Complex & Intelligent Systems
https://doi.org/10.1007/s40747-024-01409-z

ORIGINAL ARTICLE

Combining transformer global and local feature extraction for object detection

Tianping Li[1] · Zhenyi Zhang[1] · Mengdi Zhu[1] · Zhaotong Cui[1] · Dongmei Wei[1]

Received: 13 December 2023 / Accepted: 9 March 2024

Abstract

Convolutional neural network (CNN)-based object detectors perform excellently but lack global feature extraction and cannot establish global dependencies between object pixels. Although the Transformer is able to compensate for this, it does not incorporate the advantages of convolution, which results in insufficient information being obtained about the details of local features, as well as slow speed and large computational parameters. In addition, Feature Pyramid Network (FPN) lacks information interaction across layers, which can reduce the acquisition of feature context information. To solve the above problems, this paper proposes a CNN-based anchor-free object detector that combines transformer global and local feature extraction (GLFT) to enhance the extraction of semantic information from images. First, the segmented channel extraction feature attention (SCEFA) module was designed to improve the extraction of local multiscale channel features from the model and enhance the discrimination of pixels in the object region. Second, the aggregated feature hybrid transformer (AFHTrans) module combined with convolution is designed to enhance the extraction of global and local feature information from the model and to establish the dependency of the pixels of distant objects. This approach compensates for the shortcomings of the FPN by means of multilayer information aggregation transmission. Compared with a transformer, these methods have obvious advantages. Finally, the feature extraction head (FE-Head) was designed to extract full-text information based on the features of different tasks. An accuracy of 47.0% and 82.76% was achieved on the COCO2017 and PASCAL VOC2007 + 2012 datasets, respectively, and the experimental results validate the effectiveness of our method.

Keywords Object detection · Attention mechanism · Transformer · Anchor-free · Detector head

Introduction

Object detection [1] is a fundamental vision task in computer vision that focuses on categorizing objects for recognition and location localization. It has been widely used in the fields of automatic driving, security monitoring, and medical imaging [2]. Recent research on high-performance object detectors [3] has not stopped and still faces great challenges. For example, the problem of optimizing the number of parameters, the speed, and the convergence rate while maintaining high detection accuracy can be addressed. For this reason, much work has been done by researchers studying high-performance object detection algorithms.

In recent years, CNN-based [4] and Transformer-based [5] object detectors have become the current dominant architectures, and models from both architectures have achieved stunning detection results. CNN-based object detectors have evolved from original two-stage detectors (e.g., R-CNN [6] and Cascade R-CNN [7]) to single-stage detectors (e.g., YOLO [8] and RetinaNet [9]) and from anchor-based detectors (e.g., Faster R-CNN [10] and YOLOV4 [11]) to anchor-free detectors (e.g., FCOS [12] and ATSS [13]). Among them, the emergence of anchor-free detectors has led to a further increase in the speed of detection. The above CNN-based

✉ Dongmei Wei
weidongmei@sdnu.edu.cn

Tianping Li
sdsdltp@sdnu.edu.cn

Zhenyi Zhang
zzy0693@163.com

Mengdi Zhu
1342785693@qq.com

Zhaotong Cui
haotong120@163.com

[1] School of Physics and Electronics, Shandong Normal University, Jinan, Shandong, China

Published online: 15 April 2024

object detection algorithms not only have made important breakthroughs in terms of efficiency and accuracy but also have obvious advantages in extracting local and multiscale features. However, the drawback of insufficient ability to extract information globally is still prevalent, and the detection performance still somewhat deviates from that of current object detectors.

Since its introduction into the field, the transformer architecture has become a research hotspot in the field of computer vision [14]. Therefore, researchers started to introduce the Transformer to object detection and proposed a series of Transformer-based object detection algorithms. For the first time, DETR [15] used a transformer for image object detection via a tandem splicing approach involving a CNN followed by a transformer, and the 2D features extracted by the CNN were subsequently input into the transformer, which enhanced the extraction of global features. The disadvantages of these methods include poor detection of small objects, large computational parameters and slow convergence. Deformable DETR [16] fuses the sparse sampling capability of variability convolution with the powerful relational modelling capability of the transformer using a small set of sampling locations as a feature image element that highlights where the key elements are located, resulting in enhanced training effects and small object detection. The disadvantage is that the complexity and number of parameters are large, a large amount of data is needed for training, and the ability to detect small objects needs to be strengthened. Transformers [17] have received a great deal of academic attention due to their simple encoder-decoder architecture and remarkable detection results. The architecture simplifies the detection process and has a unified paradigm, and the ability to capture remote dependencies in captured objects and perform global feature extraction with robust image semantic information makes it easy to achieve highly accurate end-to-end object detection. However, there are also shortcomings in its weak ability to extract multiscale and local features. Although researchers have recently achieved good results in transformer optimization [18], this approach still suffers from the disadvantages of a large number of computational parameters, slow convergence, and a lack of ability to acquire local information (edges and texture) compared to CNN-based detection models, and it is difficult to strike a balance between the detection accuracy and the number of parameters. This shows that although the computational process is greatly simplified, the computational cost of the Transformer itself is still large, and the number of parameters is difficult to reduce. For abbreviations and full names used in this paper, please consult the glossary of terms shown in Table 1.

Table 1 Abbreviations and full names

Abbreviations	Indexes
CNN	Convolutional Neural Network
FPN	Feature Pyramid Network
SCEFA	Segmented Channel Extracted Feature Attention
AFHTrans	Aggregated Feature Hybrid Transformer
FE-Head	Feature Extraction Head
GLFT	Combining Transformer Global and Local Feature Extraction
MHA	Multi-Head Attention
MSDA	Multi-Scale Dilated Attention
FFN	Feedforward Neural Network
FE-Head	Feature Extraction Head
MFCEM	Multiscale Full Text Channel Extraction Module
GT	Global Text
AGAP	Adaptive Global Average Pooling
AGMP	Adaptive Max Average Pooling
PE	Position Encoding
FE	Feature Extractor

The Microsoft team's Mobile-Former [19] parallelizes MobileNet [20] and Transformer using a bidirectional cross-bridging approach, which achieves bidirectional fusion of local and global features and reduces the number of random tokens generated. The disadvantage is that the presence of multiple convolutions makes the patch unstable, often leading to high computational effort and performance degradation. UniFormer [21] uses convolution and transformers to extract global and local features, effectively solving redundancy and dependency problems in the model learning process. TransXNet [22] builds a novel hybrid CNN-Transformer architecture to aggregate global information and local details with excellent detection performance. Influenced by Mobile-Former, UniFormer and TransXNet, a combination of a CNN and a transformer is considered for the above problem to enhance the feature extraction capability of local and global information. Moreover, the fast speed of the anchor-free end-to-end architecture and the absence of nonextremely large value suppression are utilized for the comprehensive design of the detector. The current mainstream CNN-based anchor-free object detector is composed of three main parts: the backbone, neck, and head. The backbone network here is ResNet, and no changes are usually made to that structure. To better realize the above, this paper rethinks the structure of the Transformer. From this, it is found that the encoder-decoder structure in the Transformer, although favorable for extracting multiscale features and improving detection performance, leads to the problems of an excessive number of computational parameters and slow convergence. Moreover, TSP-FCOS [23] achieves accelerated training and an improved ability to extract multiscale features by combining FCOSs and removing the structure of the transformer decoder; however, the detection performance needs to be further improved. To reduce the number of unnecessary calculations and parameter values as well as improve the detection performance, the key components are analysed in detail, and experiments are carried out based on them. Therefore, the multi-head cross-attention part of the decoder was removed. Additionally, the ability of the Transformer to extract localized and multiscale features that result from removing this part should be mitigated. Inspired by the ability of EPSANet [24] to segment the spatial information of extracted feature maps at different scales and establish long-term dependencies between local multiscale channel attention mechanisms, the multiscale full-text channel extraction module (MFCEM) is proposed to enhance the ability to extract multiscale features. The highest layer in the output part of the backbone network is rich in local semantic information, but the attention mechanism has difficulty balancing the number of channels and parameters at this layer, resulting in many detectors being unable to extract sufficient feature information from this layer. Inspired by the fact that the feature grouping and channel substitution operations of SANet [25] can optimize the interaction of parametric quantities and channel information, the segmented channel extracted feature attention (SCEFA) method combined with MFCEM is proposed to solve the above problem.

To further improve the encoder structure of the Transformer, inspired by the ability of the Dilateformer [26] to efficiently aggregate multiscale semantic information from different receptive fields and to reduce structural redundancy, multi-scale dilation attention (MSDA) is introduced. Moreover, the convolutional and depth-separated convolutional structures of CNNs are added to enhance local information acquisition. Then, the core parts of the position encoding (PE), multi-head attention (MHA), and feed-forward neural network (FFN) structures in the transformer are extracted, and the front and back parts are combined to improve the transformer. FPNs [27] are mostly used in necking structures to address the challenges of multiscale detection but have long been characterized by a significant drawback: the inability to adequately communicate and fuse information across layers (e.g., Layer 1 and Layer 3) limits the scope of information fusion. For the cross-level information fusion approach, inspired by TopFormer's [28] powerful hierarchical feature structure for acquiring intensive prediction tasks, extensions have been developed based on its theory. By combining the above self-attention improved architecture for transformers, the aggregate feature hybrid transformer (AFHTrans) architecture incorporating convolution is designed to be added to the front end of the retained FPN. Effective exchange of information across hierarchical levels is realized by fusing global multilevel information injected into rich semantic information at a high level. The above improvements further enhance the global and local feature extraction of neck information, improve the multiscale detection performance of the model, and avoid causing an increase in the number of multiple parameters. Finally, to improve the information injection capability, an efficient injection module for detection information is designed to realize the effective fusion of the original information into high-level information.

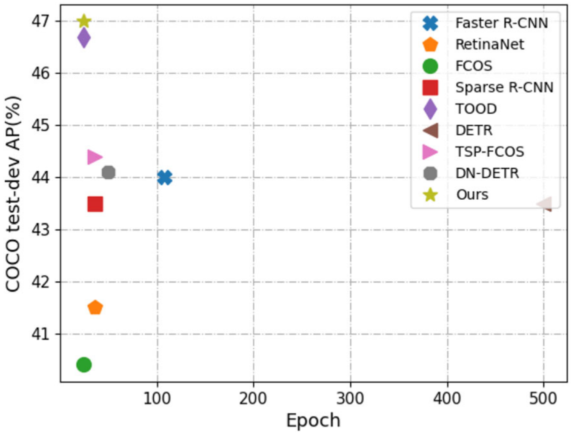


Fig. 1 Comparison of the epochs and APs of our network with those of other networks on the COCO2017 dataset

TOOD [29] proposes a task-aligned detection head that improves detection performance by aligning classification and localization tasks in a learning-based manner. However, this detection head lacks the ability to perform feature extraction on the fused full-text information. For this purpose, an anchor-free feature extraction head (FE-Head) is designed in which the fused global and local feature map information is further extracted and processed, optimizing the feature extraction capability of full-text information for different tasks at the decoupled head.

By combining the above improvements, in this paper, we design a CNN-based end-to-end anchor-free object detector that combines the Global and Local Feature Extraction (GLFT) of Transformers. The model not only enhances the global and local feature information extraction of the whole model feature map but also realizes the cross-channel information fusion of feature maps at different levels. GLFTNet outperforms other state-of-the-art detection networks in terms of the number and accuracy of computational parameters, achieving significant advantages. Comparison with other classical networks using the backbone network ResNet101. As shown in Fig. 1, our network converges quickly, and its detection accuracy is better than that of other classical networks.

In summary, the main contributions of this paper are as follows:

1. From a microscopic point of view, this approach can improve the feature extraction of local semantic information and achieve a balance between the number of channels and the number of parameters. To this end, this paper proposes the segmented channel extracted feature attention (SCEFA) module combined with the multiscale full-text channel extraction module (MFCEM), which can be used to efficiently extract high-level localized channel feature map information and establish a remote dependency on the number of long-distance channels, effectively reducing the number of parameters. The SCEFA is flexible and easy to use and can be used in many computer vision network models.
2. The ability of the Transformer to extract localized information is limited by the large number of parameters and slow convergence. Moreover, considering the FPN from a macro perspective, it lacks the ability to interact and merge information across hierarchical levels. To this end, this paper proposes an aggregate feature hybrid transformer (AFHTrans) module combined with convolution and an efficient injection module for detecting information, which not only improves the global and local feature extraction capability of the model and the cross-fertilization of feature information across hierarchical levels but also significantly reduces the number of computational parameters.
3. The feature extraction head (FE-Head) was designed to address the lack of extraction of the full-text information of the feature map by the detection head. It not only has the powerful ability to acquire full-text feature information but also balances the number of parameters and accuracy well. It can be applied to a variety of single-stage object detectors in a plug-and-play manner with significant results.
4. By combining the SCEFA, AFHTrans, and FE-Head methods, a novel GLFT object detection model incorporating a transformer is designed to improve global and local feature extraction. Because of the smaller number of parameters and faster training speed compared to those of the transformer detector, the proposed architecture allows this paper to achieve superior performance in benchmark model comparison tests, outperforming most state-of-the-art object detectors.

