methods struggle to overcome complex scenarios, e.g., low illumination, occlusions, among others.



Source publication: DMSC-Net: a multimodal pedestrian detection network based on infrared and visible image fusion

Background

Multispectral object detection has recently attracted increasing interest due to its **robust performance** by fusing information from multiple spectral bands, such as RGB (Visible spectrum) and thermal (long-wave infrared spectrum).

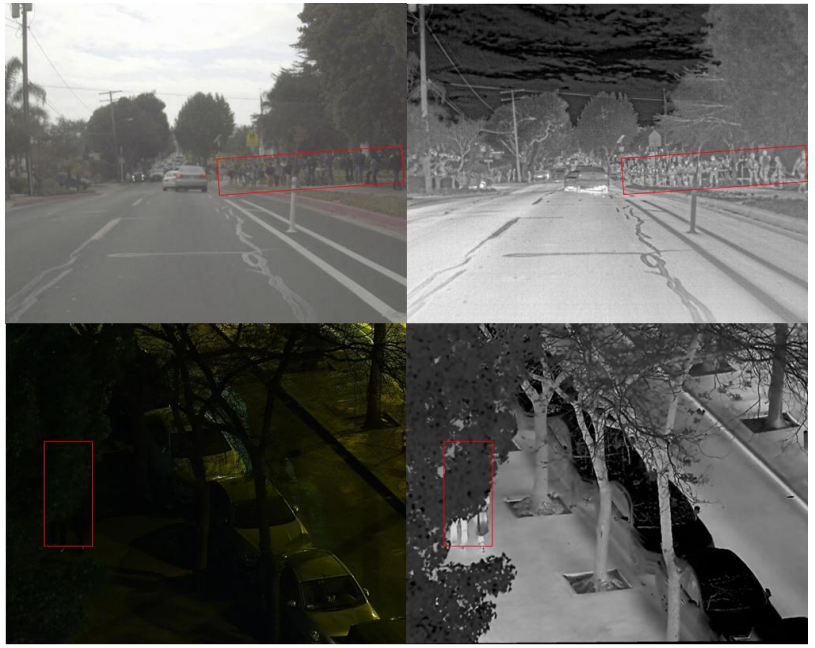
RGB images usually offer **high resolution**, rich **color** and **texture** features, but suffer from sharp performance degradation in complex scenarios such as **low light, adverse weather or occlusions**.

In contrast, thermal images can effectively **overcome these environmental limitations** but exhibit obvious **deficiencies in color and texture details.**

Background

However, scenarios where both modalities exhibit weak discriminative features (e.g., blurred textures in RGB and low contrast in thermal) cannot be resolved by complementary information alone.

Therefore, **shared features**, such as cross-modal consistent shapes and structural patterns, become essential, as they capture modality-invariant representations for reliable detection.



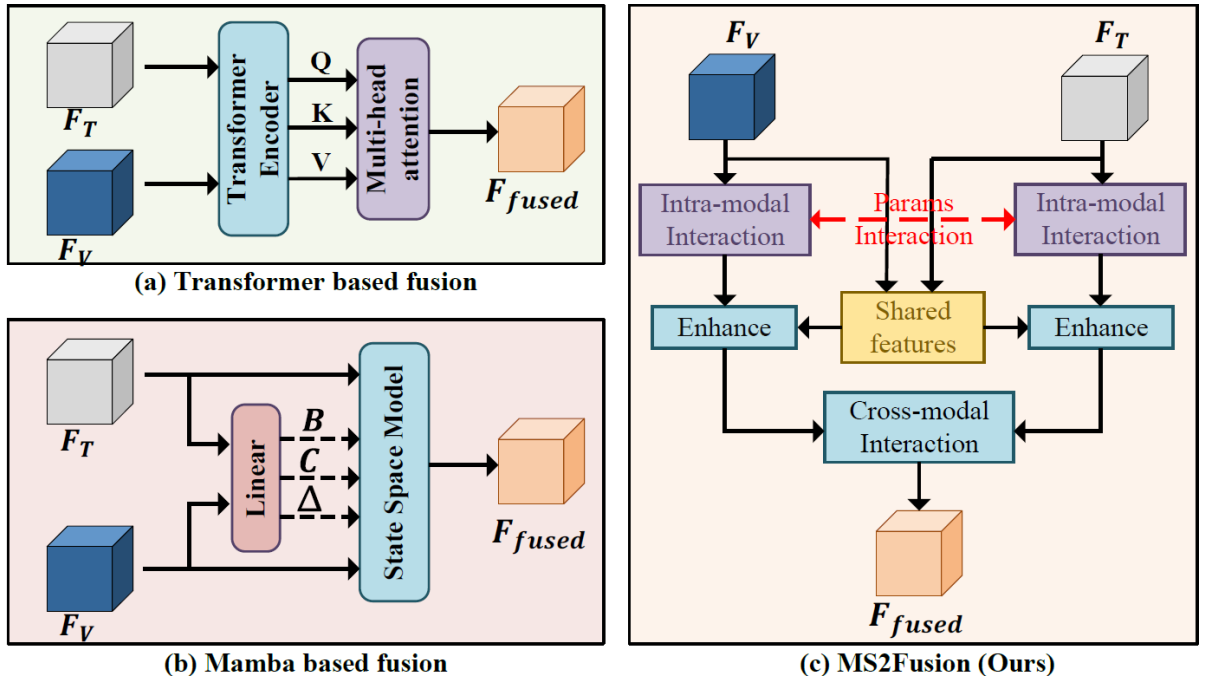
Current multispectral feature fusion for object detection presents two critical challenges:

