ISSN: 2688-8645

D-GPH: Dynamic Graph-Prior Hybrid Detector with Uncertainty-Aware Refinement

Yuxi Han

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

Abstract

As a classic paradigm of two-stage target detection, Mask R-CNN achieves robust instance awareness through regional proposal and feature alignment. In its structure, the backbone network is responsible for extracting multi-scale features from the image layer by layer, and usually uses convolutional networks such as RESNET, which is difficult to use the global context prior to suppress the background noise in the degraded environment; Neck network (neck) relies on fixed multi-scale fusion scheme, lacks the ability of adaptive expression for different image contents, and is limited by the local receptive field of convolution operator, so its global modeling is easily disturbed by uncorrelated long-range targets; The detection head carries out category determination and boundary box regression on the candidate region (ROI), and usually uses the same set of forward and loss for difficult samples and easy samples. It is not enough to distinguish the fuzzy target with low confidence and small area, which is prone to missed detection and false detection. To solve these problems, this paper proposes a dynamic graph prior hybrid detection framework and named it D-GPH. Based on the global context prior generated by the backbone, the framework innovatively introduces the dynamic prior router (DPR), which generates adaptive prior injection strength for each layer according to the current image content, and realizes dynamic modulation by layer and graph before FPN fusion. In the neck design, this paper constructs Manhattan constraint graph attention module (MC-GAT), which explicitly suppresses non local noise by introducing spatial distance penalty, and cooperatively extracts global topology and local texture details with heterogeneous fusion layer. In the detection stage, this paper designed the uncertain perceptual refining head (UAR head), established the quality evaluation mechanism by using the exponential moving average (EMA) of varifocal loss and IOU, and started the secondary refining prediction based on channel recalibration for difficult samples, so as to significantly improve the positioning accuracy of complex targets. The experimental results show that the performance of D-GPH on MS coco dataset is better than Mask R-CNN benchmark, especially in the detection of fuzzy targets and difficult samples.

Keywords

Object Detection; Mask R-CNN; OverLoCK; Dynamic Prior Router; Manhattan-constrained Graph Attention; Heterogeneous Fusion; Varifocal Loss; Uncertainty-Aware Refinement.

1. Introduction

As a basic and challenging task in the field of computer vision, target detection is widely used in key fields such as autonomous driving[1], industrial detection[2] and biometrics[3]. The core goal 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 network (CNN)[4], breakthroughs have been made in the performance of detectors, gradually leading to a technological landscape dominated by single-stage detectors[5] and multi-stage detectors[7].

ISSN: 2688-8645

Among them, multi-stage detectors based on regions, represented by Mask R-CNN[9], have long been the first choice for handling complex instance aware tasks with their accurate candidate frame extraction and multi task parallel processing mechanism. The deep optimization of its structure, especially how to enhance the feature robustness in complex environment, has become one of the core directions of current research.

The typical mask R-CNN structure is composed of backbone, neck, RPN and ROI head. Specifically, the backbone networks (such as the ResNet[10] series) extract multi-scale features through layer by layer convolution; Feature pyramid realizes feature fusion through top-down path; The regional recommendation network generates candidate regions; The ROI header achieves feature alignment, classification and regression.

This paper notes that although the previous improvement work introduced global prior and graph modeling, it still has limitations in dealing with highly challenging complex scenes. First of all, the prior injection method is too static: the modulation of the feature pyramid to the global prior is usually unified and hard coded, and the dependence of the features of each layer on the prior cannot be dynamically adjusted according to the current image content. Secondly, the graph attention mechanism lacks spatial constraints: the attention of the standard graph treats all pairs of pixels equally, resulting in long-distance background noise that is easy to participate in feature modeling, introducing interference and causing features to be too smooth. In addition, the neck network lacks the coordination of heterogeneous features, and it is easy to ignore local texture details by relying only on global graph modeling. Finally, the detection head is not sensitive to difficult samples: the traditional classification and regression branches do not explicitly model the positioning quality, and lack the secondary refining mechanism for low confidence regions.

In order to solve the above problems, this paper proposes a dynamic graph prior hybrid enhanced detection framework, called D-GPH (dynamic graph prior hybrid). D-GPH has been systematically enhanced on the basis of retaining the global perception ability of overlock[11] backbone network. Firstly, this paper designs a dynamic prior router (DPR), which generates the injection weights of each level in real time according to the image content, and realizes the adaptive prior modulation. Secondly, Manhattan constrained graph attention (MC-GAT) is proposed to suppress long-range noise by introducing physical distance penalty. At the same time, a heterogeneous fusion layer is constructed in the P5 layer, taking into account the global topology association and local detail enhancement. Finally, the uncertain aware refining head (UAR head) is designed to improve the detection accuracy by using the quality perception loss and the secondary prediction mechanism of difficult samples.

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

(1) A high performance target detection framework named D-GPH is constructed. On the basis of OverLoCK-GPH, the framework integrates dynamic prior modulation, spatial constraint graph interaction and uncertainty perception refinement, breaking the limitations of the traditional two-stage detector in the complex environment where the feature expression is rigid and vulnerable to noise interference, and providing a full link dynamic optimization scheme for the accurate detection of environment adaptation.

