Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 20, No. 2, 2026

OverLoCK-GPH: A Bio-Inspired Object Detector with Graph-Prior Modulation and Hybrid Instance Refinement

Yuxi Han

Beijing University of Civil Engineering and Architecture, Beijing 102616, China

Abstract: MASK R-CNN is a visual model based on convolutional neural networks and applied to object detection. In the Mask R-CNN architecture, the Backbone typically employs ResNet. Through continuous convolution and downsampling, it extracts texture and semantic features of the image equally layer by layer, resulting in a large amount of background noise being mistaken for useful information, which interferes with the localization of the target. In addition, the Neck adopts a simple top-down additive fusion. This fusion is static and linear, and is limited by the local receptive field of the convolutional kernel, resulting in a lack of spatial relationships in FPN, incomplete object detection, and inaccurate localization. This paper proposes an enhanced detection framework named OverLoCK-GPH. Firstly, we utilize the Overview-Net of OverLoCK to generate a global context prior, and inject it into the features at each level through a novel prior-guided feature pyramid network, achieving dynamic weight modulation in space. Secondly, we introduce the Graph Attention Block at the high-level feature extraction stage, which captures long-range semantic dependencies by modeling pixels as graph nodes. Finally, we designed a Hybrid Instance Refinement Head for detection, which suppresses background noise at the ROI level through a channel attention mechanism. Experiments demonstrate that this method significantly outperforms the benchmark model in complex scenarios, effectively addressing the issues of missed and false detections of fuzzy targets.

Keywords: Object Detection; Mask R-CNN; OverLoCK; Prior-Guided Modulation; Graph Attention; Hybrid BBox Head.

1. Introduction

Object detection as a fundamental and challenging task in the field of computer vision, is widely applied in key areas such as autonomous driving [1], crack detection [2], and cell recognition [3].The core objective of this task is to accurately assign category labels to each object in the image and define the minimum bounding rectangle. With the development of Convolutional Neural Networks (CNN) [5], breakthroughs have been made in the performance of detectors, gradually leading to a technological landscape dominated by single-stage detectors [6-7] guided by anchor boxes or points, and multi-stage detectors [8-9] based on regions. In these tasks, region-based multi-stage detectors, represented by Mask R-CNN [4], have long been the preferred choice for handling complex instance-aware tasks in both industry and academia, thanks to their precise anchor box extraction and multi-task parallel processing mechanisms. Furthermore, deep optimization and improvement of their structures have gradually become one of the core research directions in the current object detection field.

The typical network structure of Mask R-CNN consists of a backbone network, a feature pyramid, a region proposal network (RPN), and an ROI head, which work together. Specifically, backbone networks (such as the ResNet[10] series) extract multi-scale features of images through layer-by-layer convolution; The feature pyramid achieves the fusion of multi-scale features through a top-down path and lateral connections, aiming to construct feature representations that combine high resolution and strong semantic information; The region proposal network is responsible for generating and pre-filtering candidate regions on the feature map; The ROI head achieves feature alignment through RoI Align, and then utilizes parallel branches to simultaneously complete object classification, bounding box regression, and pixel-level mask generation.

This paper observes that the classic Mask R-CNN architecture exhibits certain limitations when dealing with complex contextual logic. From the perspective of feature perception, the backbone network can only capture local receptive fields and lacks the ability to provide an overview of the global semantic information, making it prone to false detections in complex backgrounds. From the perspective of feature fusion, feature pyramids often adopt a static linear stacking approach. Although this method achieves the integration of multi-scale information, it lacks an explicit prior guidance mechanism and is constrained by the locality of convolution operations, making it difficult to capture long-range spatial dependencies between effective targets. From the perspective of feature purity, the ROI head typically directly utilizes fully connected layers to process aligned features, without actively removing background noise within the region of interest. This limits the model's discriminative accuracy in signal-to-noise ratio environments. In summary, there is a significant mismatch between Mask R-CNN's traditional local convolutional modeling paradigm and its high-performance universal detection task's demand for global semantics.

To address the aforementioned issues, this paper proposes an enhanced Mask R-CNN object detection model that integrates biologically-inspired priors [12] and graph-associated modulation [11], named OverLoCK-GPH. Due to the introduction of the top-down attention mechanism of the OverLoCK backbone network, OverLoCK-GPH is able to obtain initial features with stronger global perception compared to traditional ResNet. However, simple feature extraction is insufficient for achieving precise localization in complex contexts, and the locality of convolutional

operations still limits the network's ability to model long-range spatial relationships. To alleviate this issue, this paper proposes the Prior-Guided Graph FPN (GP-FPN), which dynamically modulates multi-scale features through global priors and combines graph attention mechanisms to achieve non-local semantic enhancement; Meanwhile, this paper designs a Hybrid Instance Refinement mechanism in the detection head part, which further removes background noise from ROI features through channel adaptive recalibration.

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

(1) This paper proposes an enhanced multi-stage object detection framework, OverLoCK-GPH. This framework breaks the limitation of traditional Mask R-CNN, which relies solely on bottom-up feature extraction. By introducing the global context prior generated by the OverLoCK backbone network to guide feature fusion, it achieves a detection paradigm of "overview first, then detailed examination", significantly improving the model's localization performance in complex scenes.