- Generalization performance due to an unbalanced focus on **local complementary** features over **cross-modal shared** semantics.
- Scalable feature modelling because of the trade-off between the **receptive field size** and **computational complexity**.

Proposal

In Transformer-based method (a), FT and FV are fused through the multi-head attention mechanism, effectively integrating complementary information and enhancing performance across scenarios.

The traditional Mamba approach (b) directly mixes dual-modal features to generate B, C, and Δ parameters for SSM-based feature interaction, which may lead to modal misalignment and feature redundancy.



A Multispectral State-Space Feature Fusion framework (MS2Fusion) is proposed based on the state space model (SSM). First performs intra-modal feature interactions, then extracts cross-modal shared features, achieving better modal alignment and fusion, thereby yielding more robust and unified feature representations.

Proposal

SSM combines the advantages of RNNs and CNNs for efficient handling of long dependencies. It transforms the input sequence *x(t)* into an intermediate state *h(t)* through a state equation, then generates the output y(t) via an output equation. SSM is typically represented as a linear ordinary differential equation:

$$\begin{aligned} h^{'}(t) &= \mathbf{A} \cdot h(t) + \mathbf{B} \cdot x(t) \\ y(t) &= \mathbf{C} \cdot h(t) + \mathbf{D} \cdot x(t) \end{aligned} \tag{1}$$

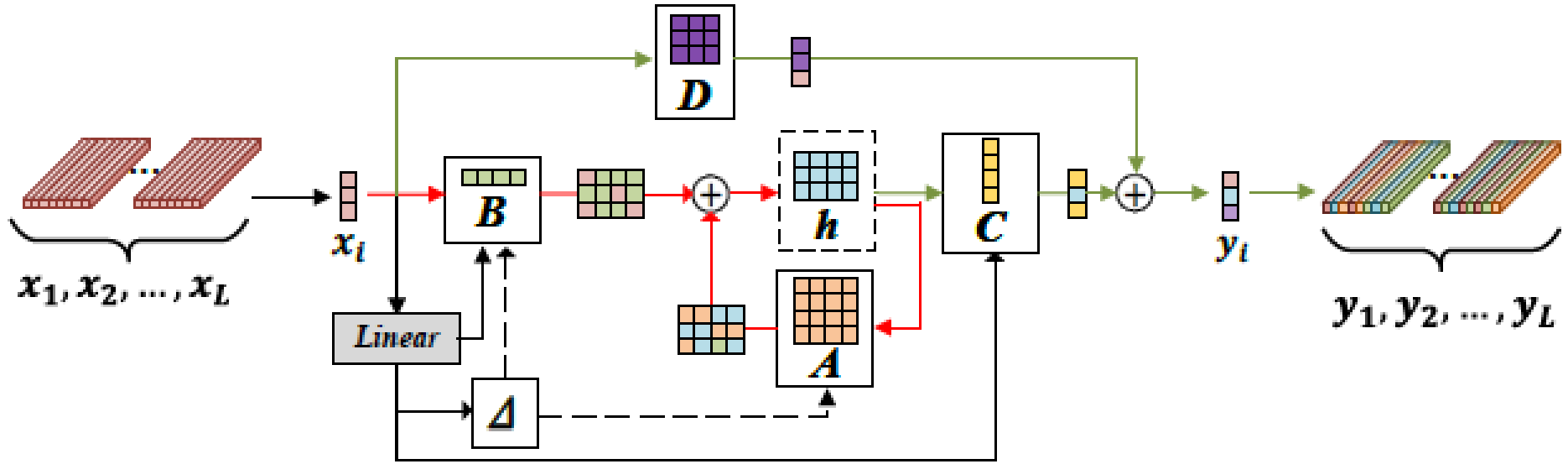
To adapt SSM for deep learning, it is discretized as:

$$\begin{aligned} h_k &= \overline{\mathbf{A}} \cdot h_{k-1} + \overline{\mathbf{B}} \cdot x_k \\ y_k &= \overline{\mathbf{C}} \cdot h_k + \mathbf{D} \cdot x_k \\ \overline{A} &= exp^{\Delta \mathbf{A}} \\ \overline{B} &= \left(exp^{\Delta \mathbf{A}} - \mathbf{I}\right) / \Delta \mathbf{A} \\ \overline{C} &= \mathbf{C} \end{aligned} \tag{2}$$

To overcome the issue of fixed parameters, SSM introduces dynamic adjustments, allowing the model to flexibly handle complex data and improve long-sequence modelling.

Proposal

Details of SSM, where the red, green and black lines correspond to the equations of Equation (2), respectively. (For an L×d dimensional input {x1, x2, ..., xL}.) The input sequence is linearly projected to generate parameters for the state equation, followed by recursive computation via matrices A, B, and C, with an optional skip connection D. The output sequence retains the same dimensionality as the input.



SSM

Proposal

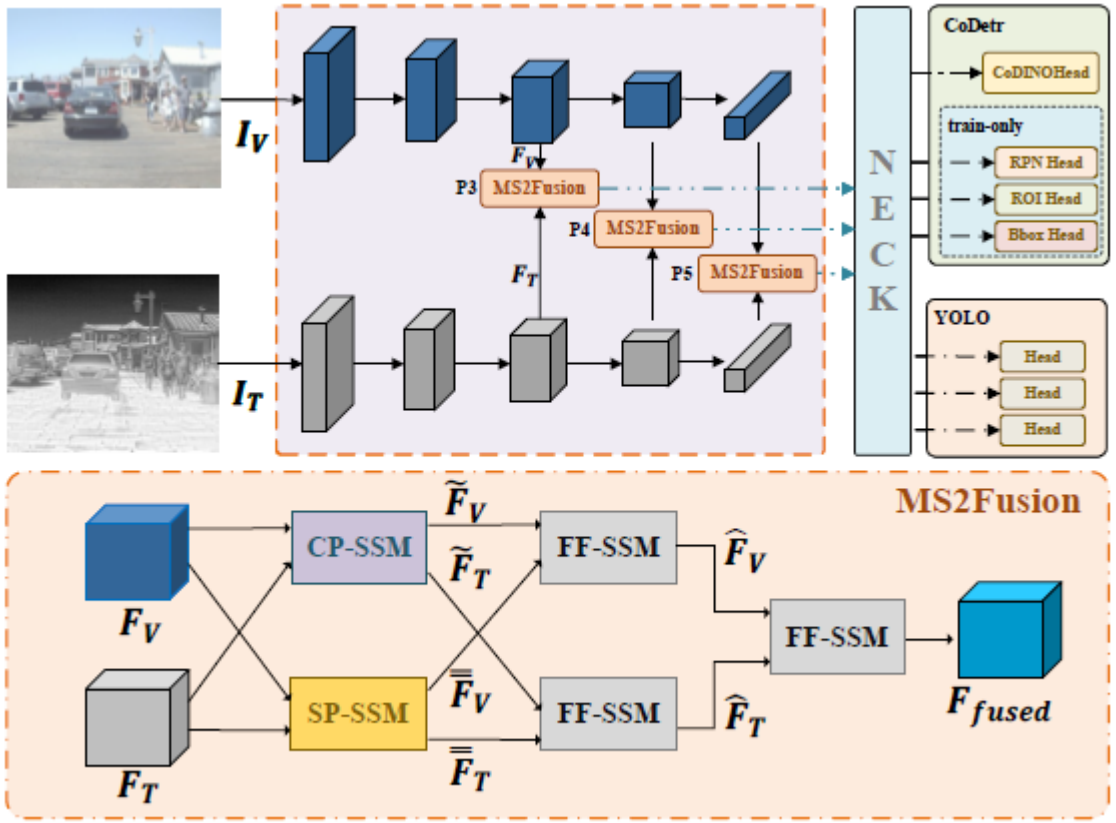
Overview of the model architecture. It consists of three main stages:

(1) **feature extraction** with two backbone networks.

(2) **cross-modal feature fusion** of P3, P4 and P5 with MS2Fusion module.

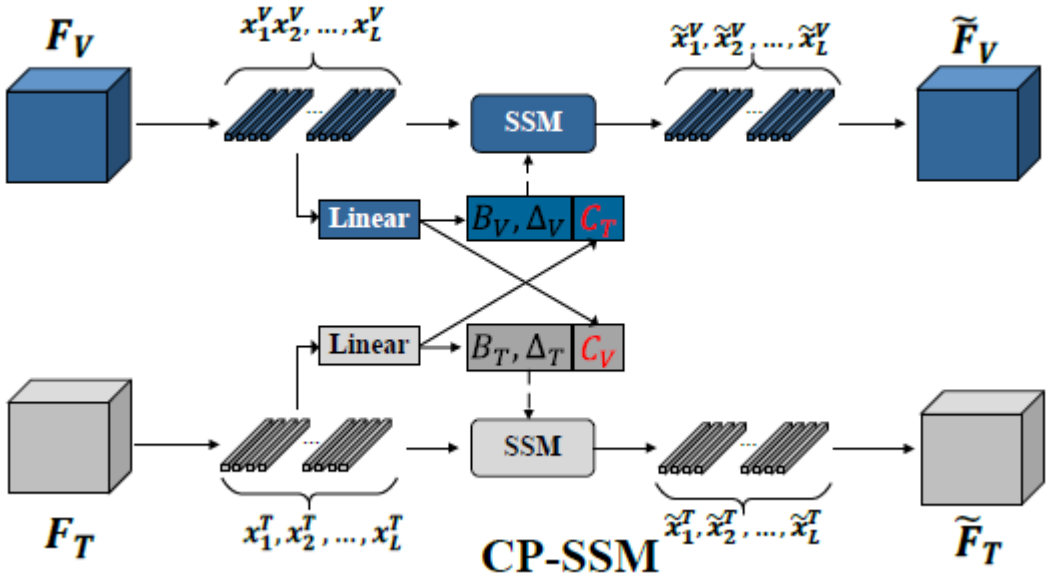
(3) The detection results are generated through the Neck and Head layers.

The MS2Fusion module employs a dual-branch architecture to process the features of two modalities, F_V and F_T. **CP-SSM** learns cross-modal complementary features, while **SP-SSM** extracts shared features. Finally, **FF-SSM** performs single brach feature enhancement and cross-modal fusion, outputting the fused feature F_{fused}.



