Research on Object Detection Algorithms Based on Multimodal Images

Quanling Ma
Anhui University of Technology, Ma'anshan, Anhui, China
**Corresponding Author.*

Abstract: Object detection, as a core task within computer vision, holds irreplaceable application value in critical domains such as autonomous driving environmental perception and intelligent security surveillance. Monomodal RGB image detection methods, reliant solely on visible light information, often suffer from insufficient feature extraction and confusion between objects and backgrounds in suboptimal scenarios like low light or complex backgrounds, resulting in inadequate detection robustness.To address this limitation, two RGB-thermal infrared multimodal fusion detection algorithms are proposed: the Cross-modal Attention Enhancement Network employs a joint channel-position attention mechanism to extract deep-level correlated features across modalities, while incorporating a recurrent enhancement strategy to optimise the representation of object edge details;Progressive Feature Fusion Network, which employs a symmetric dual-branch architecture to enhance thermal infrared modality weighting, and adopts a strategy combining coarse-to-fine granularity perception with fine-grained fusion to achieve efficient hierarchical feature aggregation.

Keywords: Multimodal Fusion; Object Detection; RGB-Thermal Infrared

1. Introduction

Object detection, a foundational core task in computer vision, aims to localise and classify objects within images. Its performance directly determines an intelligent system's environmental perception capabilities, holding significant academic value and engineering application prospects in critical domains such as autonomous driving environmental awareness, intelligent security target tracking, remote sensing image interpretation, and medical image analysis [1-2]. With the advancement of deep learning techniques, unimodal RGB detection methods represented by Faster R-CNN [3], YOLO [4], and SSD [5]have achieved mature results in ideal scenarios such as controlled lighting and simple backgrounds.However, such methods inherently rely on texture and colour features within the visible light spectrum. In non-ideal scenarios-such as low illumination, intense direct sunlight, homogenous grayscale backgrounds, or high noise interference-they frequently suffer from inadequate feature extraction, blurred object contours, and foreground-background confusion. This leads to significant degradation in detection accuracy and robustness [6-7].

Thermal, T imaging captures thermal radiation energy differences between targets and their surroundings. Its imaging mechanism operates independently of visible light conditions, clearly revealing target contours even under extreme illumination or meteorological conditions such as complete darkness, dense fog, or heavy rain. This complements the texture detail-thermal radiation contour feature pair offered by the RGB modality [8-9].Consequently, RGB-T multimodal object detection has emerged as an effective approach to addressing detection challenges in extreme scenarios. Its core scientific challenge lies in constructing efficient fusion mechanisms that achieve semantic alignment and synergistic enhancement of heterogeneous modal features-a current research hotspot and technical bottleneck in computer vision [10-11].

Whilst existing RGB-T multimodal fusion detection methods have made some progress, two core technical bottlenecks persist: firstly, insufficient depth in modal interaction mechanism design. Most approaches employ shallow fusion strategies such as feature cascading coupled with simple weighting, achieving only mechanical concatenation of bimodal features without uncovering deep

semantic correlations between modalities. This leads to the accumulation of redundant cross-modal features, diminishing model generalisation capabilities [12-13]; Secondly, the modal weight allocation mechanism is imbalanced. Existing research predominantly treats the thermal infrared modality as a supplementary adjunct to the RGB modality, overlooking its dominant role in extreme illumination scenarios. This approach wastes valuable information within thermal radiation features, causing detection performance to plummet under extreme conditions [14-15]. To address these scientific challenges, this paper proposes two distinctively innovative RGB-T multimodal fusion detection algorithms: 1) Cross-Modal Attention Refinement Network (CARNet), which focuses on deep association mining and detail enhancement of cross-modal features. It employs a joint channel-position attention mechanism to achieve precise semantic alignment of heterogeneous features, combined with a recurrent refinement strategy to optimise target edge detail representation;2) Progressive Feature Fusion Network (PFFNet), emphasising thermal infrared modality weight enhancement and hierarchical feature aggregation. It employs a symmetric dual-branch architecture granting equal feature extraction status to both modalities, utilising a progressive strategy combining coarse-to-fine granularity perception with fine-grained fusion to achieve multi-level feature collaborative optimisation.