(2) This paper designs a Guided Prior Feature Pyramid Network (GP-FPN). This module innovatively injects global prior information into each level of the feature pyramid through a spatial modulation mechanism, and utilizes a graph attention mechanism to transform pixel-level features into topological node associations, effectively capturing long-range spatial dependencies between objects and fundamentally compensating for the limited receptive field of convolutional neural networks.

(3) This paper constructs a Hybrid BBox Head. Inspired by the idea of feature mixing, this module integrates a lightweight channel adaptive recalibration mechanism in the ROI head. By nonlinearly enhancing instance-level features, it ensures that the classification and regression branches can focus on discriminative feature components, effectively suppressing background noise interference within the region of interest.

(4) This paper conducted extensive experimental verification on the MS COCO public dataset. The experimental results show that, under the standard 1x training epoch, OverLoCK-GPH achieves a significant mAP improvement compared to the benchmark Mask R-CNN model. Meanwhile, experiments demonstrate that while maintaining inference efficiency, it synergistically enhances the purity and robustness of feature representation, fully verifying the progressiveness and practical value of the model proposed in this paper in high-performance object detection tasks.

2. Related Work

There are diverse approaches to implementing object detection, such as two-stage and single-stage methods based on Convolutional Neural Networks (CNN) [5], sequential detection methods based on Recurrent Neural Networks (RNN) [21], relationship modeling methods based on Graph Neural Networks (GNN) [22], and end-to-end detection methods based on Transformer [23]. Among them, CNN-based methods have achieved outstanding performance in both accuracy and efficiency due to their powerful feature learning and multi-scale modeling capabilities. Among CNN-based detection methods, the two-stage detection paradigm represented by R-CNN has been widely applied due to its decoupling of candidate region and classification regression, facilitating the introduction of tasks such as instance segmentation: Cascade R-CNN [17] enhances the recall and precision of high-quality object bounding boxes through multi-stage cascaded refinements. In 3D point cloud detection, Fast Point R-CNN [24] further improves detection accuracy through a dual-path representation of voxels and original point clouds, as well as a two-stage framework of "lightweight convolutional initial detection + attention-based fusion of point coordinates and convolutional features for further refinement". Fast R-CNN [14] reduces redundant computations by sharing convolutional feature maps and RoI pooling. Faster R-CNN [15] further introduces a Region Proposal Network (RPN) to achieve end-to-end training; Building upon this, Mask R-CNN [16] incorporates an instance segmentation branch into Faster R-CNN [15] and employs RoI Align to mitigate quantization errors, integrating detection and instance segmentation into a unified framework. It has become a commonly used benchmark for two-stage detection and instance segmentation.

Regarding Mask R-CNN [16], numerous works have been conducted to improve it from various perspectives. In BMask R-CNN [18], Cheng enhanced the mask localization accuracy by maintaining a boundary-preserving mask head and a feature fusion block within the head, enabling mutual learning between the object boundary and the mask; Wu's RGC Mask R-CNN [19] mitigates the estimation bias of BN under small batches and the sample scarcity and overfitting caused by high thresholds by introducing Group Normalization (GN) and three-level IoU cascade training, thereby enhancing the detection and segmentation performance of Mask R-CNN. Lin, through G-Mask [20], introduces Generalized Intersection over Union (GIoU) as the bounding box regression loss and achieves superior performance by employing ResNet-101, RPN, RoI Align, and FCN mask head within the Mask R-CNN framework. In terms of more general improvement directions, lightweight convolutional backbones and backbones equipped with global priors or attention mechanisms are widely adopted to reduce computation while maintaining accuracy. In the neck region, FPN integrates multi-scale features through a top-down path and lateral connections. Subsequent work introduces attention or graph structures on this basis to enhance top-level semantics. In the detection head, the practice of applying channel attention or feature modulation on RoI features to improve discriminability has also been applied. However, research that unifies the "lightweight backbone - multi-scale prior modulation - graph attention neck - channel recalibration detection head" within the same two-stage framework, achieving similar accuracy to existing methods with fewer parameters, lower computation, and shorter training periods, remains relatively limited.

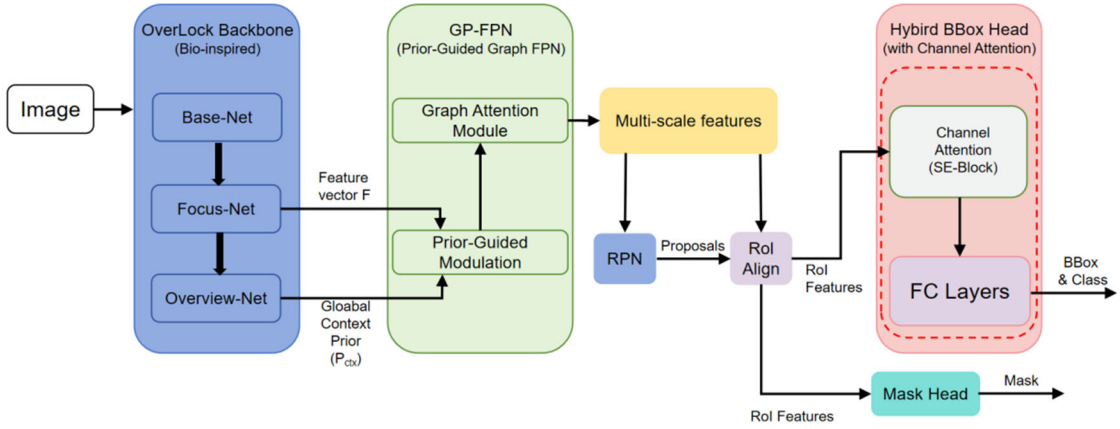
The proposed OverLoCK-GPH still follows the two-stage detection and instance segmentation framework of Mask R-CNN. To address the aforementioned shortcomings, improvements are made in the backbone, neck, and detection head. In the backbone, OverLoCK pretrained on ImageNet is used, providing multi-scale features while outputting global context priors. In the neck, a Prior-Guided Graph Feature Pyramid (GP-FPN) is designed, which spatially modulates the features at each layer using this prior on top of a standard FPN (PGSM), and introduces Graph Association Modulation (GAM) at the top layer to model long-range semantic relationships. In the detection head, a Squeeze-and-Excitation-based channel attention module is introduced to form a Hybrid Fully Connected Bounding Box Head (Hybrid FC BBox Head), which recalibrates RoI features along the