The remainder of this paper is organized as follows. First, the “Related work” of this paper is described in detail. Then, the “Methods” proposed in this paper are described. The experimental work is also described in the “Experiments” section of the paper. Finally, the “Conclusion” section provides a comprehensive summary of the work presented throughout the text.

Table 2 Work related to the “decoder–encoder” style of feature extraction of contextual information and content characterization

Related work	Content features
Faster R-CNN [10]	This work uses a convolutional layer network to extract features and a Region Proposal Network to efficiently extract candidate region context information, which, combined with the shared feature property of convolution, results in a significant improvement in detection performance. However, whether using VGGNet or ResNet, the extracted feature maps are single-layer and the resolution is relatively small, which leads to poor detection of multiscale and small objects, generating leakage and false detection
Mask R-CNN [30]	This work extracts mask information from the finer spatial layout of an object by means of pixel-to-pixel correspondences in the feature map produced by deep convolution
CA-CNN [31]	This work, to enrich the feature representation and mine the underlying contextual information between objects, integrates the feature information into a convolutional layer and maps it into a multilevel feature map to extract the contextual information, and finally fuses the multilevel feature information into a single layer. The disadvantage is that candidate regions and fusing feature information from different layers can lead to model redundancy and slow speed
PFPNet [32]	This work uses a multiscale semantic fusion module to feature-fuse high-level to shallow-level information so that there is both shallow-level object position information and high-level object semantic information

Related work

In this section, previous related work is reviewed, such as feature extraction of contextual information, transformers combined with convolution, and decoupled heads.

Feature extraction of contextual information

Feature extraction of contextual information has been widely studied by researchers as a fundamental step in many image visualizations, especially in the field of deep learning. In general, studies can be divided into two categories. The first is the “decoder–encoder” style. In Table 2, different tasks and specific content features of the “decoder–encoder” style of feature extraction of contextual information are described in detail.

Table 3 Features extract contextual information related to the “backbone” style of work as well as content features

Related work	Content features
SSD [33]	This work achieves multiscale object detection by using a convolutional layer to extract different feature maps after the backbone network, thus making full use of the sensory field information of different-sized feature maps and activating the semantic and contextual information of the object. However, there is also the disadvantage of a low detection rate for small objects
RFBNet [34]	This work proposes an Inception-like receptive field module, which mainly utilizes multibranch convolutional layers to incorporate receptive field structures into the backbone network, effectively expanding the channel receptive field and enhancing the feature extraction capability
CBNet [35]	This work presents novel backbone networks that do not require retraining of weights. By compositely connecting and integrating multiple identical backbone networks with high-level and low-level features, the scope of feature information fusion and sensory field is expanded, thus aiding the performance of high-performance detectors

Table 4 Related work on two cases of feature extraction of contextual information and content characterization

Related work	Content features
The first scenario	
CornerNet [36]	This work proposes the concept of keypoints, which enables the model to better understand the feature information of the object. Its disadvantage is that when performing object detection, it only focuses on the features of the two corners of the object, the upper-left and lower-right corners, and ignores the center features of the object, which may lead to the generation of some false positive samples. Especially when dealing with objects that cover each other, many objects do not have obvious corner features, which makes them perform poorly in the detection of covered objects
The second scenario	
PANNet [37]	This work utilizes bottom-up path augmentation and adaptive feature pooling to aggregate features between different layers, thus ensuring feature completeness and diversity. However, this also makes the network structure complex and requires more computational resources for training and inference. Moreover, some difficulties may be encountered when dealing with the small object task, as the localization information of low-resolution feature maps may be lost

The second category is the “backbone” style. As shown in Table 3, different works and specific content features of the “backbone” style of feature extraction of contextual information are detailed.

For the contextual information extracted by the features, two scenarios exist. The first scenario. When the perceptual field is too small and the object is too large, leakage may occur because the model does not sense the presence of the object. The second scenario. When the sensing range is too large and the object size is small, the network extracts background and redundant information, which leads to false detections because the model has difficulty observing tiny objects. Table 4 lists the different works and specific content features of the above two cases of feature extraction of contextual information.

Overall, there are two drawbacks to these two types of styles and scenarios. First, the affinity matrix for detecting pixel information is compared with other pixel information and lacks self-pixel information. Second, several methods are designed with redundant design operations, which results in computationally intensive and ineffective establishment of long-range channel number dependencies for localized information. To avoid the above drawbacks and to fully extract the global and local feature information of the object, three methods are proposed in this paper: the SCEFA for local information extraction on the high-level network, the AFHTrans for global and local information extraction and establishing the pixel relationship of the long-distance object, and the FE-Head for global feature information extraction on the head.

Transformer combined with convolution

The local perceptual field and shared weights provided by the CNN can be used to effectively capture the local feature information in the image and can effectively reduce the number of computational parameters. The high-dimensional feature representation of a CNN converts the information in an image into high-dimensional semantic information for better feature extraction. Moreover, CNNs are translation invariant, enabling enhanced generalization of the model. However, CNNs lack the ability to extract global features and cannot effectively capture global features or establish long-range pixel dependencies for objects. In addition, the Transformer has powerful context-aware scaling and the ability to extract global feature information but still suffers from two drawbacks: first, it is challenging in terms of training and data, and second, it occupies a large number of computational parameters. To solve the above problems, many researchers have started to combine the advantages of CNNs and transformers

Table 5 Combined related work on convolution and converters as well as content characterization

Related work	Content features
Conformer [38]	This work enhances the global and local feature learning capability by combining CNN with Vision Transformer in a hybrid network structure, which focuses on the fusion of global and local features through feature coupling units
CMT [39]	This work uses a parallel design of MobileNet and Transformer, where deep convolution is used to compress the channels to extract local information and the converter establishes dependencies between patches for global features
CoTNet [40]	This work creatively integrates the static context information aggregation of convolution with the dynamic context information aggregation of the self-attention mechanism in Transformer, which is able to improve the performance of the network without significantly increasing the computational effort
ViT-Det [41]	This work parallelizes CNN and Transformer in the shallow part of the backbone, combining the advantages of CNN and its variants in extracting high-frequency information and Transformer in extracting low-frequency information, which can enhance the ability to fuse the information and improve the accuracy of detection. However, there are the shortcomings of unnecessary structure and the inability to combine FPN, which limit the scope of the application
SMT [42]	This work performs image global and local feature extraction by combining convolutional networks with visual transformers to efficiently capture long-range dependencies
CloFormer [43]	This work rethinks the implementation of local awareness at Transformer, which enhances access to high-frequency local information through the relationship between the global shared weights of the primitive convolutional operator and the context of a particular token in self-attention

and have performed much work. As shown in Table 5, the different works and specific content features performed by the combination of the CNN and transformer are detailed.

Drawing on the merits of the above work, the AFHTrans module proposed in this paper incorporates convolution and deep convolution, MHA, the MSDA, and the FFN as core components. Convolution and deep convolution are used in front of the channel to extract local features; in the middle, MHA, MSDA, and FFN are used to expand the receptive field while extracting global information; and in the back, detection information is efficiently injected into the middle and high levels of the FPN to fuse the global and local feature information. This module effectively combines the advantages of CNNs and Transformers to enhance the global and local feature extraction capabilities, establishes the long-range interdependence of object pixels, optimizes the number of computational parameters, and achieves comprehensive and accurate detection of object information. This module effectively combines the advantages of CNNs and Transformers to enhance the global and local feature extraction capabilities, establishes the long-range interdependence of object pixels, optimizes the number of computational parameters, and achieves comprehensive and accurate detection of object information.

Decoupled head

Decoupled heads have recently become the standard structure for mainstream object detection. There have been tremendous breakthroughs in the research on decoupled heads. As shown in Table 6, the different recent works performed on decoupled heads and the content characteristics are detailed.

The above work illustrates the importance of decoupling and task characterization information processing in the head. To this end, the FE-Head method using a decoupled structure is proposed to be able to comprehensively improve the acquisition of full-text information from the head feature map, enabling the decoupled head to fully utilize the extracted feature information for classification and localization tasks.

Methods

In this study, we propose a GLFTNet-based detection model that combines the SCEFA module, the AFHTrans module, and the FE-Head. First, the paper describes GLFTNet in "GLFTNet"; then, the SCEFA module, the AFHTrans module, and the FE-Head are presented in "Segmented channel extraction feature attention module", "Feature grouping", and "Segmentation channels for extracting multiscale features", respectively.

GLFTNet

As shown in Fig. 2, this paper briefly discusses global and local feature extraction for object image detection (GLFT) in conjunction with the transformer.

First, information about the features of the image is extracted using the backbone network ResNet50 or ResNet101 to generate different layers of multiscale feature image pyramids. Next, the high-level output portion of the pyramid uses the SCEFA module to identify the importance information of the feature map. First, the SCEFA module is used to extend the channel dimension split grouping after segmentation and extraction, and the channel information and spatial information of the resulting multiscale channel feature maps are integrated into the group feature maps to obtain the global and local channel information interaction attention in the local features to adaptively differentiate the channels. The aggregated output feature maps are then subjected to channel feature conditioning operations. Finally, channel replacement operations (channel shuffling) are used for intergroup communication to output information that enriches local features. Second, the output feature maps with the same number of channels after three layers of channel compression are aggregated using the AFHTrans module, after which the spatial information about the feature maps is extracted and the global and local pixel features are augmented by the similarity of the object pixel points to establish object pixel dependencies over long distances of the model and capture the semantic dependency information of the channels at different scales. Finally, all the processed information is pooled into the second and third layers using a detection-efficient information injection module, and the full-layer information is passed horizontally into the top-down FPN for feature fusion. Finally, the fused information is passed into the FE-Head for full-text feature information extraction in the classification and regression branches.

Table 6 Related work on decoupling headers and content characterization

Related work	Content features
IoU-Net [44]	This work first introduces an additional header to predict the IOU as the localization confidence and then aggregates the localization and classification confidence into a final classification score. This approach reduces the problem of detecting mislocalization and improves the confidence scores for good-quality bounding boxes, but makes the confidence scores for rough bounding boxes lower
FCOS [12]	This work then extends the centroid branching task of the original header to lead to regression branching for effective bounding box localization under the centroid, regression, and classification tasks
Double-Head [45]	This work proposes to branch the R-CNN head to handle the classification and localization tasks separately. Although good detection results can be achieved, the shared parameters of the two tasks are reduced, and there are task conflicts
TSD [46]	This work found that the spatial deviation between the classification and localization tasks affects the training process, so it proposed a task-aware spatial separation method to divide the two tasks into separate decoupled heads, which achieved good training results
YOLOX [47]	This work introduces decoupled detection heads to the YOLO family for the first time, enhancing the performance and convergence speed of detection
RA-YOLOX [48]	This work proposes a lightweight decoupling head, while the detection head employs a structural reparameterization method to achieve significant performance improvement

Segmented channel extraction feature attention module

According to previous research [33, 49–52], the detection ability of images of objects of different sizes can be effectively improved by extracting multiscale features. Features at different scales are extracted using ResNet50 or ResNet101 to generate multiscale feature image pyramids built at different resolutions. Moreover, low-level high-resolution images contain more spatial information than high-resolution images

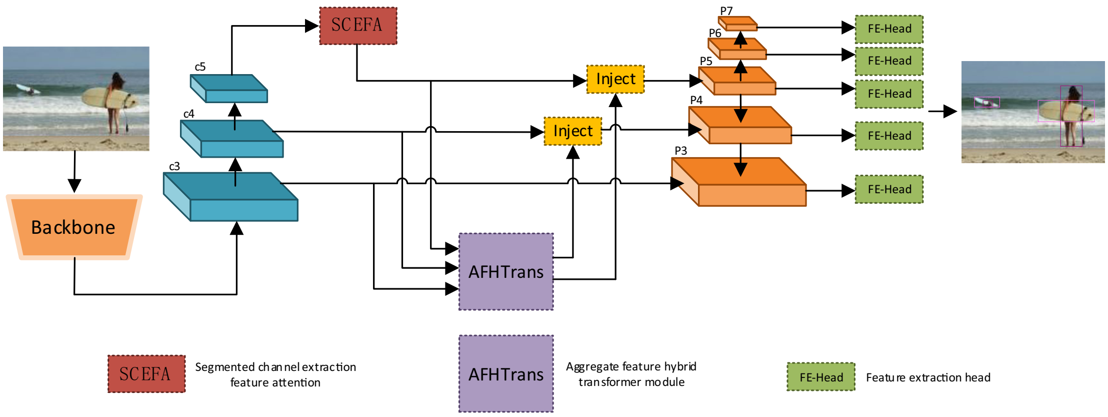


Fig. 2 Overview of the overall GLFT framework diagram

and can be used for the detection of small objects. A structure that preserves the FPN is used at the back end of the model to fuse the efficiently acquired global and local multiscale information in a top-down manner, enriching and effectively utilizing the low-level spatial information. High-level low-resolution feature maps contain additional semantic information and can be used for the detection of large objects.