(2) Dynamic prior router (DPR) is proposed. This mechanism breaks through the static paradigm of traditional prior injection, and dynamically generates scale specific modulation coefficients by modeling global context features, so as to realize on-demand prior injection driven by image content. This design significantly enhances the adaptive expression ability of the model to the changing environment background.

(3) The Manhattan constraint graph attention (MC-GAT) and heterogeneous feature fusion layer are designed. By introducing Manhattan distance penalty term into the non Euclidean space interaction, and using the physical geometry prior explicit constraint graph to connect the weights, the pollution of long-distance background noise on target features is effectively

ISSN: 2688-8645

suppressed. At the same time, the heterogeneous fusion of global topological Association and local expansion convolution enhancement is realized in the top-level feature, which further improves the discrimination and purity of feature representation.

(4) Uncertain aware refining head (UAR head) is introduced. The module integrates variable loss and moving average (EMA) mechanism of quality estimation, and triggers the attention recalibration and iterative prediction process of the secondary channel by explicitly modeling the uncertainty of the detection results. This not only improves the positioning quality of difficult samples, but also effectively solves the problem of missing detection and false detection of small targets in complex background.

(5) Rigorous and extensive experimental verification has been carried out in the MS coco public data set. The experimental results show that D-GPH achieves significant performance gain compared with the benchmark model in the standard training cycle. At the same time, ablation experiments show that the purity and robustness of feature expression are enhanced while maintaining the reasoning efficiency, which fully verifies the advancement and practical value of this model in high-performance target detection tasks.

2. Related Work

The object detection architecture covers multiple paradigms based on convolutional neural networks (CNN)[4], recurrent neural networks (RNN)[12], graph neural networks (GNN)[13], and transformers[14]. Among them, Girshick et al.'s Fast R-CNN[15] and Ren et al.'s Faster R-CNN[16] laid the foundation for two-stage detection; Subsequently, Mask R-CNN proposed by He et al. [16] became a classic benchmark in this field by adding instance segmentation branches and using RoI Align to alleviate quantization errors. In 3D point cloud detection, Fast Point R-CNN[17] further improves detection accuracy through a dual-path representation of voxels and original point clouds. In the 2D domain, subsequent improvement work, such as Cai et al.'s Cascade R-CNN[18], improved the recall rate of high-quality target boxes through multi-stage cascade refining; Cheng et al. introduced boundary learning in BMask R-CNN[19] to optimize mask localization; Wu et al.'s RGC Mask R-CNN[20] and Lin et al.'s G-Mask[21] respectively enhanced it from the perspectives of normalization strategy and GIoU loss function. However, most of the classic methods mentioned above rely on a bottom-up static feature extraction paradigm, which lacks real-time adaptability to image content heterogeneity when facing complex degraded environments such as low contrast.

To enhance multi-scale feature representation, Feature Pyramid Network (FPN) integrates deep semantics and shallow details through a top-down path. On this basis, researchers began to introduce non local modeling to break through the limitations of convolutional local receptive fields. For example, the pre benchmark work OverLoCK-GPH[22] in this article outputs global context priors and enhances features through prior space modulation (PGSM) and graph association modulation (GAM). However, this prior injection method is statically fixed across images, and standard graph attention mechanisms lack spatial geometric constraints, which can easily introduce background noise interference from a long distance.

At the same time, dynamic neural networks[23] and predictive uncertainty modeling have become important ways to improve environmental robustness. The dynamic network theory emphasizes that network weights should be dynamically generated based on input data to enhance environmental robustness. In terms of detection head design, Zhang et al. proposed VarifocalNet (VFNet)[24], which proved that jointly modeling classification confidence and localization quality (IoU) can significantly improve the performance of dense detection; Pan et al. [25] also validated the effectiveness of quality perception loss in complex biological detection. However, how to use this uncertainty estimation for closed-loop feature

ISSN: 2688-8645

recalibration of low confidence "difficult samples" in mainstream two-stage detectors is still an area that needs to be explored.

In response to the shortcomings of the existing work mentioned above, this paper proposes a dynamic graph prior hybrid framework D-GPH. This article still follows the two-stage decoupling framework of Mask R-CNN, but has undergone generational upgrades in three dimensions: feature modulation, spatial relationship modeling, and detection head refinement. At the feature injection end, a dynamic prior router (DPR) is innovatively designed to replace static prior, achieving "graph adaptive" feature modulation. In the feature fusion layer, construct Manhattan Constrained Graph Attention (MC-GAT) and heterogeneous fusion layer. On the one hand, using the Manhattan distance penalty term to impose physical locality constraints on graph interactions significantly reduces long-range background noise; On the other hand, Dilated Conv is combined with top-level feature parallelism to achieve complementarity between global long-range topology and local texture details. Introduce an uncertainty aware refinement head (UAR Head) at the output of the detection head to accurately locate difficult sample areas with high uncertainty, and apply secondary channel adaptive recalibration and prediction to them. In summary, D-GPH breaks through the static representation limitations of traditional two-stage detectors by decoupling and reconstructing feature flow, highly integrating dynamic routing mechanisms, heterogeneous graph modeling with spatial geometric constraints, and uncertainty based refinement strategies. Compared with similar cutting-edge methods, D-GPH exhibits excellent environmental adaptability without significantly increasing the number of parameters and computational burden, achieving a more competitive balance between detection performance and resource overhead, and providing an efficient solution for robust detection in complex degraded scenarios.

