PSRDET: Fast Multimodal Detection based on Prior Scene Repair for All-weather Road Sensing

Chengbo Yu[1], Dengshi Li[1], Jia Wei[1], Yihui Wang[1], and Haiyang Ye[1]

Jianghan University, No. 8, Delta Lake Road, Wuhan Economic and Technological Development Zone, Wuhan, China yuchengbo@stu.jhun.edu.cn, reallds@jhun.edu.cn, 3270584097@qq.com, wyh6038@stu.jhun.edu.cn, y0h1y8@gmail.com

Abstract. Multispectral object detection using both RGB and infrared (IR) data improves accuracy and robustness by fusing complementary information from different modalities. However, current methods lack adaptability to real-world application scenarios and are not conducive to practical deployment.Therefore, we propose a fast multimodal detection based on prior scene repair for all-weather road sensing (PSRDET): Firstly, we design a Cross-Visual State Space Fusion (CVSSF) module that maps cross-modal features to the hidden state space for asymptotic interaction. Facing the challenges of various complex environments, we propose the prior-based degradation-resistant hybrid expert (PDR-MoE) module to adaptively adjust the visible images to mitigate the image degradation problem caused by harsh environments. Finally, the module is embedded into our dual-stream backbone network. Experiments show that PSRDET achieves state-of-the-art performance on FLIR and M3FD benchmark datasets, thus offering greater safety and environmental adaptability in the field of autonomous driving.

Keywords: Multimodal object detecton · Mixture of experts · State space model · Cross-modal fusion.

1 Introduction

Object detection is crucial for autonomous driving, yet it is challenged by varying road conditions and the limitations of single-sensor systems like RGB cameras, which struggle in rain, fog, and low light. However, infrared cameras detect thermal radiation and are less affected by such conditions, making them a valuable addition to detection systems. Despite this, infrared images lack the detail of visible-light images. The key research focus now is on creating efficient fusion algorithms that combine the strengths of both infrared and visible images to improve multispectral object detection performance.

In detection tasks, CNNs [7, 6, 20, 15, 12] are favored for their fast and accurate execution, but they may be deficient in handling image correlation. Transformers [31], especially the DETR [1, 36] family, provide better global modeling capabilities through the self-attention mechanism, but this structure is computationally expensive. As a result, many hybrid models start to appear such as

EfficientFormer [11], MobileVit [19] and EdgeVit [2], which try to combine the advantages of CNN and Transformer to increase the speed and reduce the complexity. However, these hybrid models also face the challenge of performance degradation.

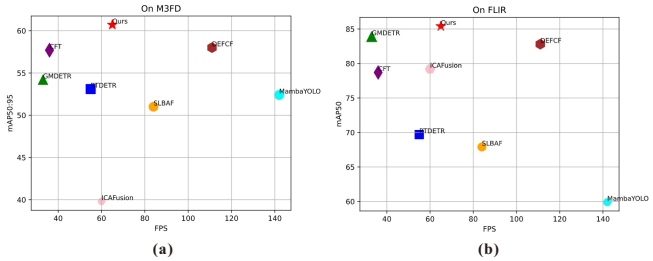

Fig. 1. Comparison of real-time object detectors on the M3FD [14] dataset (left) and the FLIR [32] dataset (right). Our method achieves a perfect balance between size-accuracy and FPS.

The YOLO family is known for its fast execution speed and real-time detection capabilities, and each new version attempts to improve performance by introducing new techniques and architectural improvements. Some recent studies incorporate Transformer structures into the YOLO model, such as ViT-YOLO [34] and YOLOS [5], but these attempts have not led to the expected performance improvements, but instead have exposed the scalability issues of the Transformer for object detection tasks. Gold-YOLO [26], on the other hand, attempt to enhance feature fusion through convolutional and attentional mechanisms, but also face limitations in performance improvement.

Recent approaches based on this state-space model (SSM), such as Mamba [8], have been able to capture long-range dependencies while allowing the model to remain linear in computational complexity. Surprisingly, researchers have successfully introduced the Mamba architecture into the vision domain with success in image segmentation tasks [18, 33]. MambaOut [30] states that Mamba has potential for object detection tasks.