For this reason, the SCEFA module is designed to extract high-level local features. As shown in Fig. 3, the SCEFA module contains four parts.

Feature grouping

To reduce the number of computational parameters, features are extracted using segmentation channel dimension split grouping. The information from the high-level local feature maps is first split into multiple groups along the channel dimension. Suppose the input is characterized by X ∈ $\mathrm{R}^{C\times H\times W}$ (the C channel number is 2048, H is the height, and W is the width), and the input X is split into Z groups along the channels:X $= [X_1, X_2, \ldots\ldots, X_Z] \in R^{C/Z\times H\times W}$.

Segmentation channels for extracting multiscale features

Each group of features is subsequently segmented into two paths, and the two paths are subjected to fine feature context information extraction via the proposed multiscale full-text channel extraction module (MFCEM), which is the core of the SCEFA module. The MFCEM of the SCEFA block, as shown in Fig. 4, uses group convolution of different kernels to extract high-level localized multiscale channel feature map information, which is capable of obtaining feature maps with different receptive fields and different resolution levels. Each branch of the group convolutional branch can learn multiscale spatial information over a wide range of horizons. A global text (GT) module is added after each branch of the group convolutional branch of the weighted path, modelling the global context while saving the number of parameters. The GT module here (shown in the gray dashed box) consists of the attention head (shown in the blue dashed box) and the transform neck (shown in the yellow dashed box). The attention head is a structure in which the output feature maps after channel normalization and Softmax [53] are cross-multiplied with the original input feature maps to make effective judgments about the importance of the channel weighting information. The transform neck is a structure that combines an extraction module, layer normalization, and Relu [54] and is able to further extract channel feature maps and effectively reduce the number of parameters in the MFCEM. The main purpose of the GT module is to effectively encode the relevant information in the feature map into the full-text attention map, thus enabling MFCEM to correctly discriminate between features of different scales or dimensions needed, enhancing the ability of the data to adapt to multiscale features, and providing more global channel information. According to CANet, only channel information coding is considered while ignoring location information, and only local information is considered to be captured, while global information over long distances is missing. Therefore, CANet [55] remedies the above deficiencies by embedding spatial attention into channel attention using two one-dimensional global pooling structures. In our MFCEM method, global textual information is extracted by segmenting channels from the input feature map via multiple group convolutions.

In detail, feature X_y will be split into two paths along the channel: $X_{y1}, X_{y2} \in R^{C/2Z\times H\times W}$. In MFCEM, first, after the convolution of two multibranch groups, the channels are divided into C/4 segments, reducing the number of computational parameters. The input feature map is compressed at different resolutions to obtain rich feature information, and cross-channel information interaction is realized while extracting multiscale spatial information. Then, the GT module follows the structure of GCNet [56], the feature maps generated by the convolution of each branch group are converted into weights to discriminate the important features in the attention head of the GT module, and the important feature map information is extracted after the transform neck, which further reduces the computational parameter cost. The MFCEM generates global and local channel information, and the detailed process equations are as follows:

$$\begin{aligned}\Phi_j &= Conv_i(k_u, g_v), \#\\ \beta_i &= \sigma(SN(\Phi_1(X_{yni}))) \times \Phi_1(X_{yni}), i = 1, 2, 3, 4,\\ \rho_i &= ET_2(\delta(LN(ET_1(\beta_i)))),\end{aligned} \tag{1}$$

Here, $X_{yni} \in R^{C/8Z\times H\times W}$, n $=$ 1, 2 and andΦ_j, $j = 1, 2$ represent the process of group convolution of the weighted paths and group convolution of the initial paths in the ith stage, respectively, where the convolution kernel size is k_u and the convolution group size is g_v. SN denotes channel normalization, and β_i represents the channel information attention map output from the attention head of the GT module. ET stands for the extracted module, and LN stands for layer normalization.ρ_i represents the channel information attention feature map of the GT module-transformed neck output.σ denotes the Softmax function, and δ denotes the Relu activation function.

Finally, the output channel feature map information of each branch is aggregated, and the obtained global multiscale channel feature map is subsequently recalibrated with

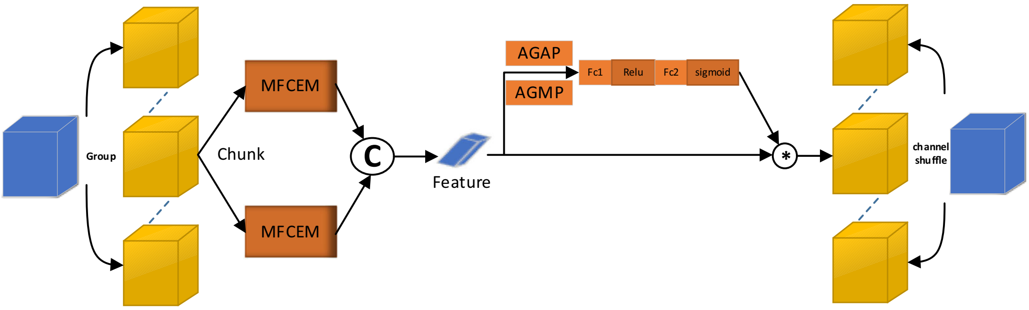


Fig. 3 Detailed structure of the SCEFA block

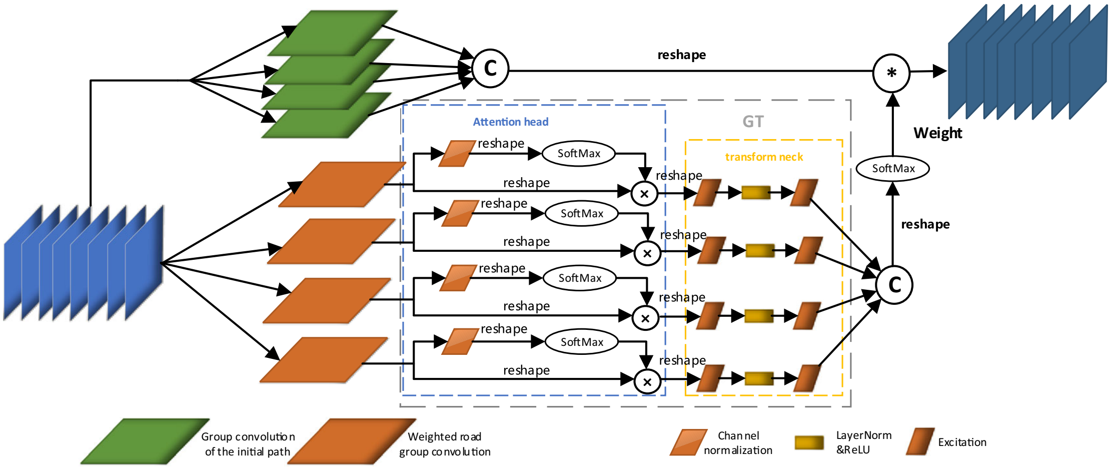


Fig. 4 Detailed structure of the MFCEM

weights and corresponding probability values via softmax, which realizes the interaction between local and global channel information. The final output feature map is then generated by multiplying the product of the recalibrated weight information with the feature maps fused by the group convolutional channels of the initial path. Overall, this approach enables the two paths to simultaneously extract multiscale spatial information for different channels, establish long-distance channel dependencies, and effectively minimize the impact of channels on the number of parameters. Finally, the overall process equation for the MFCEM is as follows:

$$\begin{aligned}\alpha_i &= \Phi_2\left(X_{yni}\right), i = 1, 2, 3, 4, \\ Y_n &= \sigma(Cat(\rho_1, \ldots, \rho_4)) * Cat(\alpha_1, \ldots, \alpha_4),\end{aligned} \tag{2}$$

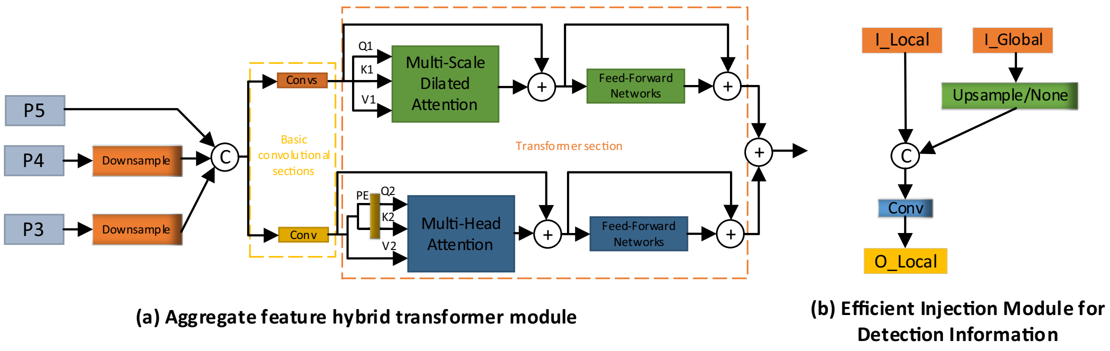
Here, $Y_n \in R^{C/2Z \times H \times W}$, $n = 1, 2$ represents the final output feature map, α_i represents the output feature map of the group convolution after the initial path of the ith stage, and Cat denotes a fully connected process.

Channel feature adjustment

As shown in Fig. 3, first, the fused output feature maps of the connected channel dimension information extracted from the segmented channels are input into adaptive global average pooling (AGAP) and adaptive global maximum pooling (AGMP) to enable the postprocessing conditioning fusion of the channel and spatial dimension information, respectively, to better adapt to the global and local channel information extracted from each group. Then, the regulated feature weight information is further extracted and processed to maximize the regulated useful feature weight information by the fully connected F1, Relu, fully connected F2, and sigmoid [57] functions. Finally, the feature weight information is multiplied by the fused original output feature map to achieve feature conditioning. The detailed flow formula for this part is as follows:

$$\begin{aligned}\gamma_m &= Cat(Y_1, Y_2), \\ O_m &= Sigmoid(F2(\delta(F1(\varphi(\gamma_m) + \omega(\gamma_m))))) * \gamma_m,\end{aligned} \tag{3}$$

Here, $\gamma_m \in R^{C/Z \times H \times W}$, m = 1, 2, ..., z.$\varphi$, ω represent AGAP and AGMP, respectively. $O_m \in R^{C/Z \times H \times W}$ denotes the final output feature map through feature conditioning.



(a) Aggregate feature hybrid transformer module

(b) Efficient Injection Module for Detection Information

Fig. 5 Main structure of AFHTrans and the structure of the efficient injection module for detecting information

Channel shuffle

Like in ShuffleNetV2, the SCEFA module follows the structure of SANet to improve the interaction of channel information between groups by putting the conditioned output features through a channel shuffling operation for interfeature group communication. The original channels are reshaped while being redistributed and mixed to ensure that the number of input channels and the number of output channels are uniform.

The convolution-based aggregated feature hybrid transformer module

This section details the combination of the aggregate feature hybrid transformer for convolution (AFHTrans) module and the efficient injection of the detection information module. Figure 5a shows the general structure of the AFHTrans module. It consists of a basic convolutional section (shown in the yellow dashed box) and a transformer section (shown in the orange dashed box). Based on the original structure of the Transformer, the core parts of its position encoding (PE), multi-head attention (MHA), and feedforward neural network (FFN) are extracted. Inspired by the use of Dilateformer [26], a combination of multi-scale dilated attention (MSDA) is introduced, which not only achieves a trade-off between computational cost and sensory wildness but also further optimizes global dependency modelling and effective aggregation of semantic multiscale information. Then, the basic convolutional section and the two attention layers are cascaded to construct a module that combines convolution and a transformer, which effectively extracts global feature information and local feature information. Thus, the AFHTrans module not only extracts multiscale channel features but also reduces global object pixel independence.