channel dimension before performing classification and regression. Overall, after the backbone extracts multi-scale features and contextual priors, GP-FPN performs multi-scale fusion and enhances the features through prior modulation and graph attention, making them available for RPN and RoI extraction. Finally, the hybrid detection head outputs bounding boxes and instance masks. Compared with methods like RetinaNet, Faster R-CNN, and DETR, which require longer training or larger backbones, OverLoCK-GPH achieves comparable detection and segmentation accuracy with fewer parameters, similar or lower computational cost, and shorter training cycles, demonstrating a balance between efficiency and performance in the proposed backbone, neck, and detection head design.

3. Methodology

3.1. OverLoCK-GPH Network Structure

The OverLoCK-GPH network structure proposed in this paper is shown in Figure 1.



Finger 1. The OverLoCK-GPH network structure

While inheriting the classic two-stage detection paradigm of Mask R-CNN, this model addresses the issues of global prior knowledge deficiency and insufficient feature purity in complex scenes. It constructs a deep perception system composed of the OverLoCK backbone network, Prior Guided Feature Pyramid Network (GP-FPN), Region Proposal Network (RPN), and Hybrid BBox Head. Specifically, the backbone network OverLoCK is responsible for preliminary feature encoding. Its input is the original image： $I \in \mathbb{R}^{3\times H\times W}$. Through the internally coordinated three sub-networks: Base-Net, Overview-Net, and Focus-Net, it simulates the biologically inspired mechanism of human vision, which involves "first taking a general overview, then focusing on details". The role of OverLoCK is to break the bottleneck of convolutional neural networks being limited to local receptive fields. The first two stages output local features C_2 and C_3, while the high-level stage generates global context features from the context branch while producing main branch features. These features are then fused by the fusion sub-network to obtain C_4 and C_5. The backbone ultimately outputs a multi-scale local feature set:

$$F = C_2, C_3, C_4, C_5$$

(the downsampling factor of C_i is 2^i).

Simultaneously, a global context prior (Context Prior) P_{ctx}, which incorporates salient information from the entire image, is explicitly generated to provide scene-level prior guidance for subsequent layers.

Subsequently, the Guided Pyramid of Features with Prior (GP-FPN) receives F and P_{ctx} as inputs, and applies standard FPN to obtain five layers of features $P2, P3, P4, P5, P6$ from F. Then, for each layer P_i, P_{ctx} is used for spatial weight modulation: P_{ctx} is bilinearly interpolated to the spatial size of P_i, mapped to a single-channel spatial weight through an independent 1×1 convolution W_i of that layer, and multiplied element-wise with P_i. The aim is to enhance feature representation through global information modulation and spatial correlation modeling. In this module, each layer of convolutional features C_i first undergoes adaptive spatial modulation under the guidance of P_{ctx} to suppress interference noise in the background region and enhance target response. Its mathematical expression is as follows:

$$P_i' = P_i \odot \sigma\left(W_i \cdot Interpolate\left(P_{ctx}, size(P_i)\right)\right)$$

Among them, $\odot$ denotes element-wise multiplication, and σ represents Sigmoid. To further compensate for the semantic loss caused by the static linear stacking in traditional FPN, GP-FPN introduces a graph attention enhancement branch on the modulated P_5 to enhance top-level semantics and long-range dependencies. This branch maps the spatial dimension in the features to a sequence of graph nodes, employs multi-head self-attention for global association modeling, and achieves non-local semantic reasoning by calculating self-attention associations between nodes, that is:

$$X_{graph} = LayerNorm\left(X + Attention(Q, K, V)\right)$$

among which:

$$Q = K = V = X$$

$$Attention(Q, K, V) = Softmax\left(\frac{Q(K)^T}{\sqrt{d}}\right)V$$

This mechanism enables the model to establish non-local semantic relationships at the P_5 scale, and inputs the modulated and graph-enhanced features into the RPN to generate candidate boxes. The fixed-size instance feature block X_{roi} is obtained through RoIAlign. Simultaneously, to address the interference of residual background noise within the RoI region on the classifier, this paper designs a Hybrid BBox Head. Before X_{roi} enters the fully connected layer, a channel recalibration mechanism based on the SE-Block structure is utilized to purify its features: a channel descriptor vector $z \in \mathbb{R}^{1\times C\times 1\times 1}$ is obtained by aggregating spatial information through global average pooling.

$$z = GlobalAvgPool(Xroi)$$

Then, two fully connected layers are used to capture the nonlinear dependencies between channels, resulting in excitation weights s

$$s = \sigma\big(W_2\delta(W_1 z)\big)$$

Where δ represents the ReLU activation function, σ denotes the Sigmoid function, and W_1 and W_2 stand for learning weights. Ultimately, the refined features $X_{roi} \odot s$ are fed into the prediction branch. This design enhances the robustness of the detection head by strengthening discriminative channels and suppressing noisy channels, while maintaining the same RPN and RoI processes as Mask R-CNN. In summary, OverLoCK-GPH achieves full-process optimization from scene understanding to instance perception by outputting scene-level priors at the backbone network, performing prior-guided modulation and P_5 map attention at the neck, and channel refinement at the head. This makes it more robust and accurate than the native Mask R-CNN on the MS COCO object detection task.