Therefore, we focus on how to apply the advantages of the SSM model to the object detection task and design effective fusion strategies. First, the detection backbone is extended to a parallel dual-stream backbone. Next, we design a Prior-Based Degeneration-Resistant Mixture of Experts module (PDR-MoE) and then a Cross Visual State Space Fusion module (CVSSF). Finally, the PDR-MoE module and the CVSSF module are embedded into the dual-stream network. Specifically this paper has three major contributions:

– We propose PSRDET, a real-time multispectral object detection model based on SSM that integrates an efficient dual-stream backbone fusion encoder. The network extension approach is generic and can be used for any single stream network.

– We designed two innovative modules, the CVSSF module is proposed to efficiently fuse information across modalities with linear complexity. The PDR-MoE module to remove the degradation effects received by visible images in harsh environments.
– We conduct experiments on two public datasets, FLIR [32] and M3DF [14] , and demonstrate that the performance and efficiency of our method outperforms other state-of-the-art methods, as shown in Fig. 1.

2 Related Work

2.1 Multispectral Object Detecton

Multispectral object detection algorithms are mainly derived from conventional object detection algorithms such as Faster R-CNN [21], RetinaNet [13], YOLOv5, and DETR [1]. The core challenge in realizing multispectral object detection lies in exploring and developing effective strategies for integrating these two different modes of image information.

Currently, fusion strategies are divided into two main categories: early fusion and feature fusion. Feature fusion strategies involve the construction of two independent dual-stream networks that operate in parallel and are responsible for extracting features with similar shapes, subsequently integrating these features, and performing object detection based on the integrated features.

In recent years, feature fusion has been widely studied as a fusion strategy. Sharma et al. [22] first proposed to use the middle layer of the network to fuse multispectral data to improve the detection performance, and then researchers developed a dual-stream network architecture that integrates multiple fusion and enhancement modules. SwinFusion [17], which is based on cross-domain long-range learning and the Swin Transformer, strengthens the fusion of information between different domains and global interactions through the attentional mechanism and compensates for the shortcomings of traditional fusion methods. PIAFusion [25] effectively integrates the information of infrared and visible images through cross-modal difference-aware fusion and halfway fusion strategy to solve the lighting problem, but the model is simple and has limited adjustment ability in complex lighting environments. DIVFusion [24] combines a low-light image enhancement task and a bimodal fusion task in a unified framework to address weak texture detail and poor visual perception due to insufficient illumination in nighttime image fusion. ICAFusion [23] not only enhances the feature representation, but also effectively reduces the complexity and computational cost of the model by introducing an iterative feature fusion module and a cross-attention mechanism.

It is not difficult to find two problems with previous multispectral detectors: (1) The inherent limitations of their own backbone network and fusion strategy constrain the performance and speed of their model in the context of complex environmental scenarios. (2) These approaches fail to account for the detrimental

impact of visible modal information degradation on the network in extreme environments. Consequently, this paper proposes the PDR-MoE module for feature enhancement and the CVSSF module for feature fusion.