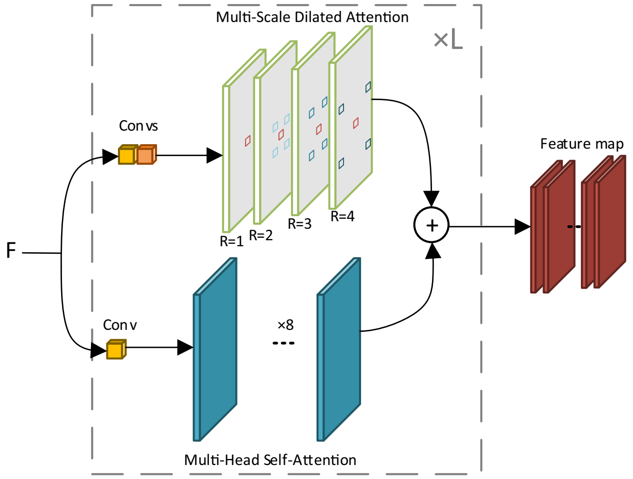


Fig. 6 Stereogram of the combined convolution and transformer module in AFHTrans

The feature extraction process is divided into two stages: local feature extraction by convolution and the establishment of remote interdependencies. By using the basic convolutional section, it is possible not only to fully extract the pixel space information in the local feature extraction stage but also to save computations. To save the computational resources of the video memory, the feature layer inputs were then downsampled (average pooling operation) to the top resolution, and the feature map information of the three layers was fused using a three-way fully connected approach. The three-dimensional structure of the convolution and transformer combination is shown in Fig. 6. L represents the number of layers in the convolution and transformer layers. Assume that the input features after full connection are F$R^{3C^{*}\times H\times W}$($C^{*}$=256). The AFHTrans module is divided into a multiscale dilated attention path (upper part in Fig. 5a) and a multihead attention path (lower part in Fig. 5a). In the multihead attention path, a conv block ordinary convolution is used to extract local features and compress the channel. In multiscale dilated attention paths, the use of a combination of Convs block ordinary convolution and depth-separated convolution [58] of residuals enables the extraction of localized features while compressing the channel and extracting spatial detail information. In this case, depth-separated convolution is used by the module to extract spatial information from the feature map and to improve the interaction of feature information in each channel. The extraction performance of the converter is improved by deep separable convolution, which also reduces the number of computational parameters. The computational formula (4) is shown below.

$$\begin{aligned} I_1 &= Conv(F) \\ I_2 &= Dw_{Conv}(I_1) + I_1, \end{aligned} \tag{4}$$

Here, $I_1,\ I_2 \epsilon R^{C^* \times H \times W}$, $Conv$ stands for ordinary convolution, and $andDw_{Conv}$ represents depth-separated convolution. In this case, the input feature maps are filtered by depth-by-depth convolution, and the input channels are integrated by point-by-point convolution.

In the remote interdependency establishment phase, the global feature acquisition capability is not only effectively improved by the transformer but also capable of capturing multiscale contextual semantic dependencies. This part of the whole process combines the standard layer and the residual structure. In the multiscale dilated attention path, first, the input feature maps go through the standard layer to generate Q_1, K_1, and V_1 to accelerate convergence. Afterwards, Q_1, K_1, and V_1 are computed by the multi-scale dilated attention and then linearly mapped by the FFN, thus expanding the receptive field and improving the mutual independence of the object pixels. In the multi head attention path, first, the feature maps of the inputs to the basic convolution section are passed through the standard layer to Q_2 and K_2, and V_2 generate to accelerate convergence. Then, Q_2 and K_2 are passed through PE to correct the feature map position information. After that, Q_2, K_2, and V_2 are computed by the multi-head attention mechanism and then linearly mapped by the FFN in the standard layer to enhance the extraction of global features and spatial information. Eventually, the feature maps of the two parts are aggregated to fuse rich local and global feature information. The formula is shown below:

$$\begin{aligned} &MSDA = Softmax\left(\frac{Q_{1ij}K_{1r}^T}{\sqrt{d_{k_1}}}\right)V_{1r},\ 1 \le i \le W;\ 1 \le j \le H, \\ &MHA = Softmax\left(\frac{Q_2K_2^T}{\sqrt{d_{k_2}}}\right)V_2, \\ &T = rFFN(\mathrm{rMHA}(\mathrm{PE}(I_1),\ \mathrm{PE}(I_1),\ I_1) + \mathrm{rMSDA}(I_2,\ I_2,\ I_2)), \end{aligned} \tag{5}$$

Here, $\mathrm{T}R^{C^* \times H \times W}$ denotes the output feature map after AFHTrans fusion, Q_{1ij} denotes the query of position (i, j) in the original feature map, and rinK_{1r} and V_{1r} denote the dilation rate of the multi-scale dilated attention. rinrMHA, rMSDA, and rFFN denote the residual structure operation. PE stands for the position encoding process. By default, a standard layer is added before each computational step in Eq. (5) to accelerate model training.

As shown in Fig. 5b, in the Injection's Detection Information Efficient Injection Module, a simple combination of structures is used to inject efficient information into the middle and high feature layers of the neck to fuse the feature maps to transmit global and local feature information. Following TopFormer's information injection module, I_Local is the input path for the local network feature layer, and I_Global is the input path for the AFHTrans module, where the fusion of mid-layer features requires the addition of an upsampling operation. Afterwards, the two feature information sources are fully fused by full connectivity, and later convolution is used to compress the channel and improve the feature extraction ability to generate the feature map output from the middle- and high-level fusion information. Therefore, the calculation formula is:

$$\mathrm{O}_{\tau} = Conv\Big(\mathrm{Cat}\Big(\mathrm{I}_{\mathrm{Local}(\mathrm{X}^*)},\ \mathrm{Interpolate/None}\big(\mathrm{I}_{\mathrm{Global}(\mathrm{T})}\big)\Big)\Big),\ \tau = 1,\ 2, \tag{6}$$

Here, $\mathrm{O}_{\tau} \epsilon R^{C^* \times H \times W}$ represents the feature map of the AFHTrans module after fusion with the middle and upper layers, Interpolate/None represents the middle layer upsampling operation and the high layer no upsampling operation.

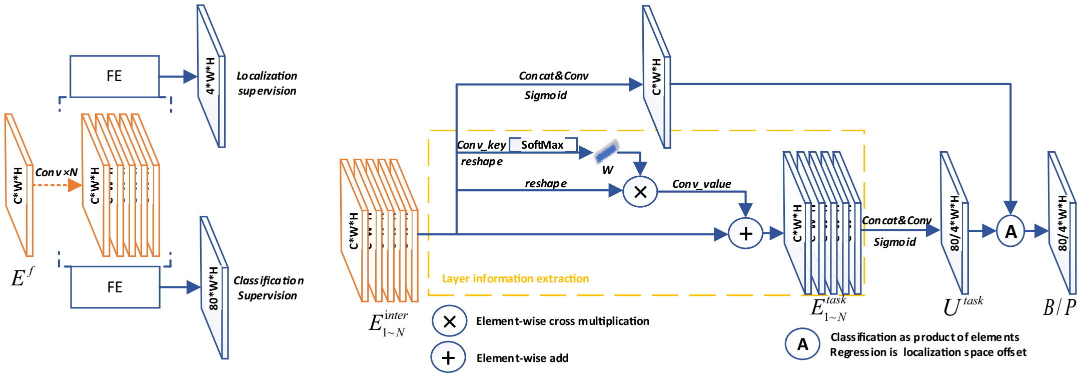
Finally, the information injected by the AFHTrans module is transferred to the FPN for feature fusion and later pooled into the feature extraction detection head for classification and localization after full connectivity. The process is as follows:

$$\Delta = \Gamma_{cls,reg}(Cat(\mathrm{FPN}(\mathrm{O}_{\tau 1},\ \mathrm{O}_{\tau 2},\ \ldots,\ \Upsilon))), \tag{7}$$

Here, $7R^{C^* \times H \times W}$ represents the feature maps of the layers that have not been injected with information by the AFHTrans module, and FPN represents the feature pyramid network fusion feature process. $\Gamma_{cls,reg}$ denotes the processing of the feature extraction detection header. $1R^{C^* \times H \times W}$ represents the final output of the detector.

Feature extraction head

Improving the traditional parallel head structure can effectively compensate for the lack of processing features in the front-end model and provide comprehensive supervision of the final feature information. After front-end model network processing, the acquisition of local and global features and long-range object pixel dependencies of the model are effectively improved. However, for the back-end head task, there is a lack of extraction of global feature information from the feature maps after the fusion and stitching of features from each layer of the FPN. Therefore, to enhance the extraction of global feature information and optimize the head structure, we consider two aspects: on the one hand, improving the information interaction of the detection head. On the other



(a)Feature extraction head(FE-Head)

(b)Feature extractor

Fig. 7 General structure of the FE head and the structure of the feature extractor

hand, the attention of the detecting head is improved to the feature information.

For this purpose, the feature extraction head (FE-Head), which consists of a feature interaction extractor and a feature extractor (FE), is designed as shown in Fig. 7(a). Following the structure of TOOD [29], to enhance the interaction of information by the detection head, a feature interaction extractor is used to learn task information interaction features from multiple convolutional layers. As shown in orange in Fig. 7(a). This design not only facilitates information interaction but also provides multiscale characterization of the different sensory fields involved in the two tasks. Assume that $\mathrm{E}^f \epsilon R^{C' \times H' \times W'}\left(C'\text{numberofchannels}, H'\text{height}, W'\text{width}\right)$ is the output feature of the FPN. The feature interaction extractor is computed as follows:

$$E_k^{inter} = \delta\left(conv_k\left(\ldots\delta\left(conv_{kj}(\mathrm{E}^f)\right)\ldots\right), \forall k, j\epsilon\{1, 2, \ldots, N\}, \quad (8)$$

Here, k and kj of $conv_k$ and $conv_{kj}$ represent the kth convolutional layer and the jth activation function of the kth convolutional layer of N consecutive convolutional layer processes, respectively. Therefore, rich multiscale features can be efficiently extracted from an FPN using a single branch.

Feature maps rich in feature information are generated through information interactions. Due to the single-branch structure, it becomes difficult to fully obtain the global feature information of the corresponding task. This is also discussed in [45, 59]. For this purpose, the features after the information interaction can be input into the two FEs to extract the features for the corresponding tasks from the direction of attention. The features required to correspond to different tasks are different. Therefore, as shown in Fig. 7b, a layer information extraction mechanism (shown in the yellow dashed box) is proposed to obtain information by subtasking task-specific features dynamically computed at the layer level. The formula is as follows:

$$E_k^{task} = \varphi_k + E_k^{inter}, \forall k\epsilon\{1, 2, \ldots, N\}, \quad (9)$$

Here, φ_k is the kth feature map of the layer information extraction $\varphi \in R^{C' \times H' \times W'}$. φ is able to obtain the dependencies between layers while extracting global features after layer information interaction feature computation:

$$\varphi = Cv(Softmax(Ck(E_k^{inter}) \times E_k^{inter})), \quad (10)$$

Here, Ck represents the channel normalization, and Cv represents the process of going through convolution, the standard layer, ReLU, and convolution. The feature map obtained after computing the corresponding task can fully extract features and reduce the dimensionality of the adjustment process to reduce the number of parameters. The E^{task} is a tandem characterization of E_k^{task}. Finally, the final extracted features for localization or classification are obtained from each E^{task}:

$$U^{task} = conv_2\left(Sigmoid\left(conv_1\left(E^{task}\right)\right)\right), \quad (11)$$

where U^{task} is a tandem feature of U_k^{tsak} and $conv_1$ and $conv_2$ are 1 × 1 convolutional layers for variable dimensions. U^{task} is converted to a dense classification score $\mathrm{P} \in R^{H \times W \times 80}$ using the sigmoid function or to a bounding box $\mathrm{B} \in R^{H \times W \times 4}$ by applying the distance-to-bbox method in [12, 13]. Finally, following the TOOD approach, the spatial distribution and task coordination learning are further adjusted for both the P and B tasks to obtain global feature information for the classification and localization tasks, respectively. Thus, high-resolution feature map information with additional edge information is extracted to improve the classification and localization accuracy of the detector.

Experiments

Datasets

Our approach is evaluated using two main datasets: the PASCAL VOC2007 + 2012 [60] and COCO2017 [61] datasets. PASCAL VOC2007 + 2012 is a dataset of a comprehensive class of scenarios. The dataset contains 27,088 images in 20 categories. Of these, 16,551 out of 27,088 images were used for training, and 4952 were used for validation. The COCO2017 dataset is a dataset of a class of common everyday objects. A total of 163,957 images were included in 80 categories. Of these, out of 163,957 images, 118,287 were used for training, 5000 for validation, and 40,670 for testing. The training and verification operations have labels with corresponding information.

Implementation details

GLFTNet uses ResNet50 or ResNet101 pretrained on the ImageNet dataset as the backbone. The SCEFA module and the AFHTrans module were combined to serve as the NECK, where the number of heads for multi-head attention and multi-scale dilated attention were set to 8 and 4, respectively. FE-Head as a Model Head. The model was trained using stochastic gradient descent (SGD) with momentum = 0.9 and weight decay = 0.0001 and a strategy of setting the learning rate in segments with gamma = 0.1 and linear warmup steps = 500. The initial learning rate was set to 0.001.