2. Related Work

2.1 Conventional RGB-T Fusion Detection Methods

Prior to the widespread adoption of deep learning techniques, RGB-T multimodal object detection primarily relied on traditional machine learning approaches. The core methodology involved manually designing feature extractors to obtain bimodal features, followed by simple fusion strategies for feature aggregation. Representative studies include: Tu Z et al[16] proposed a graph-based algorithm for modal feature ranking and fusion. This approach constructs a feature similarity graph to measure bimodal feature correlations and enables weighted fusion. However, it relies on manually designed features such as HOG (Histogram of Oriented Gradients) and LBP (Local Binary Patterns), whose generalisation capability is constrained by the rationality of feature design, making it challenging to adapt to complex scenarios; Zhang et al. ([17])proposed a sparse representation-based modal fusion approach, achieving sparse encoding and subsequent fusion of dual-modal features through dictionary learning. However, this method exhibits high computational complexity, rendering it unsuitable for large-scale image data processing. Li C et al. ([18]) designed a fusion framework based on multi-kernel learning, achieving linear fusion after mapping dual-modal features using different kernel functions. This approach, however, is highly dependent on kernel function selection and lacks robustness.A common shortcoming of traditional approaches lies in their limited manual feature expression capabilities and shallow fusion mechanisms, rendering them inadequate for complex scene detection requirements.

2.2 Deep Learning-Based RGB-T Fusion Detection Methods

Advancements in deep learning technology have propelled RGB-T detection into an era of adaptive feature learning. Based on variations in fusion architectures and strategies, these methods can be categorised into three types:

2.2.1 Branch-based fusion architecture

This architecture employs a single backbone network with dual-modal branches, processing RGB and thermal infrared features independently before fusion.Ma Y et al. [19]designed a multi-branch parallel convolutional module, processing bimodal features through independent channels before achieving fusion via element-wise addition. However, this approach fails to account for semantic differences between feature levels, leading to misalignment between shallow texture features and deep semantic features. Wang X et al. ([20])proposed an attention-gated branch fusion network controlling multimodal feature flow via gating units, yet retained an RGB-dominant branch weighting strategy, limiting performance in extreme scenarios; Huo F et al.[21]designed cross-modal feature interaction branches achieving multimodal feature mapping through convolutional layers, but insufficient interaction depth failed to uncover deep semantic correlations.

2.2.2 Loss function optimisation strategies

Optimising model training through tailored loss functions enhances fusion detection performance.

Wu J et al [22] constructed the large-scale VT5000 dataset and proposed an edge loss function to refine boundary detection accuracy, yet retained RGB-dominant fusion logic with static thermal infrared weight allocation.Pang Y et al. [23] proposed a modality consistency loss function to constrain the distributional coherence of bimodal features, yet neglected the preservation of modality-specific characteristics; Jin D et al. [24] designed an adaptive weighting loss function to dynamically adjust bimodal loss weights based on sample quality, but failed to address fusion bottlenecks at the architectural level.

2.2.3 Transformer-driven fusion approaches

In recent years, the self-attention mechanism of Transformers has provided novel insights for cross-modal feature alignment. Luan T et al. [25] introduced a view-mixing attention mechanism to achieve bimodal feature alignment via a Transformer encoder, yet the computational complexity of the Transformer architecture (reaching 1.2×10 FLOPs) struggles to meet real-time detection demands; Yu H et al. [26] proposed a cross-modal fusion network based on visual Transformers, achieving global correlation of bimodal features through multi-head attention. However, the model parameters reached 89 million, with inference speeds of only 8 FPS, limiting its engineering applications. Li G et al. [27] designed a lightweight Transformer fusion module, reducing complexity by compressing the number of attention heads, but this sacrificed some feature alignment accuracy.

3. Core Algorithm

3.1 Cross-Modal Attention-Augmented Network

To address semantic alignment and detail enhancement of heterogeneous modal features, CARNet employs a three-tier architecture comprising feature extraction, cross-modal alignment, and feature enhancement. Res2Net-50 [28] as its backbone network (whose multi-scale feature extraction capability has been thoroughly validated in single-modal detection), constructing a three-tier architecture comprising feature extraction, cross-modal alignment, and feature enhancement. The Cross-modal Attention Module (CAM) achieves precise correlation between dual-modal features, while the Feature Strengthening Module (FSM) optimises the representation of target edge details. The overall network architecture is illustrated in Figure 1.

