文章编号:1674-2869(2025)05-0539-09

基于视觉变换器动态优化的瓷砖瑕疵检测模型

罗芷萱,田 斌*,黎 曦*
武汉工程大学电气信息学院,湖北 武汉 430205

摘 要:瓷砖表面瑕疵呈现出面积较小、形状不规则、尺度多变、难与复杂花纹背景区分等特点,这使得高精度瑕疵检测的实现异常困难。为解决此问题,提出了一种高性能瓷砖表面瑕疵检测模型,该模型结合了视觉变换器(ViT)网络与YOLOv8n高效架构,在精度和计算量之间取得了良好平衡。首先,使用高效ViT网络作为模型的主干网络,增强模型对图像全局语义和局部特征的联合建模能力,同时保持计算效率。其次,使用空间到深度卷积替换普通下采样卷积,以保留更多细粒度信息。此外,在颈部网络末端添加空间通道重组卷积,使模型更好地聚焦关键特征信息。最后,引入动态头机制,通过动态调整注意力权重,提升模型对复杂瑕疵特征的检测能力。在阿里云天池官方提供的瓷砖瑕疵检测数据集上的实验结果表明,改进后的模型平均检测精度达到71.5%,较原模型YOLOv8n提升了11.6%,且在瓷砖表面微小瑕疵和密集瑕疵检测等复杂场景中表现出色,并兼具较低的部署成本,能够满足实际生产中实时质量检测的需求。

关键词:深度学习;缺陷检测;YOLOv8n;视觉变换器;自注意力机制
中图分类号:TP391.4 **文献标识码**:A **DOI**:10.19843/j.cnki.CN42-1779/TQ.202409022

A ceramic tile defect detection model based on dynamic optimization of vision transformer

LUO Zhixuan, *TIAN Bin**, *LI Xi**
School of Electrical and Information Engineering, Wuhan Institute of Technology, Wuhan 430205, China

Abstract: Ceramic tile surface defects are characterized by their small size, irregular shapes, varying scales, and the challenge of differentiating them from complex patterned backgrounds, rendering high-precision defect detection extremely difficult. To address this issue, we proposed a high-performance ceramic tile surface defect detection model that seamlessly integrates a vision transformer (ViT) network with the efficient YOLOv8n architecture, achieving a favorable balance between accuracy and computational efficiency. First, an efficient ViT network was adopted as the backbone to enhance the model's capability for joint modeling of global image context and local features while maintaining computational efficiency. Then, space-to-depth convolution replaced standard downsampling convolution to preserve more fine-grained information. And a spatial-channel reorganization convolution module was incorporated at the end of the neck network to focus on critical feature information. Finally, a dynamic head mechanism was introduced to dynamically adjust attention weights, significantly improving the model's ability to detect complex defect features. Experiments on the Alibaba Cloud Tianchi ceramic tile defect detection dataset demonstrated that this optimized model achieved a mean average precision (mAP) of 71.5%, outperforming the original YOLOv8n

收稿日期:2024-09-28
基金项目:国家自然科学基金(62367006)
作者简介:罗芷萱,硕士研究生。Email:sql9617@163.com
***通信作者**:田 斌,博士,教授。Email:121494671@qq.com
黎 曦,博士,副教授。Email:305642677@qq.com
引文格式:罗芷萱,田斌,黎曦. 基于视觉变换器动态优化的瓷砖瑕疵检测模型[J]. 武汉工程大学学报,2025,47(5):539-547,555.

by 11.6%. The proposed model excels in detecting tiny and densely distributed defects on tile surfaces while maintaining low deployment costs, fulfilling real-time quality inspection requirements in industrial production.

Keywords: deep learning; defect detection; YOLOv8n; vision transformer; self-attention mechanism

瓷砖作为建筑行业的重要装饰材料之一,市场需求巨大,2023年中国陶瓷砖产量达67.4亿m^2,远高于其他国家。消费者对瓷砖品质的要求也愈发严格。然而,复杂的制作工艺使瓷砖表面容易出现各种瑕疵,如粉团、滴墨、缺角等。尽管自动化技术已逐步应用于施釉、研磨、抛光等工序,质检阶段仍主要依赖人工[1],无论从效率还是精度方面,人工质检都难以适应现代瓷砖生产线的需求。

近年来,基于深度学习的目标检测算法被广泛应用于缺陷检测领域[2],主要包括以单阶段多框检测器(single shot multibox detector, SSD)[3]、YOLO[4]系列模型等为代表的单阶段目标检测算法和以快速区域卷积神经网络(faster region-based convolutional neural network, Faster RCNN)[5]、级联区域卷积神经网络(cascade region-based convolutional neural network, Cascade RCNN)[6]等为代表的两阶段目标检测算法[7-8]。赵楚等[9]针对花色背景较为简单的瓷砖表面瑕疵,提出了一种改进Faster RCNN的瓷砖表面瑕疵算法,其平均检测精度达到了76.3%,但每秒处理图像帧数仅为17.7,实时性较差。单阶段目标检测算法在检测速度上显著优于两阶段目标检测算法,更适用于实际应用场景。如Yu等[10]基于YOLOv5n和移动友好型视觉变换器网络提出了一种高效瓷砖表面缺陷检测算法,该算法在测试集上的平均精度为71.9%,每秒处理图像帧数达到了140,但其使用的瓷砖数据集中复杂花纹瓷砖图像较少,缺乏泛化性。余松森等[11]基于YOLOv8s,通过引入高效通道注意力机制、添加微小目标检测头等改进策略将模型平均检测精度提升了5%,但模型的参数量和计算量较大,仍有一定优化空间。