The resolution of the images in the experiment is set to 800 × 1333, and the data enhancement includes operations such as random flipping and multiscale training. Training was performed with the number of epochs set to 12 and the batch size set to 4. Experiments were performed on a single NVIDIA Tesla V100 GPU device. Our model is optimized for classification and localization tasks using two loss functions. The first classification loss function is focal loss. The second bounding box regression loss function is the GIOU [62]. Thus, the total loss function is as follows:

$$L = w_{cls} L_{cls} + w_{reg} L_{reg}, \tag{12}$$

Here, L_{cls} denotes the classification loss function, and L_{reg} denotes the boundary regression loss function. w_{cls} and w_{reg} are weights set to 1.0 and 2.0, respectively.

Evaluation metrics

In this paper, the intersection over union (IOU), average precision (AP), mean average precision (mAP), frames per second transmission (FPS), parameters (number of parameters), and number of floating point operations (FLOPs) are used as evaluation metrics for the model. IOU denotes the degree of resynthesis of the predicted frame with the true frame, which is the ratio of the intersection and concatenation of the detection result and ground truth. This metric usually sets the threshold ($\mathrm{IOU}_{threshold}$) to determine the truthfulness of the prediction frame. That is, it is usually set to 0.5, with a value greater than 0.5 considered a correct prediction and a value less than 0.5 considered an invalid detection. AP and mAP (here, the IOU is generally 0.5) are used to evaluate the ability of the algorithm to correctly detect an object and are the most important evaluation metrics for detection algorithms. Depending on the accuracy or inaccuracy of the prediction frames, the evaluation object algorithm includes four samples: true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). TP is the number of predicted frames with IOU $> \mathrm{IOU}_{threshold}$, FP is the number of predicted frames with IOU $\leq \mathrm{IOU}_{threshold}$, and FN is the number of real frames that are not detected correctly. As a result, $P_{precision}$, R_{recal}, and AP are calculated as follows:

$$\begin{aligned} &P_{precision} = \frac{TP}{TP+FP}, \\ &R_{recal} = \frac{TP}{TP+FN}, \\ &AP = \int_{0}^{1} P_{precision} d R_{recal}, \end{aligned} \tag{13}$$

AP denotes the area of the curve formed by $P_{precision}$ and R_{recal}. It specifically refers to the accuracy of a particular category among multiple categories. Based on the above description of AP, the specific formula for determining the mAP is as follows:

$$mAP = \frac{\sum_{k=0}^{C} AP_k}{C}, \tag{14}$$

The mAP is the average of the AP values after summing multiple categories. C denotes the total number of multiple categories, and k represents different categories. In particular, the AP for the COCO dataset calculates the multicategory mean by default, i.e., the mAP. The FPS is an important metric for evaluating how fast an algorithm can detect; it is the ratio of the number of frames to the elapsed time. In addition, Params and FLOPs represent the computational space complexity and computational complexity of the model, respectively.

Ablation study

In this section, ablation experiments are performed to verify the validity of our method. The validity of structures such as the SCEFA, AFHTrans, and FE-Head was verified by ablation experiments on the PASCAL VOC2007 + 2012 and COCO2017 datasets.

Table 7 Variation in the convolutional group size and convolutional kernel size in the experiments

Backbone	Kernel size	Group size	mAP
ResNet50	1, 3, 5, 7	1, 2, 4, 8	80.56
ResNet50	1, 3, 7, 9	1, 2, 2, 8	80.63
ResNet50	3, 5, 7, 9	1, 4, 8, 16	**81.79**
ResNet50	3, 5, 5, 7	1, 4, 4, 16	80.80

The bold symbols represent the best results of the convolution kernel and convolution group size experiments

SCEFA module

According to Fig. 3, the SCEFA module is able to fully extract the channel information from the split segmentation. First, the features are grouped for segmentation to decrease the number of computational parameters. Afterwards, according to Fig. 4, MFCEM can effectively extract various global pieces of information from each group of segmented channels to generate an information-rich feature map. In this case, the data are first divided into two paths to go through the group convolution in the MFCEM to compress the channels to produce spatially scaled feature maps. Here, ResNet50 is used as the backbone network. The appropriate convolution kernel and convolution group size for group convolution are determined. The performance of the MFCEM in the network is tested on the PASCAL VOC2007 + 2012 dataset using different convolutional kernel sizes and convolutional group sizes, as shown in Table 7, considering the number of channels and the need for practical experimental comparisons, based on the work of [24, 52]. Table 7 shows the experimental results. The average accuracy obtained for the first group with a convolution kernel size and convolution group size set to 1, 3, 5, and 7 and 1, 2, 4, and 8 is 80.56%. The next set of settings were 1, 3, 7, 9, and 1, 2, 2, 8, yielding a result of 80.63%, an improvement of 0.07% over the previous set. The convolution kernel sizes for the following two groups are set to 3, 5, 7, 9 and 3, 5, 5, and 7; the convolution group sizes are set to 1, 4, 8, and 16; and 1, 4, 4, and 16, respectively. The final accuracies obtained are 81.79% and 80.80%, respectively. In a comprehensive comparison, the optimal convolution kernel sizes 3, 5, 7, and 9 and convolution group sizes 1, 4, 8, and 16 were selected as the parameters for group convolution in the module. Therefore, the use of multiple convolution kernels and convolution groups of different sizes and a large range of numbers can improve the speed and accuracy of parallel training, and the appropriate adjustment of the size of the convolution kernels and convolution groups can enhance the ability to obtain feature map information. The experimental results in the table effectively prove our conclusion.

The feature map is input to the GT module after group convolution extraction of the weighted paths, which has an attention head to discriminate the important information of each branch and a transform neck to fully extract the feature information. As a result, the GT module can extract rich global textual information. Then, the channel attention weight values are generated by the Softmax function, which is effective at establishing long-range channel dependencies and calibrating weight information. To demonstrate the effect of the GT module and the Softmax comparison Sigmoid function on the experimental results, the above structures were added. Table 8 shows. A total of 81.28% of the mAP is achieved using only the GT module in the first row. The second step involves removing the GT module and using the Softmax function to determine the probability corresponding to the weights; this approach yields an mAP of 80.93%, which represents a decrease in the indicator, indicating that the GT module is indispensable. Finally, to verify the effect of Softmax on the sigmoid function, the use of the sigmoid function alone in the third row improved the accuracy by 0.33% over that of the Softmax function alone, confirming that the replacement of the sigmoid function indeed improved the accuracy. However, by combining the GT module with the sigmoid and softmax functions, the mAP reached 81.58% and 81.79%, respectively, with a difference of 0.21% between the two accuracies, and the combination of the GT module and the softmax function resulted in an improvement of 0.51% compared to the original use of the GT module alone. Therefore, the GT module and Softmax function can better extract important information about the weights so that the detection accuracy is improved because the overall models of these two parts are indispensable.

The two segmented channels are then fused dimensionally after the global and local channel information is extracted by the MFCEM of the SCEFA module. This is followed by channel feature adjustments to better accommodate the fused global and local channel information. The feature maps of the adjusted channels are then channel-swapped to perform intergroup communication to realize information interaction. To verify the effects of group segmentation, channel feature adjustment, and channel shuffling on the model effects and parameters, the above operations were performed sequentially. Table 9 shows. The first row removes the group segmentation process, and the mAP reaches 81.57%; here, the parameter is 41.54 M. The last two rows sequentially removed the channel feature adjustment and channel shuffling to obtain average accuracies of 81.45% and 81.40%, respectively, and parameters of 35.17 M and 36.25 M were obtained. The highest accuracy was achieved under channel feature adjustment and channel shuffling; therefore, these two processes provide effective extraction of feature information, but the number of parameters is large. After the latter use of group segmentation, the parameters drop by 6.37 M

Table 8 Ablation experiments validating the global text module (GT) and the Softmax comparison sigmoid function

Backbone	GT	Softmax	Sigmoid	mAP
ResNet50	√			81.28
ResNet50		√		80.93
ResNet50			√	81.26
ResNet50	√		√	81.58
ResNet50	√	√		81.79

Table 9 Ablation experiments were performed in the SCEFA module to verify the effects of group segmentation, channel feature adjustment, and channel shuffling

Backbone	GS	CFA	CS	mAP	Params (M)
ResNet50		√	√	81.57	41.54
ResNet50	√		√	81.45	35.17
ResNet50	√	√		81.40	36.25
ResNet50	√	√	√	81.79	36.41

GS represents the group segmentation process, *CFA* represents channel feature adjustment, *CS* represents channel shuffling

Table 10 Experimental comparison of the SE-Net and the SCEFA

Method	Backbone	mAP	FPS
SE-Net	ResNet50	81.25	13.75
SCEFA (Ours)	ResNet50	81.79	12.98

SE-Net denotes the substitution for channel attention, *SCEFA* denotes the substitution for segmentation channel extraction feature attention

Table 11 Comparison of modelling effects against basic convolutional sections

Method	mAP
+ Basic convolutional	81.79
− Basic convolutional	81.44

The basic convolution section includes convolution and depth-separated convolution. + Use the basic convolution section. − Does not use the basic convolution section, but needs to add a dimensionality reduction convolution to reduce the number of parameters

and 5.29 M, respectively, and the number of parameters is greatly reduced.

Finally, by adding all three processes, the mAP further increases to 81.79%, and the parameter quantization is 36.41 M, which achieves a remarkable accuracy-parameter balance. Overall, the group segmentation, channel feature adjustment, and channel shuffling processes in the SCEFA module enhance the spatial adjustment of feature information and the information interaction of feature maps and reduce the number of parameters.

Finally, the validity of the SCEFA module is fully demonstrated. The SCEFA module and its structurally similar SE-Net module were subjected to comparative experiments, the results of which are shown in Table 10. In the first line, replacing the SCEFA module with the SE-Net module achieves an mAP of 81.25% and a speed of 13.75 FPS. Similarly, the use of the SCEFA module proves to be effective by improving the accuracy by 0.54%, despite a 0.77 FPS decrease in speed. Overall, the speed improved when using the SE-Net module, but it still did not achieve the same performance results as the SCEFA module. This further demonstrated the importance of the SCEFA module for the overall model.

AFHTrans module

With respect to the AFHTrans block, it is verified that the combination of convolution and the transformer improves the model performance. For the FPN, there is a lack of information interaction across layers, so aggregated connections are used. Moreover, after three layers of aggregation, the feature map is rich in information. Therefore, convolution and dual-attention fusion joining are used to enhance the ability to acquire both local channel spatial information and global information. Among them, embedding convolution in a transformer not only enhances spatial information extraction but also improves the interaction between channel information. To this end, ablation experiments are performed on the PASCAL VOC2007 + 2012 dataset to validate the effect of the embedded convolution on the modelling effect. As shown in Table 11, the mAP reaches 81.79% with the addition of the basic convolutional section, decreases by 0.35% under no addition, and reaches 81.44%. As a result, whether the basic convolution section is added to the AFHTrans module has a significant impact on the results, effectively verifying that the combination of convolution and the transformer contributes

Table 12 Ablation experiments were performed for the multi-head attention and multi-scale dilated attention components

Backbone	BC and FFN	MHA	MSDA	mAP
ResNet50	√	√		81.41
ResNet50	√		√	81.58
ResNet50	√	√	√	81.79

BC denotes the basic convolutional section, *MHA* denotes a multi-head attention section process, *MSDA* denotes a multi-scale dilated attention section process, *FFN* denotes a feedforward neural network

Table 13 Variation in expansion rates for multi-scale dilated attention in the AFHTrans module

Backbone	Dilation rate	mAP
ResNet50	2, 4, 4, 6	81.55
ResNet50	2, 4, 6, 6	81.62
ResNet50	2, 4, 6, 8	**81.79**
ResNet50	2, 4, 8, 8	81.71

The bold text represents the best results of the experiment at different dilation rates

to the results. Therefore, convolution can acquire local features and spatial location information.

Using ResNet50 as the backbone network, the effects of the multi-head attention and multi-scale dilated attention components on the network gain were observed to verify the effectiveness of both components in the model. Table 12 shows the ablation experiments for both the multi-head attention and multi-scale dilated attention components under the use of the basic convolutional component and the feedforward neural network. The mAP reached 81.41% and 81.58% when using multi-head attention and multi-scale dilated attention, respectively. The combination of the two components amounted to 81.79%, a 0.38% improvement over the use of multi-head attention. It is well demonstrated that the combination of multi-head attention and multi-scale dilated attention enhances the model's gain effect and improves the ability to acquire global features and establish object pixel dependencies at long distances.