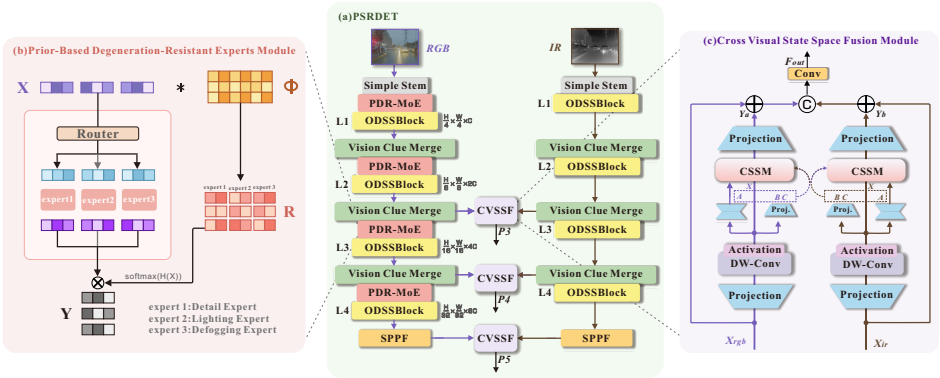

Fig. 2. The architecture of our **PSRDET**. (a) Overview: RGB and IR images are fed together into a two-stream backbone network. In this way, the ODSSBlock [28]for each modality is carried out independently to extract the respective information.The PDR-MoE module is used for visible branching, and the CVSSF module is used for multimodal feature aggregation, and these fused features are subsequently imported into the downstream detection module. (b) PDR-MoE module. (c) CVSSF module.

3 Methodology

3.1 Overall Framework

As shown in Fig. 2(a), PSRDET consists of two parallel two-stream backbone networks and neck part , in the necking part, we adopt the design of PAFPN [10]. It is worth noting that PSRDET is built on top of a real-time detector (MambaYOLO) [28], which is an end-to-end object detector built on SSM. For a pair of visible and infrared images $\{x_{ir}, x_{rgb}\}$, two SSM backbone networks are first used to extract features from the dual stream images. In the visible image streaming backbone network, we incorporate the PDR-MoE module to deal with the degradation of visible images in harsh environments. In stage L2,L3,L4, the dual streaming features are used as inputs to the fusion module. We propose the CVSSF module to incorporate the input of the multimodal features.The outputs of the fusion module are sent to the PAFPN [28].

3.2 Prior-Based Degeneration-Resistant Experts Module

As shown in Fig. 2(b), RGB visible image features $X \in \mathbb{R}^{H\times W\times C}$ are used as inputs to the module, and the input features are transformed into permutations

by a learnable matrix $\mathbf{\Phi} \in \mathbb{R}^{C\times P}$, where P is the capacity factor. The transformation generates a routing matrix $R \in \mathbb{R}^{HW\times P}$. Noise is added to the routing matrix to enhance the generalisation of the model:

$$H(x)_i = R_i + \mathcal{N}(0,1) \cdot \text{Softmax}((X \cdot \mathbf{\Phi}_{\text{noise}})_i), \tag{1}$$

where Softmax$(\cdot)$ is defined as Softmax$(z) = \log(1 + e^z)$, which is a smooth activation function used to ensure the non-negativity of the noise. $\mathcal{N}(0,1)$ denotes a random number drawn from a standard normal distribution. $\mathbf{\Phi}_{\text{noise}}$ is a learnable weight matrix. The TopK$(\cdot)$ is applied, keeping the largest K values in each $H(x)$ and setting the remaining values to negative infinity:

$$\text{TopK}(v,k)_i = \begin{cases} v_i & \text{if } v_i \text{ is in the top } k \text{ elements of } v, \\ -\infty & \text{otherwise} \end{cases} \tag{2}$$

The Softplus$(\cdot)$ is applied to the TopK$(\cdot)$ processed routing matrix to obtain the weights of each expert:

$$G(X) = \text{Softmax}(\text{TopK}(H(X),k)), \tag{3}$$

where k denotes the number of experts selected. The outputs of all the experts are multiplied with their corresponding weights and then summed to get the final output:

$$Y = \sum_{i=1}^{n} G(X)_i E_i(X), \tag{4}$$

where n indicates the number of experts, $E_i(X)$ denotes the output of the i-th expert, and Y denotes the output of the overall module. Each expert receives the same feature map as input and computes its output. Only selected experts are activated and their outputs are computed.

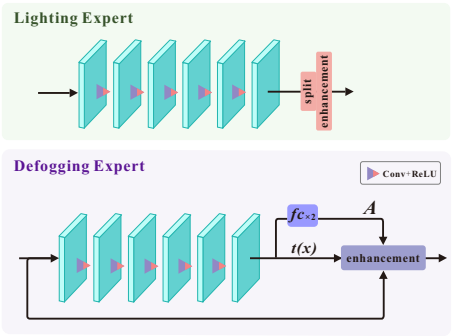

Fig. 3. The architecture of our Lighting Expert and Defogging Expert.