3. Methodology

3.1. D-GPH Network Structure

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

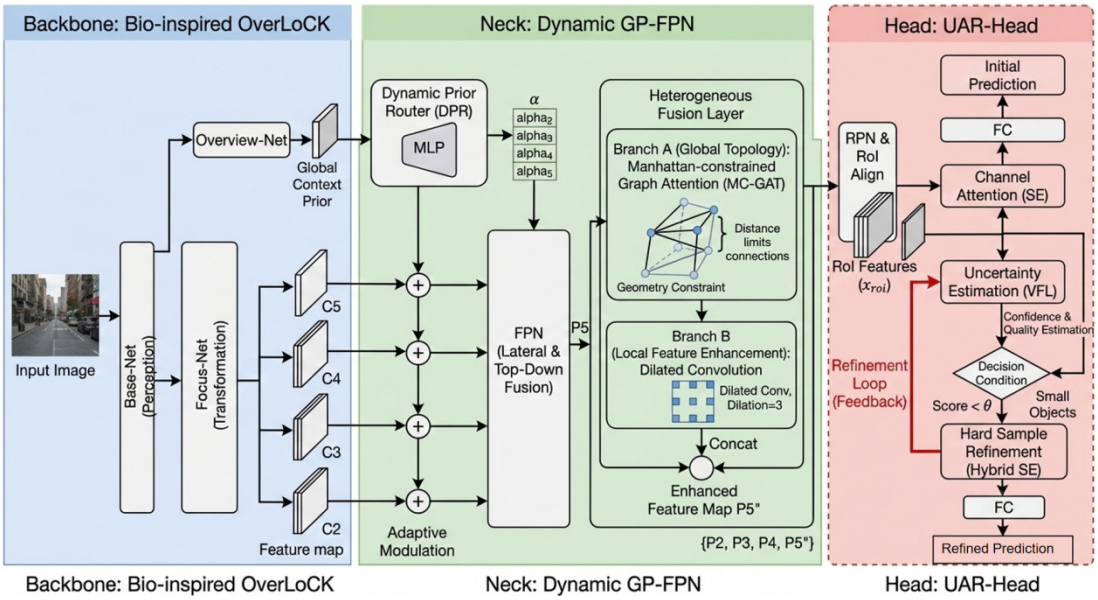

Figure 1. The D-GPH network structure

This model inherits the classic two-stage detection paradigm of Mask R-CNN while addressing the issues of global prior loss and insufficient feature purity in complex scenes. It constructs a deep perception system consisting of an OverLoCK backbone network, a prior guided graph

ISSN: 2688-8645

feature pyramid (GP-FPN), a region recommendation network (RPN), and an uncertainty aware refinement head (UAR 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 internal collaborative Base Net, Overview Net, and Focus Net sub networks, it simulates the biological inspiration mechanism of human vision of "overview first, then detail". The role of OverLoCK is to break through the bottleneck of convolutional neural networks limited to local receptive fields. In the first two stages, local features C2 and C3 are output. In the high-level stage, global contextual features are generated by the contextual branch while generating the main branch features. Then, the fusion sub network is used to obtain C4, C5, and finally, the skeleton 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 prior guided graph feature pyramid (GP-FPN) receives F and P_{ctx} as inputs. Unlike static prior injection, D-GPH first introduces a dynamic prior router (DPR): based on P_{ctx}, adaptive prior injection strengths α_i are generated for each layer, and the backbone features are modulated before being fed into the standard FPN to obtain $F'_i = F_i + \alpha_i \cdot P_i^{\mathrm{proj}}$, achieving dynamic prior injection by layer and graph. Subsequently, $\{F'_2, F'_3, F'_4, F'_5\}$ will be processed through standard FPN to obtain five layers of features $\{\mathrm{P}_2, \mathrm{P}_3, \mathrm{P}_4, \mathrm{P}_5, \mathrm{P}_6\}$. Then, P_{ctx} will be used for spatial weight modulation on each layer $\mathrm{P_i}$: P_{ctx} will be bilinear interpolated to the spatial size of $\mathrm{P_i}$, mapped to a single channel spatial weight through an independent 1x1 convolution $\mathrm{W_i}$ of this layer, and multiplied element by element with $\mathrm{P_i}$ to suppress background interference and enhance target response. Its mathematical expression is:

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

Among them,$\odot$ denotes element-wise multiplication, and σ represents Sigmoid. At the same time, to compensate for the semantic loss caused by traditional FPN static linear superposition and suppress long-distance background noise, GP-FPN introduces Manhattan Constrained Graph Attention (MC-GAT) on the modulated top layer $\mathrm{P'_5}$: mapping the spatial dimension to a sequence of graph nodes, and introducing Manhattan distance matrix penalty on the logits of multi head self attention, so that attention decays with the spatial distance between nodes, thereby suppressing far-field background and enhancing foreground correlation. Subsequently, a heterogeneous fusion layer is constructed on $\mathrm{P'_5}$: one path is the global output of MC-GAT, and the other path is the local enhancement branch (such as dilated convolution). The two paths are concatenated along the channel and convolved 1 × 1 to obtain P''_5. The final output of GP-FPN is $\{P'_2, P'_3, P'_4, P''_5, P'_6\}$, which is used by RPN to generate candidate boxes and obtain a fixed size instance feature block X_{roi} through RoI Align.