In the multiscale dilation attention section, further experiments are conducted to verify the effect of dilation rate changes on our model to validate the performance of the AFHTrans module in the model. The results of the experiments are shown in Table 13. The first and second rows of the table with expansion rates of 2, 4, 4, and 6 and 2, 4, 6, and 6 correspond to accuracies of 81.55% and 81.62%, respectively, while the third row with expansion rates of 2, 4, 6, and 8 achieves the highest average accuracy of 81.79%, and the final expansion rate of 2, 4, 8, and 8 has a decrease in accuracy of 0.08% compared to the previous row. Considering these results, as with the size law of the convolution kernel and convolution group above, more expansion events with different sizes and large ranges can effectively expand the sensory field and capture feature information. Similarly, the use of an appropriate dilation rate helps to improve the effectiveness of the model. Thus, our model uses the 2, 4, 6, and 8 best expansion rates in multiscale expansion attention.

Table 14 Experimental tests of the comparisons of different connections

Backbone	Combination	mAP
ResNet50	{1, 2}	79.97
ResNet50	{1, 3}	80.72
ResNet50	{2, 3}	81.79
ResNet50	{1, 2, 3}	81.11

Bromos 1, 2, and 3 of the combination represent connections corresponding to low, medium, and high levels, respectively

Table 15 Effect of different positions in the AFHTrans module on the effectiveness of the experiment

Method	mAP
Front	81.79
Behind	80.90

The front is at the front end of the FPN, and the behind is at the back end of the FPN

Experiments were conducted on the effectiveness of the model using different connections in the information injection section of the AFHTrans module. Table 14 shows that the accuracy is 79.97% when connecting the low and middle levels, while the accuracy rises to 80.72% when connecting the low and high levels. The best accuracy of 81.79% was achieved in connecting the middle and high levels; for this reason, this injection was subsequently used in our model. However, there was a 0.68% decrease when connecting all three levels. From the table, it can be found that it is not the case that the more connected the hierarchical layer is, the greater the effect, but the model will improve accordingly with the higher precision of the connected levels. Among other things, we believe that the specific cause is the interaction of the middle and upper levels with the FPN.

To investigate the effect of the location of the AFHTrans module in the FPN on the model, experiments were conducted to place the module in different locations in the FPN to compare the model effects. Placing the module at the front end of the FPN yielded an accuracy of 81.79%, while at the back end, the accuracy decreased to 80.90%. As shown in Table 15, the module placement at the front end can thus effectively extract rich semantic information and facilitate

Table 16 Comparison of the experimental effects on the DETR transformer and AFHTrans

Method	Backbone	mAP	FLOPs(G)	FPS
DETR Transformer	ResNet50	81.67	192	12.51
AFHTrans(Ours)	ResNet50	81.79	189	12.98

The DETR Transformer denotes the structure of the Transformer encoder and decoder using DETR, and the AFHTrans denotes the improved structure using the Transformer

Table 18 Comparing the T-Head and FE-Head experimentally

Method	Backbone	mAP	Params (M)
FCOS-Head	ResNet50	79.65	36.60
T-Head	ResNet50	82.28	36.32
FE-Head (Ours)	ResNet50	82.39	36.41

FCOS-Head denotes the use of a multibranch task decoupling header, T-Head denotes the use of a task alignment header, and FE-Head denotes the use of our detection header

further feature fusion in the FPN. Moreover, removing fused information from the back end results in useless and redundant information, leading to a decrease in the ability to obtain valid features. Therefore, the AFHTrans module is placed at the front end of the FPN.

Finally, the transformer part of DETR is very similar to the original transformer. To test the effectiveness of AFHTrans in optimizing the cost and speed of transformer computations. Thus, a comparison experiment was set up using the DETR Transformer (including the encoder and decoder parts of DETR), and the results of the experiment are shown in Table 16. The first row of the table achieves an accuracy of 81.67% with the use of the DETR transformer, with a computational cost of FLOPs and a speed of 193 G and 12.51 FPS, respectively. In the second row of the table, the replacement of AFHTrans results in a 0.12% increase in accuracy, a 3G decrease in computational cost FLOPs, and a 0.47 FPS improvement in speed. Thus, AFHTrans is effective at optimizing the computational cost and speed of the transformer.

FE-Head

The FE-Head not only has task interaction capability but also has the powerful ability to acquire global feature information. Among them, the feature interaction extractor and the feature extractor are the main parts of our detection head. The feature interaction extractor in TOOD has been proven effective, so an experimental study is not developed here. The feature extractor is analysed in the following ablation experiment.

The effects of Conv_key and Conv_value in the feature extractor as well as multiscale and random flipping on the experimental results are explored. As shown in Table 17, the first line has an mAP of 81.10% when only Conv_key is used and the parameter is 36.34 M. The second line has an mAP of 80.22% when using only Conv_value and a parameter of 36.26 M. In the third line, when both Conv_key and Conv_value are used, the precision reaches 81.79%, and the parameter rises by only 0.15 M. As a result, it can be confirmed that the Conv_key and Conv_value boosting effects are significant and reduce the number of parameters, which facilitates the extraction of feature context information. The following two rows sequentially add random flip and multiscale data enhancement operations, yielding metrics with 82.39% and 82.76% excellent accuracy, respectively. Therefore, data augmentation is used so that the richer the amount of data is, the more accurate our model will be. The latter experiments were conducted using the randomized flipped data enhancement method.

The multibranch task decoupling head (FCOS-Head [12]) of the FCOS can effectively suppress the generation of low-quality bounding boxes far from the target location without introducing any hyperparameters. TOOD’s [29] Task Alignment Head (T-Head) improves coordination of classification and localization tasks while enhancing task interactivity. It can also apply the task alignment learning method to adjust the distance between anchor points, thus realizing accurate detection of objects. Among them, FCOS-Head, T-Head, and FE-Head are all anchor-free detection heads. To validate the effectiveness of the FE-Head, the FE-Head was experimentally compared to these two detected head structures. The comparison results are shown in Table 18. In the first row of the table, the accuracy of FE-Head is 2.74% greater than that of FCOS-Head, and the number of parameters decreases by 0.19 M. In the second row of the table, on the other hand, the precision of FE-Head is 0.11% higher than that of T-Head, and the number of parameters is elevated by only 0.09 M. The detection plots of FCOS-Head, T-Head and FE-Head at a confidence level of 0.5 are visualized in Fig. 8. The cars and people in the first and second columns of the first and second rows using FCOS-Head and T-Head had varying degrees of missed detections and low category accuracy, whereas the cars and people in the first and second columns of the third row using FE-Head could be accurately detected (shown as gray circles in Fig. 8). The figure shows that the detection of objects is significantly improved, and the detection accuracy is also greatly improved. The effectiveness of the FE-Head detection head in extracting global contextual information from image objects is fully confirmed. Although the number of parameters was slightly greater than that of T-Head, the detection achieved satisfactory results.

Combination of the SCEFA, AFHTrans module and FE-Head

The combination of the SCEFA module, the AFHTrans module, and the FE-Head yielded excellent detection results when combined with GLFTNet. The entire AFHTrans module section consists of the AFHTrans module and the FPN. Among them, the AFHTrans module, which combines convolution and transformation, plays a key role in global and local feature extraction. To demonstrate the effect of the above three modules, different experimental setups (shown in Tables 19 and 20) are tested to verify that the use of the SCEFA, AFHTrans, or FE-Head can improve the detection performance. Table 19 shows that, compared to the Baseline, the SCEFA module improves the mAP by 1.38% and is 0.23 FPS faster. The AFHTrans module increased the mAP by 1.42% and the speed by 2.90 FPS. With our FE-Head, the mAP increased by 1.45%, and the speed was 0.52 FPS faster. The modules produced significant improvements in accuracy and speed, validating the effectiveness of the modules. When the three modules are combined, the detection performance reaches a high accuracy metric of 82.39%, with a slight decrease in speed of 0.68 FPS; thus, our model achieves a balance of speed and accuracy. An improvement in accuracy of 2.38% over that of the baseline case demonstrates that

Table 17 Conducting FE experiments to validate Conv_key and Conv_value ablation and exploring the impact of employing data augmentation on the experimental results

Backbone	Ck	Cv	RF	MS_{train}	mAP	Params(M)
ResNet50	√				81.10	36.34
ResNet50		√			80.22	36.26
ResNet50	√	√			81.79	36.41
ResNet50	√	√	√		82.39	36.41
ResNet50	√	√	√	√	82.76	36.41

Ck denotes Conv_key, *Cv* denotes Conv_value, *RF* denotes RandomFlip, MS_{train} denotes multiscale training

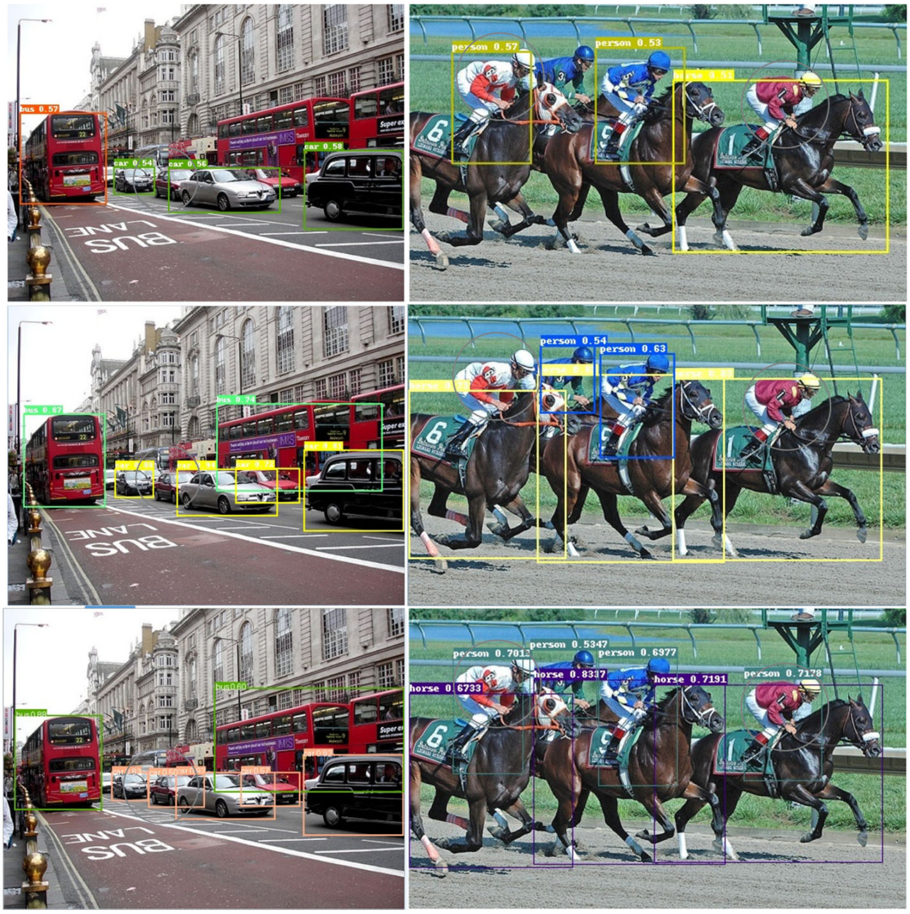


Fig. 8 The top, middle, and bottom graphs show the results of the effect plots using FCOS-Head, T-Head, and FE-Head, respectively, at a confidence level of 0.5

Table 19 Ablation experiments were performed on the PASCAL VOC2007 + 2012 dataset using the SCEFA module, AFHTrans module, and FE-Head

Method	Backbone	SCEFA	AFHTrans	FE-Head	mAP	FPS
Baseline	ResNet50				80.01	13.23
GLFT	ResNet50	√			81.39	13.46
GLFT	ResNet50		√		81.43	16.13
GLFT	ResNet50			√	81.46	13.75
GLFT	ResNet50	√	√	√	82.39	12.64

The baseline represents TOOD without using task-aligned heads, SCEFA represents the process of extracting features after segmentation channels, AFHTrans represents the convolution-based aggregated feature mixing transformer, and FE-Head represents the process of using feature extraction heads

Table 20 Comparison of the results of adding the SCEFA module, the AFHTrans module, and the FE-Head to the COCO2017 validation set in turn

Method	Backbone	AP
Baseline	ResNet50	41.8
Baseline + SCEFA	ResNet50	42.8
Baseline + SCEFA + AFHTrans	ResNet50	43.3
Baseline + SCEFA + AFHTrans + FE-Head	ResNet50	43.5

The baseline represents a TOOD that does not use task alignment headers, and + represents the process of adding modules sequentially

our method effectively improves the detection performance of the detector.

To further validate the effect, the detection plots of the TOOD, SCEFA, and AFHTrans modules were visualized in detail on the PASCAL VOC2007 + 2012 validation set. According to the first row of Fig. 9, the detection accuracy of the airplane has significantly improved, and the detection of medium-sized objects has significantly improved. The small target of a person riding a horse (marked by a red circle) in the second row was detected without any problems, and the objects for detecting the small target were thus improved. The people and chairs (marked by red circles and ellipses) in the third row are correctly detected, allowing the image to improve the problem of missed detection in complex scenes. The lamb in the fourth row (marked by a red circle) is correctly detected, further validating the adaptability of our model to detect objects at different scales. An overall improvement in detection accuracy was achieved compared to that of the baseline. Thus, combining the three modules plays an active and effective role in our model, and our model is competitive in detection applications.