Lighting Expert. Inspired by the low-light enhancement method Zero-Dce [9], the light enhancement is converted into an image-specific curve estimation

problem:

$$LE(I(x);\alpha) = I(x) + \alpha I(x)(1 - I(x)), \tag{5}$$

where $\alpha \in [-1, 1]$ is the trainable parameter of the curve.By means of the parameter α to adjust the number of curve steps and the exposure level.As shown in Fig. 3, we construct our light expert through a simple convolutional network using this LE-curve to iteratively augment the given features.

Defogging Expert. Inspired by our own light experts and atmospheric scattering models:

$$I(x) = J(x)t(x) + A(1 - t(x)), \tag{6}$$

where $t(x)$ denotes transmittance, i.e., the degree of attenuation of light as it travels through the atmosphere to the observer. A denotes atmospheric light, that is, background light caused by atmospheric scattering. All we have to do is estimate these two components through the network structure of Fig. 3. We are concerned that degradations such as fog and dark light are not uniformly distributed in the original image, so the de-degradation parameters are generated independently and uniquely for each pixel in the feature.

Algorithm 1 CSSM2D Block Process

Require: $X_{rgb} : (HW, C), X_{ir} : (HW, C)$
Ensure: $Y_a : (HW, C)$
$x_{rgb} : (HW, C) \leftarrow \text{SiLU}(\text{Conv}(X_{rgb}))$
$x_{ir} : (HW, C) \leftarrow \text{SiLU}(\text{Conv}(X_{ir}))$
$A : (C, N) \leftarrow \text{Parameter}_A$
$B : (HW, N) \leftarrow \text{Linear}_B(X_{ir})$
$C : (HW, N) \leftarrow \text{Linear}_C(X_{ir})$
$\Delta : (HW, C) \leftarrow \log(1 + \exp(\text{Linear}_\Delta(X_{ir}) + \text{Parameter}_\Delta))$
$\overline{A} : (HW, C, N) \leftarrow \exp(\Delta \otimes A)$
$\overline{B} : (HW, C, N) \leftarrow \Delta \otimes B$
$h : (C, N) \leftarrow \text{zeros}(C, N)$
$y : (HW, C) \leftarrow \text{zeros}(HW, C)$
for i in $\{0, \ldots, M-1\}$ **do**
  $h = \overline{A}[:, i, :, :] \odot h + \overline{B}[:, i, :, :] \odot x_{rgb}[:, i, :, None]$
  $y[:, i, :] = h \otimes C[:, i, :]$
end for
return Y_a

3.3 Cross Visual State Space Fusion Module

We process the visible and infrared images separately by two parallel VSS2D [16] modules to generate the CVSSF module shown in Fig. 2(c). In the CVSSF module, after scanning expansion, the two images are decomposed into a series of subimages, each representing a specific direction. From a diagonal point of view, the scanning expansion is performed along four symmetric directions: top-down,