To address the interference of residual background noise in the RoI region on the classifier and improve the discriminative ability between difficult samples and small targets, this paper designs an uncertainty aware refinement head (UAR Head). Before X_{roi} enters the shared fully connected layer, it undergoes channel attention recalibration based on SE structure: the channel description vector z is obtained through global average pooling, and then the channel weight s is obtained through two layers of fully connected and Sigmoid, that is:

$$z = GlobalAvgPool(X_{roi}), s = \sigma(W_2\delta(W_1 z))$$

Among them, δ is the ReLU activation function, σ is Sigmoid, and the purified feature is $X_{roi} \odot s$. After flattening and sharing FC, the first round of classification and regression results were obtained; For difficult samples with relative reliability below the threshold θ or corresponding proposal area less than $A_{\min}$, use the original X_{roi} to perform secondary prediction through

ISSN: 2688-8645

channel attention and shared FC, and replace the output of the RoI to improve the localization and recall of difficult and small targets. The classification branch adopts Varifocal Loss and maintains the exponential moving average (EMA) of IoU to enhance quality perception. This design improves the robustness and accuracy of the detection head through channel purification and secondary refinement of difficult samples, while maintaining the same RPN and RoI processes as Mask R-CNN.

In summary, D-GPH achieves full process optimization from scene understanding to instance perception by outputting scene level priors in the backbone network, performing dynamic prior routing (DPR) at the neck, prior guided spatial modulation (PGSM), Manhattan constrained graph attention (MC-GAT) and heterogeneous fusion, and uncertainty aware refinement (channel attention+difficult sample secondary prediction+Varifocal Loss) at the head, making it more robust and accurate than native Mask R-CNN in object detection and instance segmentation tasks.

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

ISSN: 2688-8645

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 $S{\times}S$ 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 a Prior Guided Graph Feature Pyramid Network (GP-FPN) to achieve efficient fusion of multi-scale features and deep enhancement of spatial semantics. The input of GP-FPN consists of two parts: the local feature set F from different stages of the backbone network, and the global context prior P_{ctx} generated by Overview Net. Unlike traditional feature pyramid networks that rely solely on convolution and linear upsampling for feature stacking, GP-FPN introduces dynamic prior routers (DPR), prior guided spatial modulation (PGSM), Manhattan constrained graph attention (MC-GAT), and

ISSN: 2688-8645

heterogeneous fusion layers for image adaptive prior injection based on the standard FPN five layer pyramid. This design enables each layer of features to retain high-resolution boundary details while carrying deep semantic information endowed by OverLoCK's bio inspired global priors and graph association mechanism, laying a solid theoretical and representation foundation for subsequent region proposal networks (RPNs) to generate accurate proposal boxes and detection heads for final feature purification.

The forward process of GP-FPN is organized into the following sub steps: (1) DPR is modulated before the multi-scale features F are fed into the standard FPN fusion; (2) The modulated features are fed into a standard FPN to obtain a five layer pyramid; (3) Apply PGSM to each layer; (4) Perform MC-GAT and heterogeneous fusion on the top-level modulation feature P'_5 to obtain P''_5; (5) Output the final multi-scale feature set for use by RPN and RoI Align. The following sections describe each component in order.

3.3.1. DPR

In the original GP-FPN, the global prior P_{ctx} is injected into the feature pyramid in a unified and static manner through PGSM. However, different images have varying degrees of dependence on the global context: shallow features rely more on prior knowledge for texture and edge enhancement, while deep features rely more on prior knowledge for semantic consistency constraints. To achieve prior modulation by layer and graph, this paper introduces DPR, which generates prior injection intensity for each layer based on the current image content, and injects prior into each layer before FPN fusion.

The lateral feature corresponding to the $i-th$ layer $(i = 2, \dots, 6)$ of the backbone network is F_i, and the global context prior output by Overview Net is $P_{\text{ctx}} \in \mathbb{R}^{C_{\text{ctx}} \times H_{\text{ctx}} \times W_{\text{ctx}}}$. DPR first aggregates P_{ctx} into channel level descriptors through global average pooling:

$$s = \text{GlobalAvgPool}(P_{\text{ctx}}) \in \mathbb{R}^C$$

Then, four weight vectors $\alpha = [\alpha_2, \alpha_3, \alpha_4, \alpha_5]$ were obtained through two layers of MLP with bottlenecks. The first layer compresses the channel dimension by mid_ratio, and the second layer maps it to 4 dimensions and passes it through Sigmoid:

$$h = \text{ReLU}(W_1 s + b_1), \quad \alpha = \text{Sigmoid}(W_2 h + b_2)$$

The middle layer dimension is C/mid_ratio. The scalar α_i controls the intensity of global prior injection into the $i-th$ layer. For each layer i, scale P_{ctx} to the spatial size of F_i, and project it onto the same channel dimension as F_i through a 1x1 convolution to obtain P_i^{proj}. The modulation characteristics sent to the standard FPN are

$$F'_i = F_i + \alpha_i \cdot P_i^{\text{proj}}$$

Among them, α_i is broadcasted in both spatial and channel dimensions. Therefore, the generation of α not only involves the allocation of weights to each layer, but also plays a dual role in channel attention and task adaptation: DPR allocates different α_i by layer and image, and achieves dynamic prior modulation before FPN fusion.