In addition, the SCEFA module, the AFHTrans module, and the FE-Head were added in turn to test the effectiveness of the modules even further on the COCO2017 validation set. As shown in Table 20, the effect of adding the SCEFA module on the extraction of multiscale local features of the channel increases the accuracy by 1.0% over that of the baseline. The later addition of the AFHTrans module for global and local feature extraction by the network, as well as the establishment of object pixel dependencies over long distances, improved the network by another 0.5%. Finally, the addition of FE-Head to the head for effective access to the collected feature context information yields a 0.2% boost. Based on the above results, the validity of our modules is fully verified.

Comparison with other object detectors on the COCO2017 dataset

To further validate the effectiveness of our model, comparative experiments with other research algorithms were conducted on the COCO2017 dataset. The large-scale COCO2017 dataset was used for training, which included the SCEFA module (which improves the extraction of high-level local feature map information), the AFHTrans module (which improves the independence of global and local information extraction and establishes the dependency of pixels on distant objects), and the FE-Head module (which performs the extraction of global feature contextual information at the head). The three main modules are combined to produce a new GLFTNet detection model. As shown in Fig. 1, when comparing GLFTNet with several classical networks, it can be visualized that our model achieves a leading edge. When training our model using multiple scales (480–800), as shown in Table 21, the performance of our detection network is detailed for each parameter. With the use of the backbone networks ResNet50 and ResNet101, GLFTNet achieves 44.3% and 47.0% AP on the COCO2017 test-dev dataset, respectively, outperforming the state-of-the-art detectors Sparse R-CNN [63], Anchor-DETR [64], and DAB-DETR [65]. Compared to other detectors, our model achieves greater progress in terms of improving AP metrics under different IOU sizes, effectively validating its effectiveness for improving global and local feature extraction. The parameters and calculations are also reduced from those of the baseline network, which is at the upper level among the other detection methods.

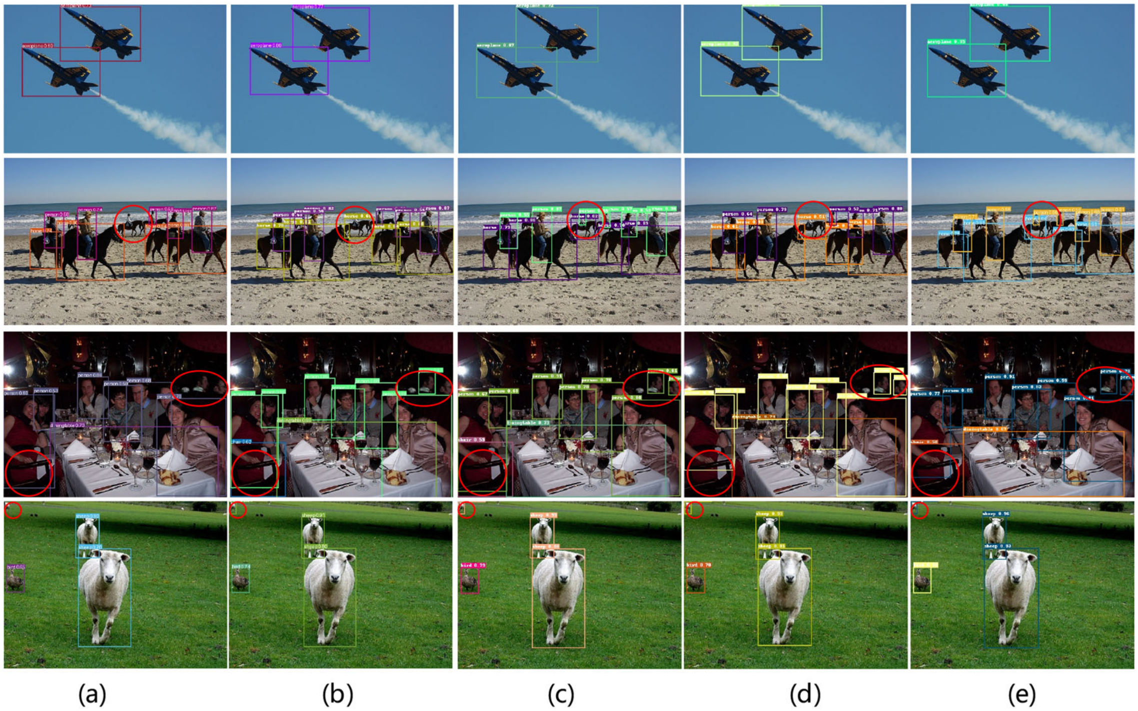


Fig. 9 Effect map results of ablation experiments. **a** Baseline, **b** SCEFA + Baseline, **c** AFHTrans + Baseline, **d** FE-Head + Baseline, **e** GLFTNet

Table 21 Comparison of GLFTNet with several state-of-the-art methods on the COCO2017 dataset

Method	Backbone	Reference	Epoch	MS_{train}	Params(M)	FLOPs(G)	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
CNN-based object detectors												
Faster R-CNN [10]	ResNet101	TPAMI 17	108		60	246	44.0	63.9	47.8	27.2	48.1	56.0
RetinaNet [9]	ResNet101	ICCV 17	36	√	56	197	40.4	60.2	43.2	24.0	44.3	52.2
Cascade R-CNN [7]	ResNet101	CVPR 18	36		77	364	42.8	62.1	46.3	23.7	45.5	55.2
FCOS [12]	ResNet101	ICCV 19	24	√	50	290	41.5	60.7	45.0	24.4	44.8	51.6
ATSS [13]	ResNet101	CVPR 20	24	√	50	292	43.6	62.1	47.4	26.1	47.0	53.6
Sparse R-CNN [63]	ResNet101	CVPR 21	36	√	125	256	43.5	62.1	47.2	26.1	46.3	59.7
TOOD [29]	ResNet101	ICCV 21	24	√	51	267	46.7	64.6	50.7	28.9	49.6	57.0
Based on the Transforner object detectors												
DETR [15]	ResNet101	ECCV 20	500	√	60	152	43.5	63.8	46.4	21.9	48.0	61.8
Deformable-DETR [16]	ResNet50	ICLR 21	50	√	40	173	43.8	62.6	47.7	26.4	47.1	58.0
TSP-FCOS [23]	ResNet101	ICCV 21	36	√	–	255	44.4	63.8	48.2	27.7	48.6	57.3
Anchor-DETR [64]	ResNet101	AAAI 22	50		58	–	43.5	64.3	46.6	23.2	47.7	61.4
DAB-DETR [65]	ResNet101	ICLR 22	50		63	174	44.1	64.7	47.2	24.1	48.2	62.9
GLFTNet(Ours)	ResNet50	–	12		36	189	44.3	62.5	48.0	25.3	47.5	57.2
GLFTNet(Ours)	ResNet101	–	24	√	55	263	47.0	65.3	51.1	28.6	50.2	58.7

The reference column represents the literature sources for comparing object detection algorithms, MS_{train} denotes multiscale training, and the data comparison results (single-model results) are provided by the corresponding official and experimental sources

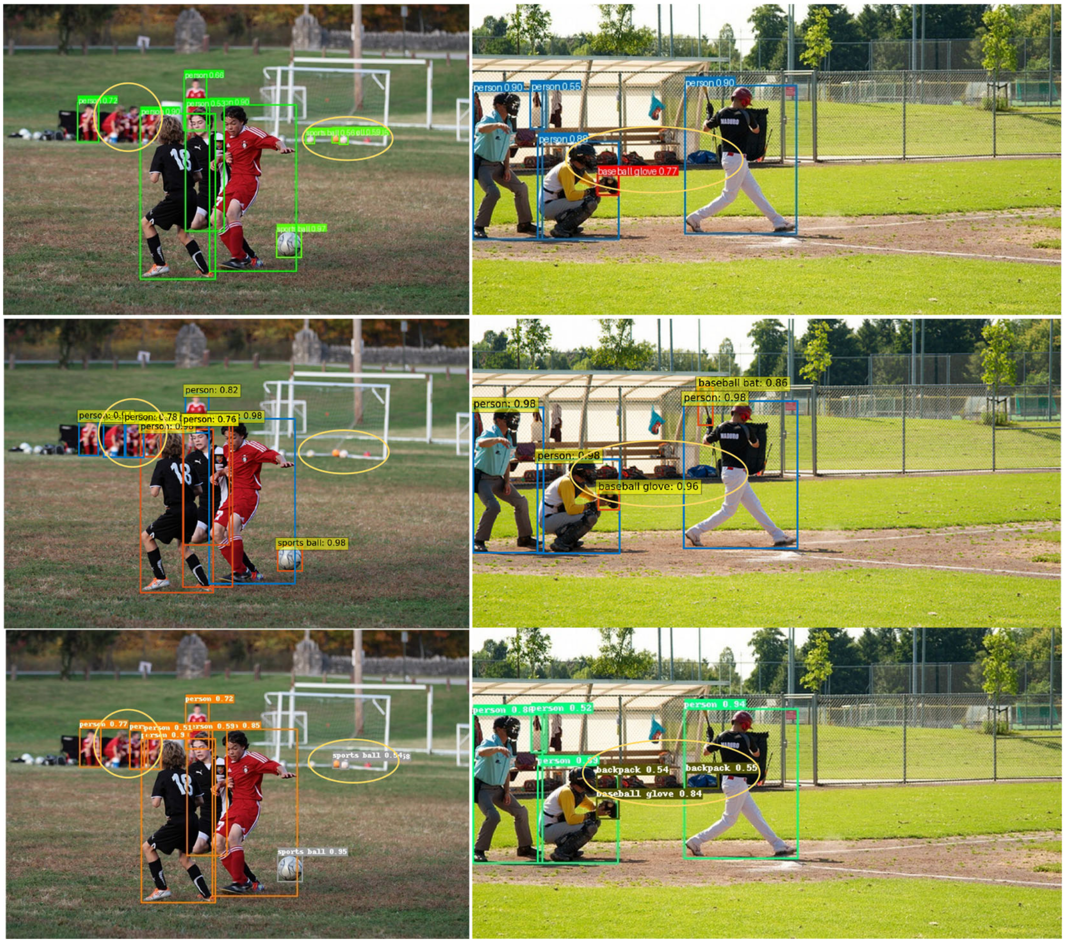


Fig. 10 The top, middle, and bottom two figures show a comparison of the effect plots for TOOD, DAB-DETR, and GLFTNet (Ours), respectively

Our method was compared with TOOD and DAB-DETR for visualizing results on the COCO test set, as shown in Fig. 10. In the first column, TOOD detects only soccer balls, and DAB-DETR detects only people. Our method successfully detects both people and soccer balls correctly (shown by the yellow circle and ellipse), so our method is significantly effective for small and overlapping object detection. Among other things, it benefits from improvements in feature extraction to enhance pixel discrimination in object regions. The backpack behind the second column (shown by the yellow ellipse) was not detected by either the TOOD or DAB-DETR methods but was successfully detected by our method, again validating the effectiveness of our method. Compared to that of the TOOD method, the detection accuracy of our model is overall greater, thus far ahead of the effectiveness of the original detection algorithm.

Conclusion

In this paper, local and global information extraction of object pixels are enhanced from micro- and macroperspectives, respectively. The acquisition of image feature information is crucial to the performance of a model and is prone to misdetection and omission when detecting objects of different scales and occlusions in an image. Inspired by this problem, this paper proposes the use of the SCEFA module to enhance local feature extraction of multiscale channel information in high-level feature maps from a microscopic perspective. AFHTrans is presented from a macro perspective. First, this approach enhances the interaction between the feature information of each layer, maximizing the computational cost. Then, self-attention and convolution are combined to enhance global and local feature extraction. Finally, long-distance dependencies between object pixels are established. In terms of modelling, the big picture proposed FE-Head. This approach not only further enhances the feature extraction of the detected head but also improves the acquisition of full-text information. This method can accurately detect objects under complex conditions, such as different scales of objects, overlaps, and similar colours. Compared with other advanced methods, our method has the best detection effect. However, the methodology of this paper is still deficient. On the one hand, there is still a need to improve the accuracy of the detection of particularly small and over obscured objects. On the other hand, our model only tested on the PASCAL VOC2007 + 2012 and COCO2017 datasets. Moreover, the effect of the model is difficult to sustain in the face of a highly variable dataset. Therefore, for the method proposed in this paper, there is a need to continuously optimize our model and expand the generalization ability of the model to achieve superior results for training on this experimental dataset and additional datasets.