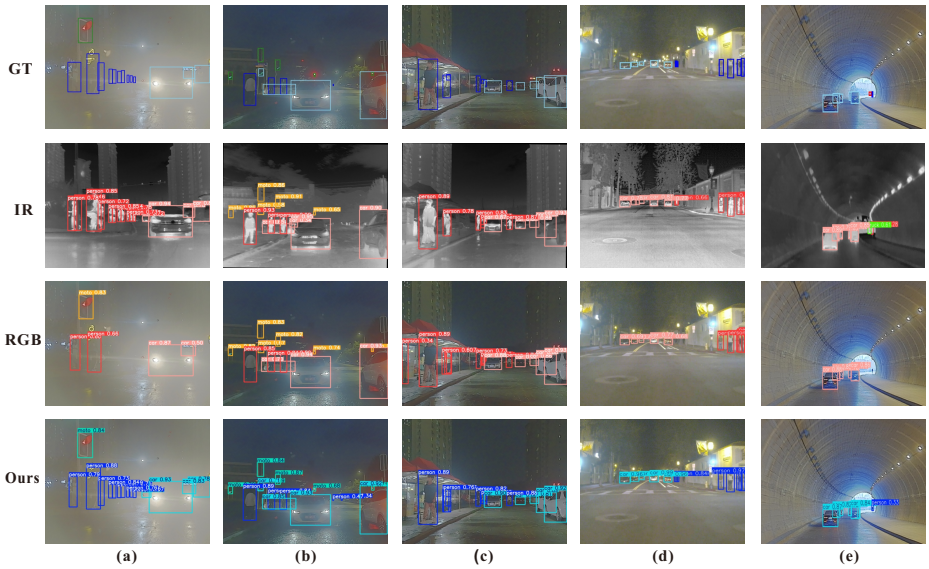

Fig. 4. Visualisation of GTs and detection results for a single model baseline and our method on both datasets.

bottom-up, left-right, and right-left. In order to perform feature extraction on the subimages of the two modes in the CSSM2D module, we use different parameters to calculate Δ, B and C for the two modes of visible and infrared images. The exact process is shown in 3.2. Subsequently, these computed parameters are exchanged and provided to each other for ssm computation. Then, the output of the parallel module is transformed into the output image F_{out} by convolution operation. In this module, we are more concerned with the information between the two cross-modalities, while in ODSSBlock, they are more concerned with their respective information.

4 Experiment

4.1 Datasets and Implementation Details

We evaluate our method on M3FD [14] and FLIR [32] datasets.

M3FD. The M3FD dataset contains a total of 4200 infrared and visible image pairs, which are used for fusion, detection and fusion-based detection tasks. In addition to this, 300 scene-independent image pairs are used for the fusion task. It covers six categories: "person", "car", "bus", "motorcycle", "lamp", and "truck".

FLIR. The FLIR dataset provides RGB and infrared image pairs for day and night scenes. However, the original dataset contains many unaligned images and only the IR images are labeled, which makes training difficult. We use the

aligned version provided by [32] , which contains 5,142 image pairs, of which 4,129 pairs are used for training and 1,013 pairs for testing. It contains 3 categories: "person", "car" and "bicycle". Subsequent references to FLIR refer to the aligned version.

Implementation Details. In this experiment,we used a detection model based on MambaYOLO [28]model and trained it on NVIDIA GeForce RTX 3090 GPU 24GB. For the training strategy, we set a batch size of 8 and a number of working threads of 8. The initial learning rate is set to 0.001 and 200 epochs of training are performed by the SGD optimiser. In the first 3 rounds of training, we implemented a learning rate warm-up strategy and applied mosaic enhancement to improve the robustness of the model to image variations, while turning it off in the last 10 rounds. All input images were resized to 640Œ640 pixels.

Table 1. Comparison of performances on The M3FD dataset and The FLIR dataset.The best accuracy is indicate in **bold**.

Data	Methods	M3FD				FLIR				Model Size/M	FPS
		P	R	mAP_{50}	$mAP_{50:95}$	P	R	mAP_{50}	$mAP_{50:95}$		
IR	MambaYOLO	83.2	70.0	77.0	51.2	80.2	64.6	73.1	38.5	11.7	147
IR	REDETR	86.7	78.5	85.4	55.1	-	-	80.5	43.6	63.1	55
RGB	MambaYOLO	85.2	71.0	78.8	52.4	71.3	53.5	59.9	29.2	11.7	142
RGB	REDETR	85.8	78.0	83.9	53.1	-	-	69.7	33.5	63.1	55
IR+RGB	SLABF	85.3	78.8	85.7	51.0	66.4	63.4	67.9	32.7	2.80	84
IR+RGB	CFT	**88.8**	84.1	**86.9**	57.7	-	-	78.7	40.2	148.2	36
IR+RGB	ICAFusion	77.9	76.0	80.8	39.8	-	-	79.2	41.4	241.3	60
IR+RGB	DFECF	86.8	**85.2**	86.7	58.0	-	-	82.8	45.3	165.1	111
IR+RGB	GMDETR	-	-	85.3	54.2	-	-	83.9	45.8	70.0	33
IR+RGB	**PSRDET(Ours)**	87.7	80.8	86.4	**60.7**	**82.9**	**76.4**	**85.4**	**51.3**	19.5	75