3.2. Backbone OverLock and Global Context Prior Generation

This paper adopts OverLoCK [12] as the backbone network to address the limitations of traditional convolutional neural networks (such as ResNet) in bottom-up stacking, which are constrained by local receptive fields and lack global understanding of complex scenes. OverLoCK mimics the cognitive mechanism of the human visual system, which involves "overview first, then look closely". Through the Deep-stage Decomposition Strategy (DDS), it decouples the feature extraction process into three collaborative subnetworks: Base-Net, Overview-Net, and Focus-Net. Its core lies in explicitly generating global context P_{ctx} and injecting it into the local feature extraction process through Context-Mixing Dynamic Convolution, achieving "guiding local convolution with global prior knowledge".

3.2.1. Deep stage decomposition and prior generation

For a given input image $I \in \mathbb{R}^{3\times H\times W}$, OverLoCK obtains multi-scale features and global priors through four stages defined by embedding layers and stacked blocks. The correspondence between its data flow and DDS is as follows.

Base-Net: Low- and mid-level feature encoding. The first two stages constitute the main body of Base-Net, performing progressive downsampling and local perception on the input. After passing through the stem, two downsampling embeddings, and their corresponding RepConv blocks, We obtain feature maps with resolutions of $H/4 \times \mathrm{W}/4$ and $H/8 \times \mathrm{W}/8$, denoted as C_2 and C_3, respectively, they correspond to the outputs of Base-Net at $H/4$ and $H/8$ in DDS. The third stage further downsamples the features to $H/16 \times W/16$, obtaining the mid-level feature map F_{mid} This mid-level feature serves as a "bifurcation point" and is simultaneously input to both Overview-Net and Focus-Net.

Overview-Net: Global Context Prior. To simulate the rapid overview capability of the human eye, F_{mid} is fed into a lightweight Overview-Net (the fourth embedding layer patch_embed4 and blocks4). This sub-network obtains a highly condensed semantic global context prior at a resolution of $H/32 \times W/32$ through further downsampling and global information aggregation: P_{ctx}

$$P_{ctx} = \Phi overview(F_{mid}) \in \mathbb{R}^{C_{ctx}\times\frac{H}{32}\times\frac{W}{32}},$$

Here, $\Phi overview$ represents the mapping of Overview-Net, and C_{ctx} is the output of the backbone. P_{ctx} explicitly encodes the information of the entire image, serving as the scene-level prior guidance for the subsequent Focus-Net and neck (GP-FPN).

Focus-Net: Prior-Guided Refined Features. Under the guidance of P_{ctx}, it performs refined modeling on mid-level features and expands the receptive field. Its inputs include the main branch features continued by F_{mid}, as well as the initial prior P_0 obtained through channel compression and upsampling. Focus-Net consists of several Dynamic Blocks, where prior fusion at both feature and weight levels occurs simultaneously: at the feature level, the current feature and current prior are concatenated in the channel dimension and modulated by a gating mechanism; at the weight level, the prior participates in the generation of dynamic convolution kernels through ContMix, enabling local convolution to have long-range dependency modeling capabilities. The output of each block is split into updated features and updated priors, which are then fused with the initial prior P_0 through learnable weighting to avoid dilution of the prior in deeper layers. Focus-Net ultimately outputs two high-level feature streams at $H/16$ and $H/32$ resolutions, which, together with C_2 and C_3 of Base-Net, form a four-stage multi-scale feature set:

$$\mathcal{F} = \{C2, C3, C4, C5\}$$

and explicitly outputs P_{ctx} for use by the neck.

3.2.2. Contextual Mixed Dynamic Convolution (ContMix)

To transform the global prior P_{ctx} into modulation capability for local features, OverLoCK introduces ContMix into the dynamic blocks of Focus-Net. This operator generates dynamic convolution kernels on a token-wise basis through the affinity between "local features and global prior", injecting global semantics while maintaining the local inductive bias of convolution. Specifically, the computation of ContMix is divided into two steps: first, establish an affinity representation between each spatial location and the global prior, and then generate dynamic convolution kernels on a token-wise basis based on this representation to convolve with local features, thus achieving global context mixing on a token-by-token basis. Below are formal descriptions of the two steps respectively.

Token-wise global context representation: Let the main branch feature within the current block be $X \in \mathbb{R}^{C\times H\times W}$ and the current context prior be $P \in \mathbb{R}^{Cp\times H\times W}$. ContMix maps X to a query matrix Q, aggregates P into $S \times S$ region centers via adaptive pooling, and then maps them to a key matrix K.

$$Q = W_q(X) \in \mathbb{R}^{C\times H\times W}$$

$$K = W_k(AvgPool_s(P)) \in \mathbb{R}^{C\times S^2}$$

Where W_k and W_q are (1×1) convolutions, and $AvgPool_s$ denotes pooling P to $S \times S$. After channels are grouped by the number of heads, the affinity matrix is calculated for each group.