3.3.2. Standard FPN and PGSM

The first step in pyramid construction is to feed the DPR modulated features into a standard FPN. In the top-down path, high-level features are upsampled and added element by element to the lateral output of the current layer, and then smoothed by a 3x3 convolution. The $i-th$ layer fusion output can be represented as:

$$P_i = \text{Conv}_{3\times3} l(\text{Upsample}(P_{i+1}) \oplus F'_i)$$

Where $\oplus$ represents element wise addition, and Upsample represents bilinear upsampling. Thus, a five layered pyramid $\{P_2, P_3, P_4, P_5, P_6\}$ is obtained.

The second step is to apply prior guided spatial modulation to all five layers of features. The core idea is to use global priors as spatial attention masks to suppress background noise. The

ISSN: 2688-8645

model compresses P_{ctx} through a lightweight convolutional layer φ_i (independent 1x1 convolution for each layer, mapping the C channel of P_{ctx} to a single channel), and aligns it to the spatial size of each layer feature P_i using bilinear interpolation U_i to generate scale dependent modulation weight maps:

$$\mathcal{M}_i = \sigma(\Psi_i(\mathrm{U}_i(P_{\text{ctx}})))$$

Where σ is Sigmoid activation. The characteristic representation after modulation is: $P'_i = P_i \odot M_i$,Where $\odot$ represents element wise multiplication. Through this explicit spatial calibration, GP-FPN can effectively highlight potential target areas from multi-scale feature maps, enhancing the model's discriminative ability in low contrast scenes.

3.3.3. MC-GAT

The original image association modulation (GAM) treats all pixel pairs equally in attention calculation, and distant background pixels may generate strong attention and introduce noise. To suppress long-range background interference while retaining non local semantic inference, this paper replaces top-level graph attention with Manhattan Constrained Graph Attention (MC-GAT) and applies penalties to attention logits through Manhattan distance matrix, causing attention to decay with spatial distance.

Flatten the modulated top-level feature P'_5 into a node sequence $X \in \mathbb{R}^{N\times C}$, where $N = H \times W$ is the number of nodes and C is the feature dimension. The Manhattan distance matrix $D \in \mathbb{R}^{N\times N}$ is defined as:

$$D_{i,j} = |x_i - x_j| + |y_i - y_j|$$

Among them, $(x_i - x_j)$ and $(y_i - y_j)$ are the spatial coordinates of pixels i and j on the feature map, respectively. The attention weight calculation in MC-GAT is:

$$\text{Attn} = \text{Softmax}\left(\frac{QK^{\top}}{\sqrt{d}} - \tau \cdot D\right)$$

Among them, Q and K are the query and key projection of X, d is the head dimension, and τ is the distance penalty strength. Subtracting $\tau \cdot D$ can reduce the attention weight of distant node pairs, thereby suppressing far-field background. To enable different attention heads to adapt to different effective receptive fields (some more local, some farther away), implement τ as a learnable parameter by head, that is, the $h-th$ head corresponds to τ_h . In the implementation, $\tau = \exp(\tau_{\log})$ is used to ensure positivity, with each head having a $\tau_{\log}$. The output of multi head self attention is calculated based on the logits after head penalty, and then connected to LayerNorm through residual connections, and restored to a spatial feature map. This output is the global branch output P_{ctx} used by the subsequent heterogeneous fusion layer.

3.3.4. Heterogeneous Fusion Layer

Relying solely on global image attention may weaken local texture and boundary information. To combine global topology with local details, this paper introduces a heterogeneous fusion layer on the top layer P'_5 . This layer consists of two branches: branch A applies MC-GAT to P'_5 , obtaining $F_{\text{global}} \in \mathbb{R}^{C\times H\times W}$; branch B uses dilated convolution for local enhancement, obtaining an effective receptive field of approximately 7×7 without increasing the number of parameters. Branch B thus complements MC-GAT in the receptive field: MC-GAT is responsible for long-range dependencies and topological associations, while dilated convolution is responsible for local texture and boundary enhancement. After concatenating the two channels along the channel dimension, they are projected back to channel C through a 1×1 convolution:

$$P''_5 = \text{Conv}_{1\times 1}(\text{Concat}(F_{\text{global}}, F_{\text{local}}))$$

Among them, $F_{\text{local}} = \text{DilatedConv}(P'_5)$ represents the output of the dilated convolution branch. Splicing reduces the channel dimension to 2C, and after 1×1 convolution, it is restored to C. Therefore, the final number of channels in P''_5 is the same as that in the other pyramid

ISSN: 2688-8645

layers. Branch B adopts dilated convolution instead of ordinary 7×7 convolution, which enlarges the receptive field without increasing parameters and complements the receptive field formed by MC-GAT, while preserving local details and avoiding redundant calculations.

3.3.5. GP-FPN Output

The final neck output of GP-FPN is a multi-scale feature set $\{P'_2, P'_3, P'_4, P''_5, P'_6\}$: where P'_2, P'_3, P'_4, P'_6 are the features modulated by PGSM (P'_6 is still referred to as P'_6 after PGSM), and the top layer is the P''_5 obtained by heterogeneous fusion. This feature set is used for RPN to generate proposals and for RoI Align to extract RoI features. In summary, GP-FPN first implements image adaptive prior injection through DPR, then constructs a pyramid using standard FPN and PGSM, and finally applies MC-GAT and local expansion branches on P'_5 and fuses them to obtain P''_5 , completing the complete process from dynamic prior routing, spatial modulation to global local heteromorphic fusion.