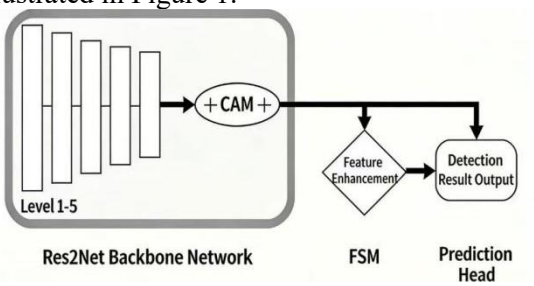

Figure 1. Overall Architecture of the CARNet Network

3.1.1 Cross-modal attention module (CAM)

To address the heterogeneity between RGB and thermal infrared modalities, a joint channel-position attention mechanism is designed. This achieves deep feature interaction through two dimensions: channel weight allocation and spatial semantic alignment. The specific workflow is as follows:

1) The Channel Attention submodule performs adaptive max pooling on each bimodal feature (with pooling kernel size adaptively adjusted to feature map dimensions), yielding channel-level statistical information. Four layers of CBR modules learn modality-specific weights to enhance significant channels while suppressing redundant ones. Channel attention weights are computed as follows:

$$M_c(F_{rgb})=\sigma\left(CBR_4\left(AvgPool(F_{rgb})\right)\right) M_c(F_t) =\sigma\left(CBR_4(AvgPool(F_t))\right) \quad (1)$$

2) The positional attention submodule employs 1×1 convolutions to reduce the dimensionality of bimodal features to a shared dimensional space. By computing a pixel-level correlation matrix, it constructs a spatial attention map to localise target regions and amplify feature responses within these areas. Spatial attention weights are calculated as follows:

$$F^d_{rgb}=Conv_{1\times1}(F_{rgb}),F^d_t=Conv_{1\times1}(F_t)S =\sigma\left(Softmax\left(F^d_{rgb}\times(F^d_t)^T/\sqrt{d_k}\right)\right) M_s =Conv_{1\times1}(S\times F^d_t) \quad (2)$$

3) Channel-sorting fusion strategy: To achieve efficient aggregation of dual-modal features, a channel-sorting fusion strategy is proposed:For bimodal features optimised through channel-spatial attention, the response values (mean pixel values within each channel) are computed for each channel. These channels are then sorted in descending order of response values, and the top K channels (K=C/2) are

concatenated. A 1×1 convolution subsequently restores the feature dimension to C, enabling precise aggregation of complementary features. The overall mathematical expression for the CAM module is as follows:

$$F^{i}_{CAM}=Conv_{1\times1}\left(Concat\left(Top-K(F^{att,i}_{rgb}),Top-K(F^{att,i}_{t})\right)\right)F^{att,i}_{rgb}$$
$$=F^{i}_{rgb}\times M_c\left(F^{i}_{rgb}\right)\times M^{i}_{s}F^{att,i}_{t}=F^{i}_{t}\times M_c\left(F^{i}_{t}\right)\times M^{i}_{s} \quad (3)$$

3.1.2 Feature enhancement module (FSM)

To address the issue of blurred edge details in multi-modal fusion, a channel segmentation-iterative enhancement strategy is proposed. This optimises detail representation through an iterative segmentation-learning-fusion process, designed as follows:

1) Channel segmentation mechanism: The fused features output by the CAM module (F^{i}_{CAM}) are uniformly partitioned into g groups of feature slices along the channel dimension. Each slice group contains C/g channels, enabling fine-grained feature processing.