Global context blending per token: Each row of affinities is aggregated into a S×S convolution kernel weight through a learnable linear layer ($W_d \in \mathbb{R}^{S^2\times K^2}$), and then normalized by Softmax to obtain a dynamic kernel per position:

$$D_g = Softmax(A_g W_d) \in \mathbb{R}^{HW\times K^2}$$

Ultimately, the output of ContMix is the result of performing convolution bitwise using these dynamic kernels X, ensuring that the convolution kernel at each position carries global information encoded by prior knowledge.

Through the aforementioned mechanism, Focus-Net is able to perceive distant context when extracting local details,

providing more robust multi-scale features and explicit priors P_{ctx} for subsequent GP-FPN and detection heads.

3.3. Guided Prior Feature Pyramid Networks (GP-FPN)

After obtaining significant prior information with a "global overview" perspective generated by the OverLoCK backbone network, this paper designs the Prior-Guided Graph Feature Pyramid Network (GP-FPN), aiming to achieve efficient fusion of multi-scale features and deep enhancement of spatial semantics. The input of GP-FPN consists of two parts: a local feature set F from different stages of the backbone network, and a global context prior P_{ctx} generated by the Overview-Net. Unlike traditional feature pyramid networks (FPNs) that only perform feature stacking through convolution and linear upsampling, GP-FPN introduces, on the basis of the standard FPN's five-layer pyramid, the Prior-Guided Spatial Modulation (PGSM) and the hierarchical Graph Association Modulation (GAM) modules. This design enables each layer of features to not only possess high-resolution boundary details but also carry deep semantic information imparted by the biologically-inspired global prior and graph association mechanism from OverLoCK, thus laying a solid theoretical and representational foundation for the subsequent Region Proposal Network (RPN) to generate accurate proposal boxes and the Hybrid Head to perform final feature purification.

The forward process of GP-FPN is divided into three steps. The first step involves feeding only the multi-scale features F into the standard FPN. Assuming the lateral features corresponding to the ith layer $(i = 2, \dots, 6)$ are $\mathcal{L}i$, in the top-down path, the high-level features are upsampled and element-wise added to the lateral output of the current layer, followed by 3×3 convolution smoothing. The fused output can be expressed as

$$P_i = Conv_{3\times3}(\mathcal{L}i \oplus Up(P_{i+1}))$$

Where $\oplus$ denotes element-wise addition, and Up represents upsampling. This results in a five-layer pyramid of features. In the second step, perform Prior Guided Spatial Modulation (PGSM) on each of the five layers. In the third step, apply Graph-Affine Modulation (GAM) only to the modulated top-layer features, while keeping the other four layers unchanged. Finally, GP-FPN outputs a multi-scale feature set P for use by RPN and RoI Align.

The core logic of PGSM lies in utilizing global prior information as a spatial attention mask to explicitly suppress noise in complex backgrounds. Specifically, the model first compresses the channels of P_{ctx} through a lightweight convolutional layer Ψ_i (implemented as independent 1×1 convolutions for each layer, mapping the C_{ctx} channels of P_{ctx} to a single channel), and aligns it to the spatial dimensions of each layer's features P_i using a bilinear interpolation operator Up_i, generating a scale-specific modulation weight map

$$M_i = \sigma(\Psi_i(Up_i(P_{ctx})))$$

where σ is a Sigmoid activation function; in implementation, P_{ctx} is first bilinearly interpolated to the spatial size of P_i, and then passed through the layer's (1×1) convolution and Sigmoid. Subsequently, the modulated feature representation is denoted as

$$P_i' = P_i \otimes M_i$$

where $\otimes$ denotes element-wise multiplication. Through this explicit spatial calibration, GP-FPN can effectively "illuminate" potential target areas from multi-scale feature maps, significantly enhancing the model's target discrimination ability in low-contrast scenarios .

Meanwhile, to compensate for the limitations of convolutional operators in capturing long-range topological associations and non-local semantics between targets, GP-FPN embeds a hierarchical Graph Association Modulation (GAM) module in the top-level feature path. This module abstracts the modulated highest-level features into an undirected graph structure $G = (V, \mathcal{E})$: each pixel on the feature map is mapped to a set of graph nodes V, and its node features correspond to a sequence representation of $\mathbb{R}^{N\times D}$, where N is the number of nodes and $D = C$ is the feature dimension. To achieve complex semantic reasoning across regions while maintaining computational efficiency, GAM employs an attention mechanism based on Graph Transformer. For a given input node sequence X_{in}, its transformation process can be defined as

$$X_{out} = LayerNorm(X_{in} + Softmax(\frac{(X_{in}W_Q)(X_{in}W_K)^{\top}}{\sqrt{d}})(X_{in}W_V))$$

where W_Q, W_K, W_V are learnable query, key, and value mapping matrices, respectively, and d represents the head dimension.

X_{out} is reshaped into a spatial feature map with shape $[B, C, H_5, W_5]$, which yields the top-level output P_5 of the GAM. This graph association mechanism breaks the local receptive field limitation of Euclidean space, allowing the model to establish logical connections between target objects and their environmental backgrounds, as well as between the components of heterogeneous targets, across the entire graph. This greatly improves the detection robustness of long-scale objects and sparsely distributed targets.