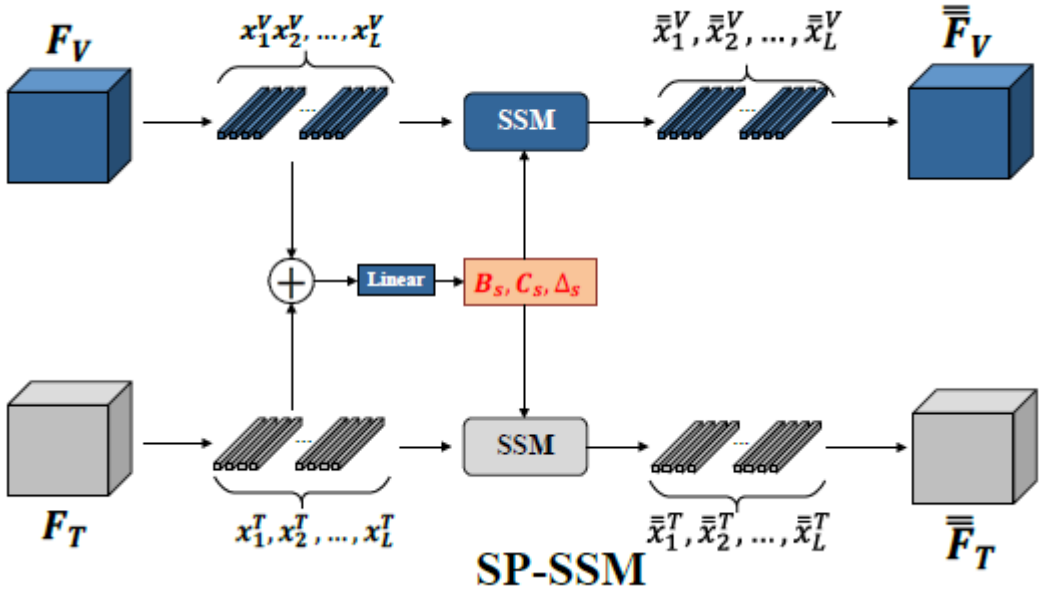
Proposal

CP-SSM module achieves global feature integration of RGB and thermal modalities through a dynamic parameter interaction mechanism, enhancing cross-modal contextual awareness while preserving modality-specific characteristics.



Proposal

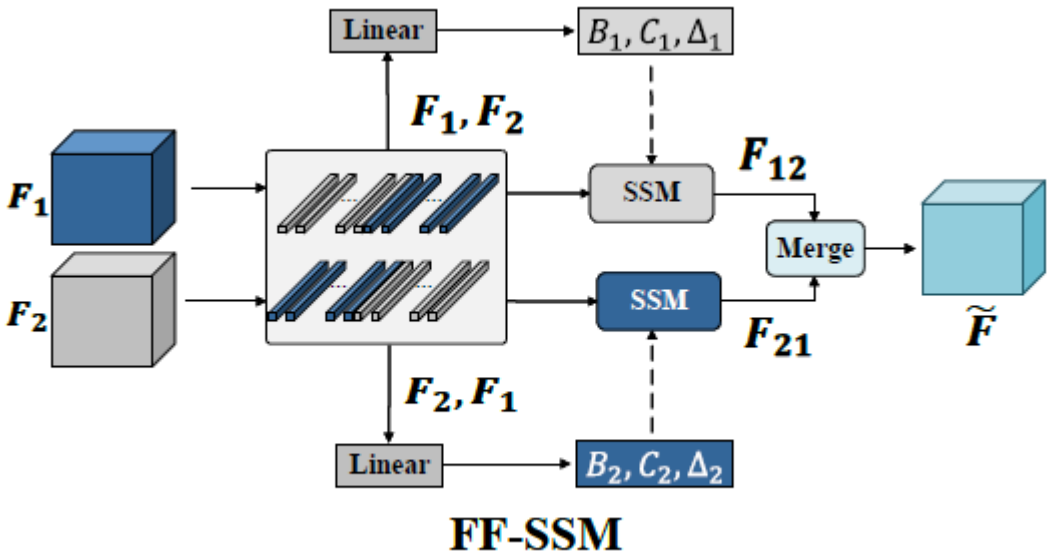
SP-SSM is designed to learn shared, modality-invariant representations that are consistent across both RGB and thermal inputs.



The CP-SSM employs dynamic parameter interaction mechanisms where modality-specific parameters are selectively exchanged or combined to capture complementary cross-modal dependencies, while the SP-SSM strictly maintains identical parameters across modalities to enforce feature consistency in shared spaces.

Proposal

FF-SSM modules further integrate these learned features in a hierarchical manner, adaptively combining complementary and shared information to achieve more effective cross-modal fusion.



FF-SSM

Experiments

All experiments are conducted using **PyTorch** on a computer equipped with an **Intel i7-9700 CPU**, **64 GB RAM**, and a **Nvidia RTX 3090 GPU** with **24 GB** of memory.

For all ablation studies, the number of epochs is set to **60**.

The batch size is set to **4**, and the SGD optimizer is used with an initial learning rate of **0.01** and a momentum of **0.937**.

The weight decay factor is set to **0.0005**, and we employ a cosine learning rate decay schedule. The input size for all training images is **640×640**, while the input size for testing is **640×512**.