4.2 Comparison With State-of-the-Art Methods

FLIR. On the FLIR dataset, we compare our proposed method with previous work, including unimodal detection methods: MambaYOLO [28], RT-DETR [35], trained on the merged FLIR [32] dataset, and multimodal detection methods: GMDETR [29], SLBAF [3], CFT [4], DEFCF [27], ICAFusion [23]. Table 1 shows the results of these existing methods and our proposed method on the FLIR dataset. Our method improves on MambaYOLO by achieving a leap in performance at the cost of a small increase in model size and a small decrease in computational efficiency. Compared to previous state-of-the-art multimodal methods, mAP_{50} improves by 2.6% and $mAP_{50:95}$ improves by 5.5%. And $mAP_{50:95}$ improves by an additional 12.8% over the MambaYOLO baseline in the IR modality. This shows that our model effectively integrates IR and RGB information and achieves a perfect balance between performance and efficiency.

M3FD. On the M3FD dataset, the following baselines still follow the setup and results of the previous work, i.e. the baselines are trained and tested on their respective modalities (IR or RGB). Specific results are shown in Table 1, where

our improved method improves mAP_{50} by 7.6% and $mAP_{50:95}$ by 8.3% compared to the unimodal MambaYOLO. Compared to the state of the art multimodal object detection methods, $mAP_{50:95}$ improves by 2.7%. This is despite being slightly lower than individual state-of-the-art methods in the mAP_{50} metric. However, we show a significant improvement in terms of high IoU AP, indicating that our model has a higher accuracy in human bounding box localisation, which is an important advantage in autonomous driving scenarios. It is also shown that the fusion of IR and RGB information in our model is also effective on the M3FD dataset.

Table 2. Ablation experiments on our PSRDET network module on the M3FD dataset.

CVSSF	**MoE**	**Data**	**Method**	P	R	mAP_{50}	$mAP_{50:95}$	FPS
		IR	MambaYOLO	83.2	70.0	77.0	51.2	153
		RGB	MambaYOLO	85.2	71.0	78.8	52.4	153
		IR+RGB	dual-stream MY	85.4	74.7	82.2	56.6	87
✓		IR+RGB	+CVSSF	**89.0**	76.5	84.3	58.6	80
✓	✓	IR+RGB	PSRDET	87.7	**80.8**	**86.4**	**60.7**	75

4.3 Ablation Experiments

Visualization. Fig. 4 shows the advantages of the PSRDET model. It is difficult for a network trained only on a single modality to accurately predict boxes based on inputs from both modalities at the same time. For example, when the image in Fig. 4(a) is fed into the baseline network trained on the RGB image, the network is unable to accurately predict the crowd of people submerged in the fog, and the object is incorrectly labelled as a human in Fig. 4(c). Meanwhile, when the corresponding infrared image in Fig. 4(a) is fed into the baseline network trained on the infrared image, the moto is missed. In addition, as can be seen in Fig. 4(b), the GTS in the dataset does not label distant persons. However, our method detects distant persons not included in the GTs.

Module Ablation study. Table 2 shows the ablation experiments of the PSRDET module, evaluating the mAP50 and mAP50:95 performance of the model with different modules and different modal input data on the M3FD [14] dataset. When we extend the backbone network of the baseline MambaYOLO to a two-stream backbone network and introduce two-modal data, both metrics are significantly improved, suggesting that the multimodal information makes a significant contribution to the detection task. After further introducing our CVSSF module, both metrics are further improved, indicating that our proposed fusion module fuses information from different sources more effectively. Finally, adding our PDR-MoE module to construct our PSRDET module achieves the best of both metrics. However, due to the continuous expansion of the network, our

method is slow compared to the baseline. However, for some existing multimodal detection models, our method is more accurate and efficient.