3.4. UAR Head and Multi Task Loss

After extracting fixed sized RoI features from the feature map output by GP-FPN using RPN and RoI Align, the detection head performs precise classification and bounding box regression on each candidate region. This article constructs a UAR Head and uses Varifocal Loss and IoU exponential moving average (EMA) to achieve quality aware classification on the basis of a hybrid BBox head. It also performs a second channel attention and prediction on difficult samples (low confidence or small area) to refine their scores and boxes. The following provides input conventions and forward flow, channel attention (SE), Varifocal loss and EMA, shared fully connected layers and classification/regression branches, bounding box encoding and decoding, and multitasking loss in sequence.

3.4.1. Input and Forward Process

The candidate regions output by RPN are sent to RoI Align along with the feature maps of GP-FPN; RoI Align crops and resamples each candidate box to obtain RoI features $x_{\text{roi}} \in \mathbb{R}^{N\times C\times H_{\text{roi}}\times W_{\text{roi}}}$, where N is the number of RoIs in the current batch, C is the number of neck output channels, and $(H_{\text{roi}}, W_{\text{roi}})$ is the RoI spatial size.

First pass: For each RoI, the original feature x_{roi} is flattened by the channel attention module (SE), and then passed through a two-layer shared fully connected network to obtain the first round of category $logits$cls_score$^{(1)}$ and bounding box offset bbox_pred$^{(1)}$. Then perform sample determination: If any of the following conditions are met, mark the RoI as a difficult sample and enter secondary refinement: (1) Softmax the maximum category probability $p_{\max} < \theta$ (θ is the threshold) for cls_score$^{(1)}$; (2) The proposal area corresponding to the RoI is less than the minimum area threshold $A_{\min}$ (as given in pixels or relative image area ratio). The latter addresses the phenomenon of "high score but low accuracy" (high classification but inaccurate bounding box) that often occurs with extremely small targets in underwater or low contrast scenes. Forcing secondary refinement of small targets can improve localization and recall.

Secondary refinement: For the above difficult samples, take their original x_{roi} and apply channel attention and shared FC again to obtain cls_score$^{(2)}$ and bbox_pred$^{(2)}$, which are used to replace the first round of cls_score and $bbox\_pred$ of the RoI.

Output: The final cls_score and $bbox\_pred$ for all RoIs are given by the first pass for easy samples and the second pass for difficult samples, used for loss calculation and inference.

3.4.2. Squeeze-and-Excitation

The channel attention module globally aggregates RoI features in the spatial dimension to obtain channel descriptors, and then generates channel weights through lightweight two-layer MLP and Sigmoi, and scales the original features by channel. The calculation is divided into three steps: Squeeze, Excitation, and Scale.

ISSN: 2688-8645

Squeeze: Perform global average pooling on each channel of RoI feature $x \in \mathbb{R}^{C\times H\times W}$, compress the spatial dimension into a scalar, and obtain channel level vector $z \in \mathbb{R}^{C}$:

$$z = \text{GlobalAvgPool}(x), \quad z_c = \frac{1}{HW}\sum_{h,w} x_{c,h,w}$$

Excitation: Learning channel dependencies through two fully connected layers and nonlinear activation. If the compression ratio is r and the intermediate layer dimension is C/r, we have:

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

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: Broadcast s along the spatial dimension and multiply it element by element with x to obtain the recalibrated RoI features:

$$\tilde{x} = s \odot x$$

Where$\odot$ denotes the channel-wise multiplication broadcast along the spatial dimension. So far, channel attention has completed the channel dimension adaptive recalibration of ROI features without changing the size of feature space, providing more discriminative input for the subsequent full connection layer.

3.4.3. Varifocal Loss and EMA

Variable loss (VFL) is used to replace cross entropy in classification branches to emphasize high-quality positive samples and focal weight is used for negative samples. Let the mass fraction of the positive sample be q (such as the IOU of the prediction box and the real box), then the loss weight of the positive sample is Q; the focal weight is used for the negative sample. The simplified form is:

$$\text{Positive sample: } L_+ = -q \cdot (1-p)^{\gamma} \cdot \ln p$$
$$\text{Negative sample: } L_- = -\alpha \cdot p^{\gamma} \cdot \ln(1-p)$$

Where P is the prediction probability of the target class, γ is the focus parameter, and α is the balance factor. The quality q can be obtained from the IOU during training, or from the quality estimation of EMA smoothing.

EMA: header maintenance scalar q_{ema}(or mass estimation per ROI), updated as follows:

$$q_{\text{ema}} \leftarrow \lambda \cdot q_{\text{ema}} + (1-\lambda) \cdot q_{\text{batch}}$$

Where q_{batch} is the mean value (or other quality statistics) of positive sample IOU in the current batch. EMA is used to stabilize the quality estimation, and can participate in the adaptation of VFL weight or difficult sample threshold.