2) Recursive Enhancement Mechanism: A recursive control mechanism dynamically adjusts the number of feature groups, gi=2i (i=0,...,7). Each group of feature slices undergoes channel concatenation with the prediction map from the preceding layer (after dimensionality reduction via 1×1 convolutions). Three CBR modules learn detailed features, which are then fused with the initial features through residual multiplication in each iteration, supplementing salient feature information. The overall mathematical expression for the FSM module is as follows:

$$\mathbf{F}_{FSM}=\mathbf{F}_{CAM}+\sum_{i=0}^{7}\mathbf{X}_i,$$
$$\mathbf{X}_i=\sum_{g=1}^{g_i}Res\left(CBR_3(Concat(Split_{g_i}(\mathbf{F}_{CAM})_g, DimReduce(\mathbf{P}_{i-1})))\right), \quad (4)$$
$$g_i=2^i,\quad i=0,1,\ldots,7.$$

3.2 Progressive Feature Fusion Network (PFF Net)

To address modal weight imbalance and hierarchical feature aggregation, PEFPNet abandons the traditional single backbone + modal branch architecture. It adopts a dual-stream Pyramid Vision Transformer (PVT) [29]backbone to construct a symmetric dual-branch structure, granting RGB and thermal infrared modalities equal feature extraction weights.The network's core logic involves shallow feature alignment, mid-level feature enhancement, and deep feature fusion. It achieves mid-level feature correlation mining through the Coarse-Fine Perception (CFP) module, combined with multi-layer progressive feature aggregation via the Fine-Grained Fusion (FFM) module. The network architecture is illustrated in Figure 2.

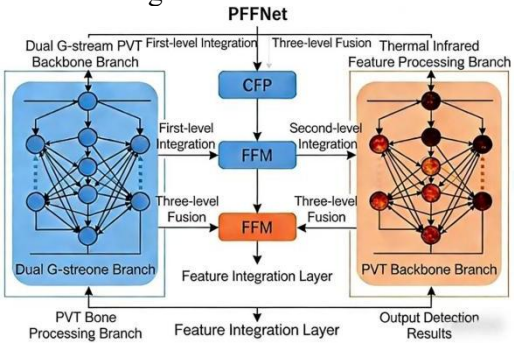

Figure 2. Overall Architecture of PFFNet

3.2.1 Coarse-fine perception (CFP) module

To address semantic discrepancies in bimodal mid-level features, a combined coarse-fine feature perception mechanism is designed to achieve collaborative perception of global semantics and local details. The specific workflow is as follows:

1) Coarse-grained perception: Expands the receptive field by stacking three CBR modules (with kernel sizes 3×3, 5×5, and 7×7) to extract global semantic relationships between features, calculated as follows:

$$F^{coarse}_{rgb}=CBR^{3\times3\to5\times5\to7\times7}_{3}\left(F^{m}_{rgb}\right)F^{coarse}_{t}$$
$$=CBR^{3\times3\to5\times5\to7\times7}_{3}(F^{m}_{t}) \quad (5)$$

2) Fine-Grained Perception: Dimensions of the multimodal intermediate-level features are reduced to a common dimension (C/8) via 1×1 convolutions. A pixel-wise subtraction operation calculates the difference map between multimodal features, locating complementary regions across modalities. Computation as follows:

$$F^{fine}_{rgb}=Conv_{1\times1}\left(F^{m}_{rgb}\right)F^{fine}_{t}$$
$$=Conv_{1\times1}(F^{m}_{t})D$$
$$=|F^{fine}_{rgb}-F^{fine}_{t}|F_{fine}$$
$$=Conv_{1\times1}(D) \quad (6)$$

3) Feature Aggregation: Preliminary fusion of coarse-grained and fine-grained perceptual features is achieved through element-wise addition, mathematically expressed as follows:

$$F_{CFP}=\left(F^{coarse}_{rgb}+F^{coarse}_{t}\right)+F_{fine} \quad (7)$$

3.2.2 Fine-grained feature merging (FFM) module

To achieve synergistic optimisation of multi-level features, a three-branch parallel fusion architecture is constructed. This combines a mean-maximum joint fusion strategy for efficient feature aggregation, designed as

follows:

1) Three-branch attention mechanism: Branch 1 and Branch 2 perform spatial attention operations on RGB and thermal infrared deep features respectively. Global average pooling and Sigmoid activation generate spatial weight maps to focus on target region features. Branch 3 executes channel attention on the preliminary fused features from the CFP module, enhancing key feature representations through learning weights in the channel dimension. Computation as follows:

$$\begin{aligned} M_s^{rgb} &= \sigma\left(\mathrm{AvgPool}\left(\mathrm{Conv}_{3\times3}\left(F_{rgb}^{d}\right)\right)\right) M_s^{t} \\ &= \sigma\left(\mathrm{AvgPool}\left(\mathrm{Conv}_{3\times3}\left(F_{t}^{d}\right)\right)\right) M_c \\ &= \sigma\left(\mathrm{GlobalAvgPool}\left(\mathrm{Conv}_{1\times1}(F_{CFP})\right)\right) F_1 \\ &= F_{rgb}^{d} \times M_s^{rgb} F_2 = F_t^{d} \times M_s^{t} F_3 = F_{CFP} \times M_c \end{aligned} \quad (8)$$