GP-FPN takes backbone multi-scale features $\mathcal{F}$ and global priors P_{ctx} as inputs. It first obtains $\{P_2, \dots, P_6\}$ through a standard FPN, then performs prior-guided spatial modulation on each layer using PGSM to obtain $\{P'_2, \dots, P'_6\}$. Finally, only $\{P'_5\}$ is subjected to GAM to obtain $\{P''_5\}$, and the output is $P = \{P'_2, P'_3, P'_4, P''_5, P'_6\}$.

3.4. Hybrid BBox Head & Multi-task Loss

After the RPN and RoI Align operators extract fixed-size RoI features from the feature map P output by GP-FPN, the detection head aims to perform accurate classification and bounding box regression on each candidate region. This paper constructs a Hybrid BBox Head, whose core innovation lies in the introduction of a channel attention module based on the Squeeze-and-Excitation (SE) mechanism before entering the shared fully connected layer, which dynamically recalibrates the RoI features. The enhanced features are then fed into a Shared Dual-FC, and the final predictions are output through parallel classification and regression branches. The following sections sequentially present the input conventions and overall forward flow of the Hybrid BBox Head, a formal description of channel attention, shared fully connected layers and classification/regression branches, bounding box encoding and decoding, as well as multi-task losses.

3.4.1. Input and forward process

The candidate regions output by the RPN and the output P from GP-FPN are inputted into RoI Align together; RoI Align crops and resamples from P according to each candidate box to obtain the RoI features $\mathrm{X_{roi}} \in \mathrm{R^{N\times C\times H_r\times W_r}}$. Here, N is the

number of RoIs in the current batch, C is the number of output channels of the neck, which is consistent with the channel dimension of P, and $(\mathrm{H_r} = \mathrm{W_r})$ represents the spatial size of the RoI. In the forward propagation stage, our method first recalibrates the original features through a channel attention module; subsequently, the recalibrated features undergo flattening processing and are mapped by a two-layer shared fully connected network, transforming into high-order shared representations. Finally, this representation is simultaneously fed into parallel detection heads, which calculate the class probability distribution and position offset vectors, respectively, and generate the final detection targets in conjunction with post-processing operators such as Non-Maximum Suppression (NMS).

3.4.2. Squeeze-and-Excitation

The channel attention module performs global aggregation on the RoI features in the spatial dimension to obtain channel descriptors. Then, it generates channel weights through a lightweight two-layer MLP and Sigmoid, and scales the original features channel by channel, thereby highlighting channels that are beneficial for the current RoI discrimination and suppressing redundant or noisy channels. Its computation is divided into three steps: Squeeze, Excitation, and Scale.

Squeeze: Perform global average pooling on each channel $c \in \{1, \ldots, C\}$, compressing the spatial dimension into a single scalar to obtain a channel-level statistical vector:

$$(\boldsymbol{z} \in \mathbb{R}^C): \ \boldsymbol{z}_c = \frac{1}{H_r \times W_r} \sum_{i=1}^{H_r} \sum_{j=1}^{W_r} [X_{roi}]_{c,i,j}$$

Excitation: The dependency relationship between channels is learned through two fully connected and nonlinear activation layers, and mapped to weights in the range of $[0,1]$. With a compression ratio of S and an intermediate layer dimension of $C_{mid} = max(C/r, 16)$, we have:

$$s = \sigma(W_2\, \delta(W_1 z + b_1) + b_2)$$

where $W_1 \in \mathbb{R}^{C_{mid} \times C}$, $W_2 \in \mathbb{R}^{C_{mid} \times C}$ are learnable weights, δ represents ReLU activation, and σ represents Sigmoid. $s \in \mathbb{R}^C$ represents the channel weight vector.

Scale: After broadcasting s along the spatial dimension, multiply it element by element with X_{roi} to obtain the RoI features after channel recalibration:

$$X'_{roi} = X_{roi} \otimes s$$

Where $\otimes$ denotes the channel-wise multiplication broadcast along the spatial dimension. So far, the channel attention has achieved adaptive recalibration of the channel dimension of the RoI features without changing the size of the feature space, providing more discriminative inputs for the subsequent fully connected layer.

3.4.3. Shared fully connected layer and classification and regression branch

Flatten X'_{roi} along the spatial dimension to obtain $x \in \mathbb{R}^{N \times (C \cdot H_r \cdot W_r)}$, and then apply a two-layer shared fully connected network to obtain a shared representation $h \in \mathbb{R}^{N \times F}$, where

$$h = W_h^2 \delta(W_h^1 x + b_h^1) + b_h^2$$

Where $W_h^1 \in \mathbb{R}^{N \times (C \cdot H_r \cdot W_r)}, W_h^2 \in \mathbb{R}^{N \times F}$, δ denotes ReLU.

This shared representation is fed into both the classification branch and the regression branch simultaneously.

In the classification branch, the classification layer outputs K category logits (unnormalized scores) for each RoI: $o_{cls} = hW_{cls} + b_{cls}$, $\in \mathbb{R}^{N \times F}$, where the number of categories is K, Wcls is the weight of the classification layer, and $\in \mathbb{R}^{N \times F}$. During training, cross-entropy loss is used to supervise the comparison between o_{cls} and the true category labels.