3.4.4. Shared Full Connection Layer, Classification/Regression Branch and Bounding Box Encoding/Decoding

After the ROI feature $\tilde{x}$ recalibrated by channel attention is flattened along the spatial dimension, the two-layer shared fully connected network is input to obtain the high-dimensional shared representation. Let the flattened vector get the intermediate representation h through the first layer full connection and ReLU activation, and then get the shared representation z_{shared} through the second layer full connection and ReLU activation, in the form of:

$$h = \delta(W_1 \cdot \text{flatten}(\tilde{x}) + b_1), \quad z_{\text{shared}} = \delta(W_2 h + b_2)$$

Where δ represents the ReLU activation function, W_1 and W_2 are learnable weight matrices, and b_1 and b_2 are biases. The shared representation feeds both the classification branch and the regression branch.

ISSN: 2688-8645

In the classification branch, the classification layer outputs Logits (non normalized score) of K categories for each ROI, that is, cls_score $= W_{\text{cls}} z_{\text{shared}} + b_{\text{cls}} \in \mathbb{R}^K$, where K is the number of categories and W_{cls} is the weight of the classification layer. During the training, the varifocal loss described in section 3.4.3 is used to monitor the matching of cls_score and the real category label.

In the regression branch, the category related regression strategy is adopted: each category corresponds to a set of bounding box offsets. The regression layer outputs bbox_pred $= W_{\text{reg}} z_{\text{shared}} + b_{\text{reg}}$, that is, each category corresponds to a 4-dimensional offset (t_x, t_y, t_w, t_h), a total of 4K dimensions. During training, L1 loss or smooth L1 loss is used to monitor the difference between the predicted offset and the real offset after coding. The physical coordinates of the bounding box are encoded in the form of delta, and the coding rules are given below; When decoding, the final frame coordinates are obtained by the offset output from the candidate frame and the regression branch according to the decoding rules.

In order to stabilize the regression scale and avoid the gradient difference caused by candidate frames with different scales, the detection head predicts the normalized offset relative to the candidate frame rather than the absolute coordinates. Let the candidate box be (x_a, y_a, w_a, h_a), representing the horizontal and vertical coordinates and width and height of the center respectively; The ground truth is (x_g, y_g, w_g, h_g). The coding target is defined as:

$$t_x = \frac{x_g - x_a}{w_a}, \quad t_y = \frac{y_g - y_a}{h_a}, \quad t_w = \ln\frac{w_g}{w_a}, \quad t_h = \ln\frac{h_g}{h_a}$$

During training, the above target offset will be normalized according to the preset mean and standard deviation before participating in the loss calculation, that is, the real label in the loss is the normalized coded value, and the predicted value is directly output by the regression layer. In the reasoning stage, according to the four-dimensional offset $(\hat{t}_x, \hat{t}_y, \hat{t}_w, \hat{t}_h)$ between the candidate box and the regression branch output (corresponding to the output of the category predicted by ROI), the physical coordinates of the final boundary box are decoded according to the following rules:

$$x = x_a + \hat{t}_x \cdot w_a, \quad y = y_a + \hat{t}_y \cdot h_a, \quad w = w_a \cdot \exp(\hat{t}_w), \quad h = h_a \cdot \exp(\hat{t}_h)$$

Where (x, y, w, h) is the center coordinate and width height after decoding. Through the above encoding and decoding methods, the regression branch has the same optimization goal on the candidate boxes of different scales, which is conducive to improving the positioning accuracy.

The total loss at the detection end is composed of RPN stage loss and ROI stage loss, which is weighted sum. Record the classification loss and boundary box regression loss of RPN as L_{RPN}, the ROI classification loss as L_{VFL} using varifocal loss, the ROI boundary box regression loss as L_{bbox}, and the mask loss of instance segmentation as L_{mask}. The total loss can be written as:

$$L = L_{\text{RPN}} + \lambda_{\text{cls}} L_{\text{VFL}} + \lambda_{\text{reg}} L_{\text{bbox}} + \lambda_{\text{mask}} L_{\text{mask}}$$

Where $\lambda_{\text{cls}}, \lambda_{\text{reg}}, \lambda_{\text{mask}}$ are the loss weights of each branch. During training, the classification and regression of RPN, UAR head and optional mask header are jointly optimized through the above multi task losses.

To sum up, UAR head achieves uncertainty perception refining through three designs: first, it uses varifocal loss in the classification branch and cooperates with IOU's EMA to make the model have differential modeling ability for high-quality positive samples and difficult samples; Secondly, the channel attention based on squeeze and exception is introduced before sharing the full connection layer to calibrate the channel dimension of ROI features, suppress the background channel and enhance the discrimination channel; Third, for the difficult samples that meet the requirements of "confidence below the threshold θ " or "corresponding proposal area is less than A_{min}", the original ROI feature is used to predict the second time through the channel attention and sharing FC again, and the second time result is used to replace the category score and boundary box of the ROI, so as to improve the positioning and recall of

ISSN: 2688-8645

difficult targets and small targets. UAR head cooperates with DPR of neck, MC-GAT and heterogeneous fusion layer to form a complete detection link of D-GPH from "dynamic prior modulation" to "uncertainty perception refinement", so as to enhance positioning accuracy and classification robustness in complex scenes, low contrast and other applications.