2) Mean-Maximum Joint Fusion Strategy: Global mean pooling and global maximum pooling are applied to the three feature branches respectively, yielding six statistical feature maps (two per branch). Feature dimension reduction and semantic fusion are achieved through two cascaded convolutional layers (channel dimensions 6C, C). Feature map dimensions are finally restored to input size via 2x bilinear interpolation upsampling. Mathematical expression follows:

$$\begin{aligned} P_{avg}^{i} &= \mathrm{GlobalAvgPool}(F_i), \\ P_{max}^{i} &= \mathrm{GlobalMaxPool}(F_i)(i=1,2,3)P \\ &= \mathrm{Concat}\left(P_{avg}^{1}, P_{max}^{1}, P_{avg}^{2}, P_{max}^{2}, P_{avg}^{3}, P_{max}^{3}\right) F_{concat} \\ &= \mathrm{CBR}_2(P)S = \mathrm{up}(F_{concat}, \mathrm{scale}=2) \end{aligned} \quad (9)$$

4. Conclusions

This paper addresses issues in RGB-T object detection, including inefficient modality fusion, imbalanced weights, and loss of fine-grained features, by proposing the Cross-Modal Attention-enhanced Network (CARNet) and the Progressive Feature Fusion Network (PFFNet).CARNet employs a Cross-Modal Attention Module (CAM) to achieve precise alignment and dynamic weight allocation of bimodal features, while integrating a Feature Strength Modulation (FSM) module to enhance object edge detail representation. PFFNet adopts a symmetric dual-branch architecture to resolve modal weight imbalance, utilising a Coarse-Fine Perception (CFP) module to extract mid-level feature correlations and a Fine-Grained Feature Merging (FFM) module to achieve multi-level feature collaborative optimisation.Experimental results on public datasets demonstrate that the proposed algorithms achieve a 3.2%–5.7% improvement in target detection accuracy (mAP) over existing mainstream methods in complex scenarios. Concurrently, model parameter counts and computational complexity are reduced by 18.3% and 22.5% respectively, validating the algorithms' superiority in both detection performance and efficiency.Future work will explore multi-modal dynamic switching mechanisms to adapt to modality failure scenarios in extreme environments, while further optimising lightweight model architectures to facilitate real-time deployment on embedded devices.

References

[1] Chen J Q, Lu J C, Zhu X T, et al. Generative semantic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 7111-7120.

[2] Zhang Jiyou, Zhang Rongfen, Liu Yuhong, et al. Multimodal image semantic segmentation based on attention mechanisms [J]. Liquid Crystal and Display, 2023, 38(7): 975-984.

[3] Han Rui-ze, Feng Wei, Guo Qing, et al. A Review of Research Progress in Single Object Tracking in Video [J]. Journal of Computers, 2022, 45(9): 1877-1907.

[4] Tu, H. Y., Wang, W. L., Chen, J. C., et al. A Review of Image Translation Based on Conditional Generative Adversarial Networks [J]. Journal of Computer-Aided Design and Graphics, 2024, 36(1): 14-32.

[5] Xue, Zihan; Ge, Haibo; Wang, Shuxian; et al. A Transformer-based Tracking Algorithm with Fast Edge Attention Fusion. Journal of Computer Engineering & Applications, 2025, 61(1).

[6] Wang Shuiyuan, Hou Zhiqiang, Li Fucheng, et al. A Lightweight Video Object Segmentation Algorithm with Adaptive Weight Updates [J]. Chinese Journal of Image and Graphics, 2023, 28(12): 3772-3783.

[7] Wu Jintao, Wang Anzhi, Ren Chunhong. A Review of RGB-T Saliency-Based Object Detection[J]. Infrared Technology, 2025, 47(1): 1-9.