Acknowledgements The authors gratefully acknowledge the financial support from the National Natural Science Foundation of China (Grant Nos. 61472220, 61572286).

Author contributions Formal analysis, Zhenyi Zhang and Tianping Li; Methodology, Zhenyi Zhang and Tianping Li; Supervision, Dongmei Wei; Writing original draft, Tianping Li and Zhenyi Zhang.

Funding National Natural Science Foundation of China, Grant/Award Numbers: 61472220, 61572286.

Data availability Data related to the current study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors have no conflicts of interest in the publication of this paper.

References

1. Zou Z, Chen K, Shi Z et al (2023) Object Detection in 20 Years: A Survey. Proc IEEE 111:257–276. https://doi.org/10.1109/JPROC.2023.3238524
2. Pathak AR, Pandey M, Rautaray S (2018) Application of Deep Learning for Object Detection. Procedia Comput Sci 132:1706–1717. https://doi.org/10.1016/j.procs.2018.05.144
3. Arulprakash E, Aruldoss M (2022) A study on generic object detection with emphasis on future research directions. J King Saud Univ - Comput Inf Sci 34:7347–7365. https://doi.org/10.1016/j.jksuci.2021.08.001
4. Dhillon A, Verma GK (2020) Convolutional neural network: a review of models, methodologies and applications to object detection. Prog Artif Intell 9:85–112. https://doi.org/10.1007/s13748-019-00203-0
5. Vaidwan H, Seth N, Parihar AS, Singh K (2021) A study on transformer-based Object Detection. In: 2021 International Conference on Intelligent Technologies (CONIT). IEEE, Hubli, India, pp 1–6
6. Girshick R, Donahue J, Darrell T, Malik J Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In: arXiv preprint arXiv:1311.2524
7. Cai Z, Vasconcelos N (2018) Cascade R-CNN: Delving Into High Quality Object Detection. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, Salt Lake City, UT, pp 6154–6162
8. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You Only Look Once: Unified, Real-Time Object Detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Las Vegas, NV, USA, pp 779–788
9. Lin T-Y, Goyal P, Girshick R, et al Focal Loss for Dense Object Detection. In: arXiv preprint arXiv:1708.02002
10. Ren S, He K, Girshick R (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems. In: arXiv preprint arXiv:1506.01497
11. Bochkovskiy A, Wang C-Y, Liao H-YM (2020) YOLOv4: Optimal Speed and Accuracy of Object Detection. In: arXiv preprint arXiv:2004.10934
12. Tian Z, Shen C, Chen H, He T (2019) FCOS: Fully Convolutional One-Stage Object Detection. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, Seoul, Korea (South), pp 9626–9635
13. Zhang S, Chi C, Yao Y, et al (2020) Bridging the Gap Between Anchor-Based and Anchor-Free Detection via Adaptive Training Sample Selection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Seattle, WA, USA, pp 9756–9765
14. Liu Y, Zhang Y, Wang Y, et al (2023) A Survey of Visual Transformers. IEEE Trans Neural Netw Learn Syst 1–21. https://doi.org/10.1109/TNNLS.2022.3227717
15. Vedaldi A, Bischof H, Brox T, Frahm J-M (2020) End-to-End Object Detection with Transformers. In: Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I. Springer International Publishing, Cham.
16. Zhu X, Su W, Lu L, et al (2021) Deformable detr: Deformable transformers for end-to-end object detection. In: arXiv preprint arXiv:2010.04159
17. Vaswani A, Shazeer N, Parmar N, et al Attention is All you Need. In: arXiv preprint arXiv:1706.03762
18. Ivanov A, Dryden N, Ben-Nun T, et al Data Movement Is All You Need: A Case Study on Optimizing Transformers. In: arXiv preprint arXiv:2007.00072
19. Chen Y, Dai X, Chen D, et al (2022) Mobile-Former: Bridging MobileNet and Transformer. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, New Orleans, LA, USA, pp 5260–5269
20. Harjoseputro Y, Yuda IgnP, Danukusumo KP (2020) MobileNets: Efficient Convolutional Neural Network for Identification of Protected Birds. Int J Adv Sci Eng Inf Technol 10:2290. https://doi.org/10.18517/ijaseit.10.6.10948
21. Li K, Wang Y, Gao P, et al (2022) Uniformer: Unified transformer for efficient spatiotemporal representation learning. In: arXiv preprint arXiv:2201.04676

22. Lou M, Zhou H-Y, Yang S, Yu Y (2023) TransXNet: Learning Both Global and Local Dynamics with a Dual Dynamic Token Mixer for Visual Recognition. In: arXiv preprint arXiv:2310.19380
23. Sun Z, Cao S, Yang Y, Kitani K (2021) Rethinking Transformer-based Set Prediction for Object Detection. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, Montreal, QC, Canada, pp 3591–3600
24. Zhang H, Zu K, Lu J, et al (2023) EPSANet: An Efficient Pyramid Squeeze Attention Block on Convolutional Neural Network. In: Wang L, Gall J, Chin T-J, et al (eds) Computer Vision – ACCV 2022. Springer Nature Switzerland, Cham, pp 541–557
25. Zhang Q-L, Yang Y-B (2021) SA-Net: Shuffle Attention for Deep Convolutional Neural Networks. In: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, Toronto, ON, Canada, pp 2235–2239
26. Jiao J, Tang Y-M, Lin K-Y et al (2023) DilateFormer: Multi-Scale Dilated Transformer for Visual Recognition. IEEE Trans Multimed 25:8906–8919. https://doi.org/10.1109/TMM.2023.3243616
27. Lin T-Y, Dollar P, Girshick R, et al (2017) Feature Pyramid Networks for Object Detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Honolulu, HI, pp 936–944
28. Zhang W, Huang Z, Luo G, et al (2022) TopFormer: Token Pyramid Transformer for Mobile Semantic Segmentation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, New Orleans, LA, USA, pp 12073–12083
29. Feng C, Zhong Y, Gao Y, et al (2021) TOOD: Task-aligned One-stage Object Detection. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, Montreal, QC, Canada, pp 3490–3499
30. He K, Gkioxari G, Dollar P, Girshick R (2017) Mask R-CNN. In: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, Venice, pp 2980–2988
31. Gong Y, Xiao Z, Tan X et al (2020) Context-Aware Convolutional Neural Network for Object Detection in VHR Remote Sensing Imagery. IEEE Trans Geosci Remote Sens 58:34–44. https://doi.org/10.1109/TGRS.2019.2930246
32. Kim S-W, Kook H-K, Sun J-Y et al (2018) Parallel Feature Pyramid Network for Object Detection. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y (eds) Computer Vision – ECCV 2018. Springer International Publishing, Cham, pp 239–256
33. Liu W, Anguelov D, Erhan D, et al (2016) SSD: Single Shot MultiBox Detector. pp 21–37
34. Deng L, Yang M, Li T, et al (2019) RFBNet: Deep Multimodal Networks with Residual Fusion Blocks for RGB-D Semantic Segmentation. In: arXiv preprint arXiv:1907.00135
35. Liang T, Chu X, Liu Y et al (2022) CBNet: A Composite Backbone Network Architecture for Object Detection. IEEE Trans Image Process 31:6893–6906. https://doi.org/10.1109/TIP.2022.3216771
36. Law H, Deng J CornerNet: Detecting Objects as Paired Keypoints. In: arXiv preprint arXiv:1808.01244
37. Liu S, Qi L, Qin H, et al Path Aggregation Network for Instance Segmentation. In: arXiv preprint arXiv:1803.01534
38. Peng Z, Huang W, Gu S, et al (2021) Conformer: Local Features Coupling Global Representations for Visual Recognition. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, Montreal, QC, Canada, pp 357–366
39. Guo J, Han K, Wu H, et al (2022) CMT: Convolutional Neural Networks Meet Vision Transformers. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, New Orleans, LA, USA, pp 12165–12175
40. Li Y, Yao T, Pan Y, Mei T (2023) Contextual Transformer Networks for Visual Recognition. IEEE Trans Pattern Anal Mach Intell 45:1489–1500. https://doi.org/10.1109/TPAMI.2022.3164083
41. Li Y, Mao H, Girshick R, He K (2022) Exploring Plain Vision Transformer Backbones for Object Detection. In: Avidan S, Brostow G, Cissé M, et al (eds) Computer Vision – ECCV 2022. Springer Nature Switzerland, Cham, pp 280–296
42. Lin W, Wu Z, Chen J, et al Scale-Aware Modulation Meet Transformer. In: arXiv preprint arXiv:2307.08579
43. Fan Q, Huang H, Guan J, He R (2023) Rethinking Local Perception in Lightweight Vision Transformer. In: arXiv preprint arXiv:2303.17803
44. Jiang B, Luo R, Mao J et al (2018) Acquisition of Localization Confidence for Accurate Object Detection. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y (eds) Computer Vision – ECCV 2018. Springer International Publishing, Cham, pp 816–832
45. Wu Y, Chen Y, Yuan L, et al (2020) Rethinking Classification and Localization for Object Detection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Seattle, WA, USA, pp 10183–10192
46. Song G, Liu Y, Wang X (2020) Revisiting the Sibling Head in Object Detector. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Seattle, WA, USA, pp 11560–11569
47. Ge Z, Liu S, Wang F, et al (2021) YOLOX: Exceeding YOLO Series in 2021. In: arXiv preprint arXiv:2107.08430
48. Zhao Z, He C, Zhao G et al (2023) RA-YOLOX: Re-parameterization align decoupled head and novel label assignment scheme based on YOLOX. Pattern Recognit 140:109579. https://doi.org/10.1016/j.patcog.2023.109579
49. Qin J, Huang Y, Wen W (2020) Multi-scale feature fusion residual network for Single Image Super-Resolution. Neurocomputing 379:334–342. https://doi.org/10.1016/j.neucom.2019.10.076
50. Ma W, Wu Y, Cen F, Wang G (2020) MDFN: Multi-scale deep feature learning network for object detection. Pattern Recognit 100:107149. https://doi.org/10.1016/j.patcog.2019.107149
51. Li Y, Chen Y, Wang N, Zhang Z-X (2019) Scale-Aware Trident Networks for Object Detection. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, Seoul, Korea (South), pp 6053–6062
52. Li T, Wei Y, Liu M et al (2023) Refined Division Features Based on Transformer for Semantic Image Segmentation. Int J Intell Syst 2023:1–15. https://doi.org/10.1155/2023/6358162
53. Jang E, Gu S, Poole B. Categorical reparameterization with gumbel-softmax. In: arXiv preprint arXiv:1611.01144
54. Xu B, Wang N, Chen T, et al. Empirical evaluation of rectified activations in convolutional network. In: arXiv preprint arXiv:1505.00853
55. Hou Q, Zhou D, Feng J (2021) Coordinate Attention for Efficient Mobile Network Design. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Nashville, TN, USA, pp 13708–13717
56. Cao Y, Xu J, Lin S, et al (2019) GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). IEEE, Seoul, Korea (South), pp 1971–1980
57. Tanaka M (2020) Weighted sigmoid gate unit for an activation function of deep neural network. Pattern Recognit Lett 135:354–359. https://doi.org/10.1016/j.patrec.2020.05.017
58. Khan ZY, Niu Z (2021) CNN with depthwise separable convolutions and combined kernels for rating prediction. Expert Syst Appl 170:114528. https://doi.org/10.1016/j.eswa.2020.114528
59. Wei X, Zhang L, Zhang J et al (2024) Decoupled Sequential Detection Head for accurate acne detection. Knowl-Based Syst 284:111305. https://doi.org/10.1016/j.knosys.2023.111305
60. Everingham M, Van Gool L, Williams CKI et al (2010) The Pascal Visual Object Classes (VOC) Challenge. Int J Comput Vis 88:303–338. https://doi.org/10.1007/s11263-009-0275-4

61. Lin T-Y, Maire M, Belongie S et al (2014) Microsoft COCO: Common Objects in Context. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T (eds) Computer Vision – ECCV 2014. Springer International Publishing, Cham, pp 740–755
62. Rezatofighi H, Tsoi N, Gwak J, et al (2019) Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Long Beach, CA, USA, pp 658–666
63. Sun P, Zhang R, Jiang Y et al (2023) Sparse R-CNN: An End-to-End Framework for Object Detection. IEEE Trans Pattern Anal Mach Intell 45:15650–15664. https://doi.org/10.1109/TPAMI.2023.3292030
64. Wang Y, Zhang X, Yang T, Sun J (2022) Anchor DETR: Query Design for Transformer-Based Detector. Proc AAAI Conf Artif Intell 36:2567–2575. https://doi.org/10.1609/aaai.v36i3.20158
65. Liu S, Li F, Zhang H, et al (2022) DAB-DETR: Dynamic Anchor Boxes are Better Queries for DETR. In: arXiv preprint arXiv:2201.12329

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.