4.4 Runtime Analysis

We show in Table 1 the total number of learnable parameters (Model size) and frames per second (FPS) on the RTX3090. We did not use any inference optimiser. Due to the efficient fusion encoder, the parameters of PSRDET (19.5M) do not grow significantly and can be described as lightweight, while still maintaining real-time operational capabilities (FPS = 75).

5 Conclusion

In this work, we propose a new multispectral object detection algorithm PSRDET. The utilisation of the CVSSF module helps to efficiently extract and fuse features from each modality, and the application of the PDR-MoE module removes degradation of the visible image and improves the robustness of the model in harsh real-time detection environments. Experimental results on two datasets show that our proposed method outperforms existing multispectral target detection models and facilitates practical deployment.

In future work, we plan to integrate more information modalities, such as depth maps, to solve the problem that infrared images are difficult to detect icy targets such as large rocks and pits in winter snowfall and icy environments. With this multimodal fusion approach, our model will be more adaptable to automated driving tasks, thus providing stronger human safety.

Acknowledgements: This work was supported by the Educational Scientific Research Project of Hubei Provincial Institute of Higher Education (Project No. 2024XD198), the Key Project of the Joint Fund of the National Natural Science Foundation of China - Guangxi (Grant No. U22A2035), the Hubei Provincial Key R & D Project (Project No. 2024BCB010), the Open Project of the National Clinical Research Center for Acupuncture and Moxibustion (Grant No. NCRCOP2024014), and the Applied Basic Frontier Project of Wuhan Science and Technology Plan (Grant No. Wuke [2020] 25).

References

1. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
2. Chen, Z., Zhong, F., Luo, Q., Zhang, X., Zheng, Y.: EdgeViT: Efficient Visual Modeling for Edge Computing, p. 393405 (Jan 2022). https://doi.org/10.1007/978-3-031-19211-1_33, https://doi.org/10.1007/978-3-031-19211-1$_3$3
3. Cheng, X., Geng, K., Wang, Z., Wang, J., Sun, Y., Ding, P.: Slbaf-net: Super-lightweight bimodal adaptive fusion network for uav detection in low recognition environment. Multimedia Tools and Applications **82**(30), 47773–47792 (2023)

4. Fang, Q., Han, D., Wang, Z.: Cross-modality fusion transformer for multispectral object detection. SSRN Electronic Journal (Sep 2022). https://doi.org/10.2139/ssrn.4227745, http://dx.doi.org/10.2139/ssrn.4227745
5. Fang, Y., Liao, B., Wang, X., Fang, J., Qi, J., Wu, R., Niu, J., Liu, W.: You only look at one sequence: Rethinking transformer in vision through object detection. Neural Information Processing Systems,Neural Information Processing Systems (Dec 2021)
6. Girshick, R.: Fast r-cnn. In: 2015 IEEE International Conference on Computer Vision (ICCV) (Dec 2015). https://doi.org/10.1109/iccv.2015.169, http://dx.doi.org/10.1109/iccv.2015.169
7. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (Jun 2014). https://doi.org/10.1109/cvpr.2014.81, http://dx.doi.org/10.1109/cvpr.2014.81
8. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces (Dec 2023)
9. Guo, C., Li, C., Guo, J., Loy, C.C., Hou, J., Kwong, S., Cong, R.: Zero-reference deep curve estimation for low-light image enhancement. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1780–1789 (2020)
10. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics yolo: Software for object detection. https://github.com/ultralytics/ultralytics (January 2023)
11. Li, Y., Hu, J., Wen, Y., Evangelidis, G., Salahi, K., Wang, Y., Tulyakov, S., Ren, J.: Rethinking vision transformers for mobilenet size and speed (Dec 2022)
12. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection. In: 2017 IEEE International Conference on Computer Vision (ICCV) (Oct 2017). https://doi.org/10.1109/iccv.2017.324, http://dx.doi.org/10.1109/iccv.2017.324
13. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection. In: 2017 IEEE International Conference on Computer Vision (ICCV) (Oct 2017). https://doi.org/10.1109/iccv.2017.324, http://dx.doi.org/10.1109/iccv.2017.324
14. Liu, J., Fan, X., Huang, Z., Wu, G., Liu, R., Zhong, W., Luo, Z.: Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5802–5811 (2022)
15. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.: Ssd: Single shot multibox detector
16. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Liu, Y.: Vmamba: Visual state space model
17. Ma, J., Tang, L., Fan, F., Huang, J., Mei, X., Ma, Y.: Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer. IEEE/CAA Journal of Automatica Sinica **9**(7), 1200–1217 (2022)
18. Ma, J., Li, F., Wang, B.: U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722 (2024)
19. Mehta, S., Rastegari, M.: Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer
20. 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 p. 11371149 (Jun 2017). https://doi.org/10.1109/tpami.2016.2577031, http://dx.doi.org/10.1109/tpami.2016.2577031
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 p. 11371149 (Jun 2017). https://doi.org/10.1109/tpami.2016.2577031, http://dx.doi.org/10.1109/tpami.2016.2577031