Additionally, mosaic augmentation is used for data enhancement.

Experiments

Compared to Transformer, MS2Fusion achieves:

66.6% reduction in FLOPs and 70.4% fewer parameters

while delivering better detection accuracy (+8.3 mAP50 points)

(2) When benchmarked against CNN baseline, it maintains 26.0% lower computational costs and 18.4% parameter reduction, while achieving significant improvements of +9.0 mAP50 points.

Table 1: Comparison of different multispectral feature fusion methods(YoloV5 framework, FLIR dataset)

	GFLOPs	Params (M)	FPS	mAP@0.5	mAP@0.75	mAP
CNN	190.3	159.7	30.8	74.3	23.8	32.5
Transformer	421.9	440.6	19.2	75.0	24.1	33.4
MS2Fusion	140.8	130.3	29.7	83.3	33.0	40.3

Experiments

Table 2: Comparison on the FLIR-align dataset. ('-' indicates missing values. '‡' symbol denotes experimental results obtained using the CoDetr framework, with input images resized to a fixed resolution of 640×640 pixels.)

Methods	mAP@0.5	mAP	Bicycle	Car	Person
MMTOD-CG [49]	61.4	-	50.3	70.6	63.3
MMTOD-UNIT [49]	61.5	-	49.4	70.7	64.5
CMPD [50]	69.4	-	59.9	78.1	69.6
CFR [26]	72.4	-	57.8	84.9	74.5
GAFF [27]	72.9	37.5	-	-	-
BU-ATT [51]	73.1	-	56.1	87.0	76.1
BU-LTT [51]	73.2	-	57.4	86.5	75.6
UA_CMDet [52]	78.6	-	64.3	88.4	83.2
CFT [2]	78.7	40.2	-	-	-
CSAA [53]	79.2	41.3	-	-	-
ICAFusion [8]	79.2	41.4	66.9	89.0	81.6
CrossFormer [54]	79.3	42.1	-	-	-
MFPT [55]	80.0	-	67.7	89.0	83.2
MMFN [34]	80.8	41.7	65.5	91.2	85.7
RSDet [56]	81.1	41.4	-	-	-
MiPa [57]	81.3	44.8	-	-	-
UniRGB-IR [58]	81.4	44.1	-	-	-
CPCF [32]	82.1	44.6	-	-	-
GM-DETR [59]	83.9	45.8	-	-	-
Fusion-Mamba(yolov5) [60]	84.3	44.4	-	-	-
Fusion-Mamba(yolov8) [60]	84.9	47.0	-	-	-
TFDet [33]	86.6	46.6	-	-	-
DAMSDet[30]	86.6	49.3	-	-	-
Ours	83.3	40.3	74.9	89.8	85.1
Ours‡	**87.8**	**49.7**	**79.6**	**93.4**	**90.2**

Experiments

Table 9: Effect of different backbones and detection frameworks.

Num	Framework	Fusion Model	Person	Car	Bicycle	mAP@0.5	FPS
1	YOLOv5	Baseline	83.8	88.9	67.3	80.0	43.2
2		MS2Fusion	85.1	89.8	74.9	83.3 (+3.3)	24.4
3	CoDetr	Baseline	88.0	91.7	73.8	84.5	5.9
4		MS2Fusion	90.2	93.4	79.6	87.8 (+3.3)	4.8

Table 10: Effect of different backbones.