4. Experiment

4.1. Experimental Setup and Dataset Introduction

In this paper, the proposed model is tested on MS coco 2017 dataset using mmdetection detection library. Following the common settings, the training set containing about 118000 images and the verification set of 5000 images are used for network parameter training. The experiment uses geforce RTX 4090 GPU for parallel training and CUDA acceleration. The training strategy includes: using adamw as the optimizer to train 12 epochs, the initial learning rate is 1×10^{-4} , the batch size is set to 6, the weight attenuation is 0.05, and the learning rate is stepped attenuation every 8 or 11 epochs. Data enhancement uses conventional strategies such as random flipping, random clipping, and multi-scale training; During training, the short edge resolution is set to 800, and the long edge multi-scale range is [800, 1333]. The image size during verification and test is 1333 × 800.

4.2. Network Hyperparameter Configuration

This paper uses the pre trained OverLoCK as the backbone network of Mask R-CNN. The neck adopts GP-FPN, which outputs five scale feature layers; The dimension reduction ratio of MLP intermediate layer of DPR is mid_ratio=4. The attention of the top chart adopts MC-GAT: the distance penalty intensity τ is set to per head learnable, the initial value τ_{init} =0.1 (using log parameterization $\tau = \exp(\tau_{\log})$), the number of headers is 8, and dropout is 0.1. In the heterogeneous fusion layer of P5, branch B uses expanded convolution (3 × 3, dilation=3, The effective receptive field is about 7 × 7). The detection head adopts the uncertain perceptual refining head (UAR head): 256 input channels, 1024 shared fully connected hidden layer dimensions, 7 × 7 ROI feature space dimensions, and 16 channel attention (SE) compression ratio. The determination conditions of difficult samples are as follows: (1) classification confidence threshold θ =0.5 (secondary refining is triggered when $p_{max} < \theta$); (2) Lower limit of area $A_{\min}$ (forced secondary refining when the proposed area is less than this threshold). EMA attenuation coefficient λ =0.96; The focus parameter γ and balance factor α of varifocal loss are set as usual. The bounding box code is in deltaxywh format, and the standard deviation of regression target is set to (0.1, 0.1, 0.2, 0.2). The number of categories is 80, which is consistent with COCO.

In terms of loss function, the sum of classification loss and boundary box regression loss of RPN is recorded as L_{RPN}; The ROI classification branch adopts varifocal loss, which is recorded as L_{VFL}; ROI boundary box regression uses L1 loss L_{bbox}; The instance segmentation branch uses the cross entropy loss L_{mask}. The total loss is the weighted sum of the losses of each branch:

$$L = L_{\text{RPN}} + \lambda_{\text{cls}} L_{\text{VFL}} + \lambda_{\text{reg}} L_{\text{bbox}} + \lambda_{\text{mask}} L_{\text{mask}}$$

$\lambda_{\text{cls}}, \lambda_{\text{reg}}, \lambda_{\text{mask}}$ are the corresponding weights (in this paper, 1.0 is taken).

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 carried out on the coco 2017 validation set. The input resolution is set to 800 × 1333, and other settings are consistent with 4.1 and 4.2. The experimental results show that in terms of bounding box detection, this method achieves 39.5 map, 59.0 ap50 and 42.9 AP75; In the aspect of instance segmentation, 38.1 map, 57.4 ap50 and 40.4 AP75 are implemented. Under the premise of less parameters, controllable computational complexity and short training cycle, the method still maintains high detection and segmentation accuracy, and achieves a good balance between lightweight and accuracy. In general, the proposed method achieves the same detection and segmentation performance as the commonly used advanced models with less parameters, moderate computational complexity and short training time, which verifies the effectiveness, efficiency and practicability of the proposed method.

5. Conclusion

In this paper, a lightweight target detection and instance segmentation model D-GPH is proposed, which takes overflow as the backbone network, DPR, PGSM, MC-GAT and the priori guided map feature pyramid (GP-FPN) of heterogeneous fusion layer as the neck, and UAR head as the detection head. Through the cooperation of lightweight backbone, dynamic graph prior pyramid and uncertainty perception refining head, the design can reduce the amount of parameters and computational overhead while maintaining the ability of multi-scale feature expression and detection, and help to shorten the training cycle and accelerate convergence. The experimental results show that, on the premise of achieving the detection and segmentation performance equivalent to the common settings, the model has obvious advantages in terms of parameters, computational complexity and the number of training rounds required, and has good application potential in resource constrained or deployment efficiency sensitive scenarios. However, under the same training settings, the detection and segmentation accuracy of this model is still inferior to the advanced methods using larger backbone or longer training cycle. Future work will focus on the verification of more low contrast data sets, making DPR/EMA and difficult sample threshold fully adaptive, and lightweight MC-GAT and heterogeneous fusion to further reduce delay.

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] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[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] 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.

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

[13] 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.

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

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] Han Y, et al. OverLoCK-GPH: A Bio-Inspired Object Detector with Graph-Prior Modulation and Hybrid Instance Refinement[J]. (Placeholder for your original OverLoCK-GPH paper).

[23] Han Y, Huang G, Song S, et al. Dynamic neural networks: A survey[J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 44(11): 7436-7456.

[24] Zhang H, Wang Y, Dayoub F, et al. Varifocalnet: An iou-aware dense object detector [C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 8514-8523.

[25] Pan G, Wang X, Li Z, et al. An underwater biological target detection algorithm based on improved RT-DETR[J]. Fishery Modernization, 2025, 52(5): 107-116.