22. Sharma, M., Dhanaraj, M., Karnam, S., Chachlakis, D.G., Ptucha, R., Markopoulos, P.P., Saber, E.: Yolors: Object detection in multimodal remote sensing imagery. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **14**, 1497–1508 (2020)
23. Shen, J., Chen, Y., Liu, Y., Zuo, X., Fan, H., Yang, W.: Icafusion: Iterative cross-attention guided feature fusion for multispectral object detection. Pattern Recognition **145**, 109913 (2024)
24. Tang, L., Xiang, X., Zhang, H., Gong, M., Ma, J.: Divfusion: Darkness-free infrared and visible image fusion. Information Fusion **91**, 477–493 (2023)
25. Tang, L., Yuan, J., Zhang, H., Jiang, X., Ma, J.: Piafusion: A progressive infrared and visible image fusion network based on illumination aware. Information Fusion **83**, 79–92 (2022)
26. Wang, C., He, W., Nie, Y., Guo, J., Liu, C., Han, K., Wang, Y.: Gold-yolo: Efficient object detector via gather-and-distribute mechanism (Sep 2023)
27. Wang, C., Sun, D., Yang, J., Li, Z., Gao, Q.: Dfecf-det: All-weather detector based on differential feature enhancement and cross-modal fusion with visible and infrared sensors. IEEE Sensors Journal (2023)
28. Wang, Z., Li, C., Xu, H., Zhu, X.: Mamba yolo: Ssms-based yolo for object detection. arXiv preprint arXiv:2406.05835 (2024)
29. Xiao, Y., Meng, F., Wu, Q., Xu, L., He, M., Li, H.: Gm-detr: Generalized muiltispectral detection transformer with efficient fusion encoder for visible-infrared detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5541–5549 (2024)
30. Yu, W., Wang, X.: Mambaout: Do we really need mamba for vision? arXiv preprint arXiv:2405.07992 (2024)
31. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection
32. Zhang, H., Fromont, E., Lefevre, S., Avignon, B.: Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In: 2020 IEEE International conference on image processing (ICIP). pp. 276–280. IEEE (2020)
33. Zhang, M., Yu, Y., Jin, S., Gu, L., Ling, T., Tao, X.: Vm-unet-v2: rethinking vision mamba unet for medical image segmentation. In: International Symposium on Bioinformatics Research and Applications. pp. 335–346. Springer (2024)
34. Zhang, Z., Lu, X., Cao, G., Yang, Y., Jiao, L., Liu, F.: Vit-yolo:transformer-based yolo for object detection. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW) (Oct 2021). https://doi.org/10.1109/iccvw54120.2021.00314, http://dx.doi.org/10.1109/iccvw54120.2021.00314
35. Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J.: Detrs beat yolos on real-time object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16965–16974 (2024)
36. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition (Oct 2020)