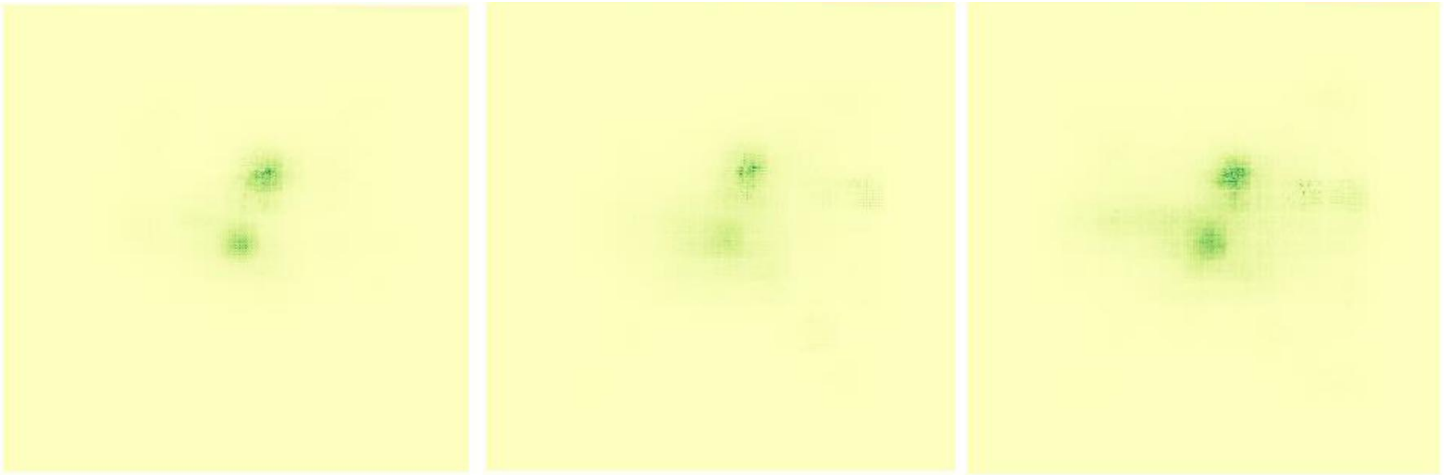
Num	Backbone	Fusion Model	Person	Car	Bicycle	mAP@0.5
1	VGG16	Baseline	78.9	87.7	51.2	72.6
2		MS2Fusion	79.0	87.8	53.8	73.6 (+1.0)
3	ResNet50	Baseline	76.8	85.7	44.6	69.0
4		MS2Fusion	80.2	88.4	52.8	73.8 (+4.8)
5	CSPDarkNet53	Baseline	83.8	88.9	67.3	80.0
6		MS2Fusion	85.1	89.8	74.9	83.3 (+3.3)

Table 11: Effect of different fusion modules.

Num	CP-SSM	SP-SSM	FF-SSM	Person	Car	Bicycle	mAP@0.5	Params(M)
1				83.8	88.9	67.3	80.0	72.7
2	✓			83.8	89.8	72.7	82.1 (+2.1)	84.5
3		✓		85.6	90.4	70.6	82.2 (+2.2)	90.4
4			✓	84.9	90.1	72.4	82.5 (+2.5)	86.0
5	✓	✓		84.9	90.3	72.9	82.7 (+2.7)	117.0
6	✓		✓	85.5	90.0	72.5	82.7 (+2.7)	97.8
7		✓	✓	85.5	90.3	72.9	82.9 (+2.9)	103.7
8	✓	✓	✓	85.1	89.8	74.9	83.3 (+3.3)	130.2

Experiments

Quantitative analysis demonstrates that MS2Fusion achieves significantly broader receptive field coverage compared to the others, successfully integrating local details with global context. This capability addresses the fundamental constraint of standard CNN in modeling long range dependencies while maintaining computational efficiency



(a) CNN-based fusion (b) Transfomer-based fusion (c) MS2Fusion

Experiments

Figure 14: Visual comparison among different SOTA methods and Ours.

Limitations

- Distant objects: The **method struggles to accurately distinguish the object from its surroundings.**

- Occluded objects: The **method tends to overlook occluded objects, leading to missed detections.**

- False positives with similar thermal appearance: **objects with thermal characteristics similar to those of a person result in false-positive detections**. This challenge is exacerbated by the low resolution of the images, which hampers the model's ability to differentiate between actual objects of interest and irrelevant background features.

These failure cases demonstrate the critical need to improve our model's capability to handle three key challenges: **distant small object detection**, **occluded object detection**, and **discrimination between thermally similar objects**, especially in low-resolution images.

Conclusions

A robust multispectral object detection framework should dynamically leverage both complementary and shared features to handle diverse real-world challenges.

Unlike conventional approaches that rely on rigid fusion strategies or suffer from computational inefficiency, our method leverages dynamic state space models to efficiently capture long-range dependencies while maintaining low computational complexity, overcoming the drawbacks of both CNN-based and Transformer-based techniques.

Beyond the current scope, the proposed fusion methodology will be extended to other multimodal sensing tasks under computational and bandwidth constraints, supported by our ongoing construction of a multispectral UAV–maritime dataset for large-scale validation.