When using class-related regression in the regression branch, each class corresponds to a set of bounding box offsets. The regression layer outputs $(o_{reg} \in \mathbb{R}^{N \times (4K)})$, which means 4-dimensional offsets $(\Delta x, \Delta y, \Delta w, \Delta h)$ for each class:

$o_{reg} = hW_{reg} + b_{reg}, W_{reg} \in \mathbb{R}^{F \times 4K}$

During training, L1 loss is used to supervise the predicted offset and the encoded true offset. The physical coordinates of the bounding box are encoded by Delta in the following text, and the rule is obtained by decoding the candidate box and o_{reg}.

3.4.4. Boundary box encoding and decoding

To stabilize the regression scale, the detection head predicts normalized offsets relative to the candidate box rather than absolute coordinates. Assuming the candidate box is (x_a, y_a, w_a, h_a) and the ground truth box is (x_b, y_b, w_b, h_b), the encoding targets are:

$$\Delta x = (x_b - x_a)/w_a\,, \Delta y = (y_b - y_a)/h_a$$
$$\Delta w = log^{(w_b/w_a)}\,, \Delta h = log^{(h_b/h_a)}$$

During training, the target offsets are normalized by mean and standard deviation, meaning that the ground truth labels are $(\Delta x/0.1, \Delta y/0.1, \Delta w/0.2, \Delta h/0.2)$ when calculating the loss, and the predicted values are directly output by the regression layer. During inference, the decoded values are

$$\hat{x} = x_a + w_a(\widehat{\Delta x} \cdot 0.1)$$
$$\hat{y} = y_b + h_b(\widehat{\Delta y} \cdot 0.1)$$
$$\hat{w} = w_a \cdot exp\big(\widehat{\Delta w} \cdot 0.2\big)$$
$$\hat{h} = h_a \cdot exp(\widehat{\Delta h} \cdot 0.2)$$

where $\Delta x, \Delta y, \Delta w, and\ \Delta h$ are the 4-dimensional outputs of the regression branch corresponding to the respective categories.

In summary, the Hybrid BBox Head achieves channel dimension recalibration of RoI features with extremely low parameter overhead. This mechanism works in conjunction with the Spatial Modulation and Graph-associated Modulation of GP-FPN to construct a multi-dimensional feature enhancement system, significantly improving the model's localization accuracy and classification robustness in complex scenarios.

4. Experiment

4.1. Experimental setup and dataset introduction

This paper conducts extensive experiments on the proposed model using the mmdetection detection library on the MS COCO2017 dataset.

Following common practice, this paper trains network parameters using a training set containing 118K images and a validation set of 5K images. In our experiments, GeForce RTX 4090 GPUs were used for parallel training, and CUDA acceleration was configured for computation. The training strategy applied in this paper includes: using AdamW as the optimizer to train network parameters, with a training duration of 12 epochs, an initial learning rate of 2×10^{-4}, a batch size set to 6, a weight decay of 0.05, and a learning rate decay strategy at stages 8 and 11 epochs. This paper employs standard data augmentation strategies, including random flipping, random cropping, multi-scale training, etc. The training image size is set to 800 on the short side, with a multi-

scale range of [800~1333] on the long side. The validation and test image sizes are both set to 1333×800. It should be noted that, based on the main experiment, this paper can extend the training duration to 36 epochs and adjust the learning rate decay nodes, batch size, and other configurations as needed for comparative experiments.

4.2. Network hyperparameter configuration

This paper adopts the OverLoCK network pre-trained as the backbone network for Mask R-CNN. Referring to the settings of commonly used detection frameworks, this paper uses a Feature Pyramid Network with Graph Attention as the neck, which outputs feature layers at five scales; the number of graph attention heads is set to 8. The detection head employs a Hybrid Fully Connected Bounding Box Head, with an input channel count of 256, a fully connected hidden layer dimension of 1024, a Region of Interest (RoI) feature space size of 7×7, and a channel compression ratio of 16 for the Squeeze-and-Excitation (SE) module. The bounding box encoding adopts the DeltaXYWH format, and the standard deviation of regression targets is set to (0.1, 0.1, 0.2, 0.2). The number of categories is 80, consistent with the COCO dataset. In terms of loss functions, the Region Proposal Network (RPN) classification loss and bounding box regression loss are respectively represented by L_{rpn_cls} and L_{rpn_bbox}. The Region of Interest (RoI) classification branch employs cross-entropy loss L_{cls}, the RoI bounding box regression employs L1 loss L_{bbox}, and the RoI instance segmentation branch employs cross-entropy loss L_{mask}. The weights of all three are set to 1.0, and the total loss is:

$$L_{loss} = L_{rpn\_cls} + L_{rpn\_bbox} + L_{cls} + L_{bbox} + L_{mask}$$

4.3. Evaluation indicators

This article adopts the standard Average Precision (AP) as the evaluation metric for model performance, and reports AP at different intersection-over-union (IoU) thresholds, including AP50 and AP75. Among them, mAP denotes the mean of AP over IoU thresholds from 0.50 to 0.95 with a step size of 0.05; AP50 and AP75 denote the AP at IoU thresholds of 0.50 and 0.75, respectively. This article reports the above indicators for both bounding box detection (bbox) and instance segmentation (segm) to characterize the model's performance in detection and segmentation. In addition, the number of training cycles (epochs), the total number of network parameters (Params), and the computational complexity in billions of floating-point operations (GFLOPs) of the proposed model are reported to describe its training cost and efficiency.