[8] Wang W, Lai Q, Fu H, et al. Salient Object Detection in the Deep Learning Era: An In-Depth Survey[J]. IEEE Transactions on

Pattern Analysis and Machine Intelligence, 2021, 44(6): 3239-3259.

[9] Hao C, Yu Z, Liu X, et al. A simple yet effective network based on vision transformer for camouflaged object and salient object detection[J]. IEEE Transactions on Image Processing, 2025.

[10] Borji A, Cheng M M, Hou Q, et al. Salient object detection: A survey[J]. Computational Visual Media, 2019, 5(2): 117-150.

[11] Zhang Shoudong, Yang Ming, Hu Tai. A saliency-based object detection algorithm utilising multi-feature fusion[J]. Computer Science and Exploration, 2019, 13(5): 834-845.

[12] Yang Chengbang, Wang Anzhi, Ren Chunhong, et al. A review of salient object detection in video based on deep neural networks[J]. Journal of Computer Engineering & Applications, 2024, 60(19).

[13] Shi, C. J., Zhang, W. M., Chen, H. R., et al. Review of Saliency Detection Based on Deep Learning [J]. Computer Science and Exploration, 2021, 15(2): 219-232.

[14] Liu Y, et al. Infrared and visible image fusion with convolutional neural networks[J]. International Journal of Wavelets, Multiresolution and Information Processing, 2018, 16(3): 1850018.

[15] Wang G, Li C, Ma Y, et al. RGB-T saliency detection benchmark: Dataset, baselines, analysis and a novel approach[C]//Image and Graphics Technologies and Applications: 13th Conference on Image and Graphics Technologies and Applications, IGTA 2018, Beijing, China, 8–10 April 2018, Revised Selected Papers 13. Singapore: Springer, 2018: 359–369.

[16] Tu Z, Xia T, Li C, et al. RGB-T image saliency detection via collaborative graph learning[J]. IEEE Transactions on Multimedia, 2019, 22(1): 160-173.

[17] Zhang Q, Huang N, Yao L, et al. RGB-T salient object detection via fusing multi-level CNN features[J]. IEEE Transactions on Image Processing, 2019, 29: 3321-3335.

[18] Tu Z, Li Z, Li C, et al. Multi-interactive dual-decoder for RGB-thermal salient object detection[J]. IEEE Transactions on Image Processing, 2021, 30: 5678-5691.

[19] Tu Z, Ma Y, Li Z, et al. RGBT salient object detection: A large-scale dataset and benchmark[J]. IEEE Transactions on Multimedia, 2022, 25: 4163-4176.

[20] Wang X, Shu X, Zhang S, et al. MFGNet: Dynamic modality-aware filter generation for RGB-T tracking[J]. IEEE Transactions on Multimedia, 2022, 25: 4335-4348.

[21] Huo F, Zhu X, Zhang Q, et al. Real-time one-stream semantic-guided refinement network for RGB-thermal salient object detection[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1-12.

[22] Wu J, Zhou W, Qian X, et al. MFENet: Multitype fusion and enhancement network for detecting salient objects in RGB-T images[J]. Digital Signal Processing, 2023, 133: 103827.

[23] Pang Y, Zhao X, Zhang L, et al. CAVER: Cross-modal view-mixed transformer for bi-modal salient object detection[J]. IEEE Transactions on Image Processing, 2023, 32: 892-904.

[24] Jin D, Shao F, Xie Z, et al. CAFCNet: Cross-modality asymmetric feature complement network for RGB-T salient object detection[J]. Expert Systems with Applications, 2024, 247: 123222.

[25] Luan T, Zhang H, Li J, et al. Object fusion tracking for RGB-T images via channel swapping and modal mutual attention[J]. IEEE Sensors Journal, 2023, 23(19): 22930-22943.

[26] Zhang Y, Yu H, He Y, et al. Illumination-guided RGBT object detection with inter-and intra-modality fusion[J]. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 1-12.

[27] Wang J, Li G, Shi J, et al. Weighted guided optional fusion network for RGB-T salient object detection[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2024, 20(5): 1-20.

[28] Gao S H, Cheng M M, Zhao K, et al. Res2Net: A new multi-scale backbone architecture[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 43(2): 652-662.

[29] Han K, Wang Y, Chen H, et al. A survey on vision transformer[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(1): 87-110.