4.4. Performance

In order to verify the effectiveness and efficiency of the proposed method, experiments were conducted on the COCO 2017 validation set in this paper. The experiment uses OverLoCK as the backbone network, a Feature Pyramid Network (FPN) with graph attention as the neck, and a hybrid fully connected detection head; Set the input resolution to 800 × 1333, and the rest of the settings are consistent with sections 4.1 and 4.2. The experimental results are as follows: In terms of bounding box detection (bbox), this method achieved 39.3 mAP, 59.1 AP50, and 42.7 AP75; In terms of instance segmentation (SEGM), 37.6 mAP, 57.5 AP50, and 40.2 AP75 were achieved. The model has a parameter count of 36.07M and a computational complexity of 187.64 GFLOPs, reflecting a lightweight design. On the premise of fewer parameters, controllable computational complexity, and shorter training period, this method still maintains high detection and segmentation accuracy, achieving a good balance between lightweight and accuracy. Overall, the method proposed in this article achieves detection and segmentation performance comparable to common advanced models with fewer parameters, moderate computational complexity, and shorter training time, verifying the effectiveness, efficiency, and practicality of the proposed method.

5. Conclusions

This article proposes a lightweight object detection and instance segmentation model: using OverLoCK as the backbone, graph attention feature pyramid network (FPN) as the neck, and a hybrid fully connected detection head. This design utilizes the collaboration of lightweight backbone, graph attention FPN, and hybrid detection head to maintain multi-scale feature expression and detection capabilities while reducing model parameter count and computational overhead. It also helps to shorten training cycles and accelerate convergence. The experimental results show that, while achieving detection and segmentation performance comparable to common settings, this model has significant advantages in parameter quantity, computational complexity, and required training epochs, and has good application potential in resource limited or deployment efficiency sensitive scenarios. However, under the same training settings, the detection and segmentation accuracy of this model is still inferior to advanced methods that use larger backbones or longer training cycles, which is related to the insufficient modeling of target secondary information by graph attention mechanisms and hybrid heads. Future work will focus on more comprehensive representation of target features and more efficient utilization of multi-scale information to further enhance the performance of lightweight detection and segmentation models.

References

[1] Feng D, Harakeh A, Waslander S L, et al. A review and comparative study on probabilistic object detection in autonomous driving[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(8): 9961-9980.

[2] Yao H, Liu Y, Li X, et al. A detection method for pavement cracks combining object detection and attention mechanism[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(11): 22179-22189.

[3] Waithe D, Brown J M, Reglinski K, et al. Object detection networks and augmented reality for cellular detection in fluorescence microscopy[J]. Journal of Cell Biology, 2020, 219(10): e201903166.

[4] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.

[5] Purwono P, Ma'arif A, Rahmaniar W, et al. Understanding of convolutional neural network (cnn): A review[J]. International Journal of Robotics and Control Systems, 2022, 2(4): 739-748.

[6] Li C, Li L, Jiang H, et al. YOLOv6: A single-stage object detection framework for industrial applications[J]. arXiv preprint arXiv:2209.02976, 2022.

[7] Ren J, Chen X, Liu J, et al. Accurate single stage detector using recurrent rolling convolution[C]//Proceedings of the IEEE

conference on computer vision and pattern recognition. 2017: 5420-5428.

[8] Liao G, Gao W, Jiang Q, et al. Mmnet: Multi-stage and multi-scale fusion network for rgb-d salient object detection[C]//Proceedings of the 28th ACM international conference on multimedia. 2020: 2436-2444.

[9] Ouyang W, Luo P, Zeng X, et al. Deepid-net: multi-stage and deformable deep convolutional neural networks for object detection[J]. arXiv preprint arXiv:1409.3505, 2014.

[10] Koonce B. ResNet 50[M]//Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization. Berkeley, CA: Apress, 2021: 63-72.

[11] Wu Minqi, Yang Yuanhua, Li Hang, etc Lightweight Underwater Small Target Detection Based on Graph Transformer and RT-DETR [J/OL]. Computer Applications, 1-12 [2026-02-16] https://link.cnki.net/urlid/51.1307.TP.20251030.1441.004.

[12] Lou M, Yu Y. Overlock: An overview-first-look-closely-next convnet with context-mixing dynamic kernels[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2025: 128-138.

[13] Sherstinsky A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network[J]. Physica d: Nonlinear phenomena, 2020, 404: 132306.

[14] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[15] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(6): 1137-1149.

[16] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.

[17] Cai Z, Vasconcelos N. Cascade r-cnn: Delving into high quality object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6154-6162.

[18] Cheng T, Wang X, Huang L, et al. Boundary-preserving mask r-cnn[C]//European conference on computer vision. Cham: Springer International Publishing, 2020: 660-676.

[19] Wu M, Yue H, Wang J, et al. Object detection based on RGC mask R-CNN[J]. IET Image Processing, 2020, 14(8): 1502-1508.

[20] Lin K, Zhao H, Lv J, et al. Face Detection and Segmentation Based on Improved Mask R-CNN[J]. Discrete dynamics in nature and society, 2020, 2020(1): 9242917.

[21] Sherstinsky A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network[J]. Physica d: Nonlinear phenomena, 2020, 404: 132306.

[22] Scarselli F, Gori M, Tsoi A C, et al. The graph neural network model[J]. IEEE transactions on neural networks, 2008, 20(1): 61-80.

[23] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[24] Chen Y, Liu S, Shen X, et al. Fast point r-cnn[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 9775-9784.