相对于传统的图像处理技术,深度学习模型往往具有更强的学习能力,但瓷砖表面瑕疵形态多样,且瓷砖背景花纹往往比较复杂,这使得实现高精度的实时瑕疵检测十分困难。为此,本文提出一种高性能的轻量级瓷砖表面瑕疵检测模型,旨在提高模型检测精度的同时尽可能减少模型参数量,主要贡献如下:

(1)采用高效视觉变换器网络[12]作为模型的主干网络。该网络使用了三明治布局块和级联分组注意力机制,其中三明治布局块通过高效的层间连接和计算优化大幅提高了内存访问效率,级联分组注意力机制通过分组多头自注意力机制和逐层级联显著减少了计算冗余。该网络的特征提取能力显著优于原YOLOv8n主干网络,在保持较快推理速度的同时,大幅提高了检测精度。

(2)为进一步增强网络的特征融合能力,采用空间到深度卷积替代普通下采样卷积[13]以保留更多细粒度信息,有效提升了小目标的检测精度。并在颈部网络末端添加空间通道重组卷积[14]模块,进一步优化空间和通道特征,使模型的特征表达更加精准。

(3)引入动态检测头机制[15],尺度感知、空间位置感知和任务感知注意力模块的有机结合,使得模型能够根据不同尺度、空间位置和任务要求灵活调整注意力分布,动态聚焦图像中复杂多样的瑕疵特征,显著增强了模型对复杂样本的检测能力。

1 模型架构与优化设计

瓷砖表面缺陷尺度形状多变、小目标占比高,且难以在复杂花纹背景中准确区分,直接使用轻量级目标检测算法错检、漏检率过高,而大型目标检测模型又难以满足工业场景下的部署条件及实时性要求。针对这一问题,本文基于YOLOv8n模型架构,提出了一种结合动态自注意力机制和视觉变换器网络的改进YOLO模型(yolo with dynamic self-attention mechanism and vision transformer,DViT-YOLO),模型整体结构如图1所示。

1.1 高效视觉变换器主干网络

本文采用高效视觉变换器网络作为模型的主干网络,通过有机结合卷积操作的局部特征提取能力和自注意力机制的全局建模能力,使得模型能够在保持计算效率的同时具备更强的特征提取能力。

高效视觉变换器网络子模块作为高效视觉变换器网络的核心结构,主要由前馈网络层-级联分组注意力模块-前馈网络层交替排列组成,其结构如图2所示,整体计算流程可以表示为:

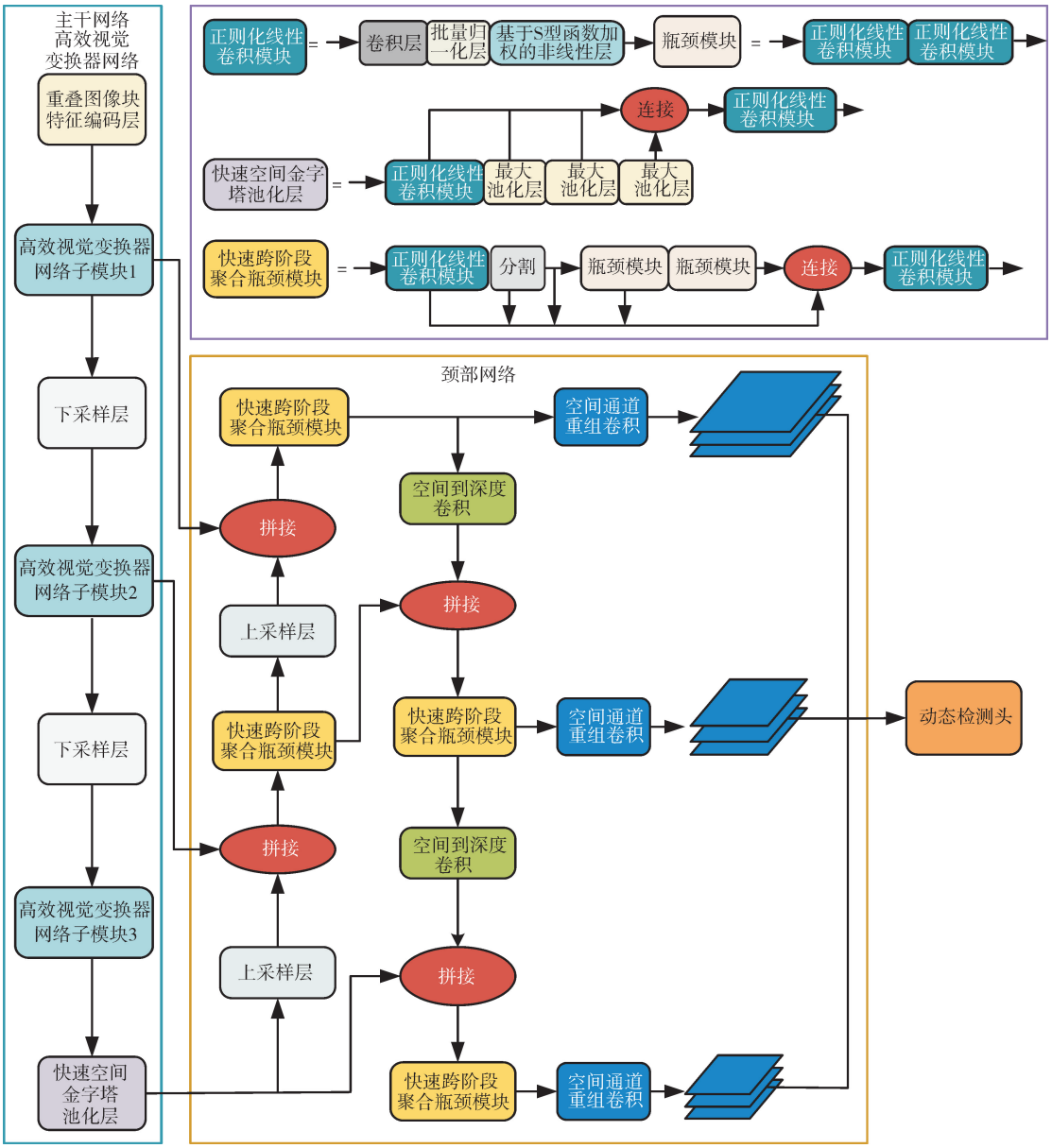

图 1　DViT-YOLO 模型整体结构

Fig. 1　Overall structure of DViT-YOLO model

$$\boldsymbol{X}_{i+1} = V - V - D - \theta[V - V - D(\boldsymbol{X}_i)] \quad (1)$$

其中：$\boldsymbol{X}_i$ 表示第 i 个高效视觉变换器网络子模块的输入特征，D 代表深度卷积，V 代表点卷积，$\theta(\cdot)$ 表示级联分组注意力计算，$\boldsymbol{X}_{i+1}$ 为输出特征。

级联分组注意力模块首先将输入特征 $\boldsymbol{X}$ 划分为 4 个子集，每个子集分别进行自注意力计算。通过分别计算每个子集 $\boldsymbol{X}_j$ 的查询向量 $\boldsymbol{Q}_j = \boldsymbol{X}_j\boldsymbol{W}_j^{\mathrm{Q}}$、键向量 $\boldsymbol{K}_j = \boldsymbol{X}_j\boldsymbol{W}_j^{\mathrm{K}}$、值向量 $\boldsymbol{V}_j = \boldsymbol{X}_j\boldsymbol{W}_j^{\mathrm{V}}$，得到对应的自注意力输出 $\boldsymbol{A}_j$，其中 $\boldsymbol{X}_j$ 表示输入特征 $\boldsymbol{X}$ 的第 j 个子集，$\boldsymbol{W}_j^{\mathrm{Q}}$、$\boldsymbol{W}_j^{\mathrm{K}}$、$\boldsymbol{W}_j^{\mathrm{V}}$ 为对应查询、键和值空间的投影矩阵。最后，将所有子集自注意力计算的输出结果在通道维度上进行拼接，并通过线性变换投影到输出特征维度。具体计算流程如下：

$$\boldsymbol{A}_j = \Theta(\boldsymbol{Q}_j, \boldsymbol{K}_j, \boldsymbol{V}_j) = E\left(\frac{\boldsymbol{Q}_j\boldsymbol{K}_j^{\mathrm{T}}}{\sqrt{d_K}}\right)\boldsymbol{V}_j \quad (2)$$

$$\boldsymbol{Y} = L(T[\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{A}_3, \boldsymbol{A}_4]) \quad (3)$$

其中，Θ 表示自注意力计算函数；E 表示指数归一化函数，即将所有输入向量通过指数运算后进行归一化处理；d_K 表示键向量 $\boldsymbol{K}_j$ 的维度；T 表示拼接操作，即将每个子集的自注意力输出 $\boldsymbol{A}_j$ 沿着分割时的维度拼接；L 表示线性变换操作，用于调整特征维度以匹配输出要求；$\boldsymbol{Y}$ 为最终输出特征，其维度和输入特征 $\boldsymbol{X}$ 一致。

级联分组注意力模块使得每个自注意力计算单元只需要关注部分输入特征，从而大幅提高了计算效率。此外，如图 2 所示，级联分组注意力模

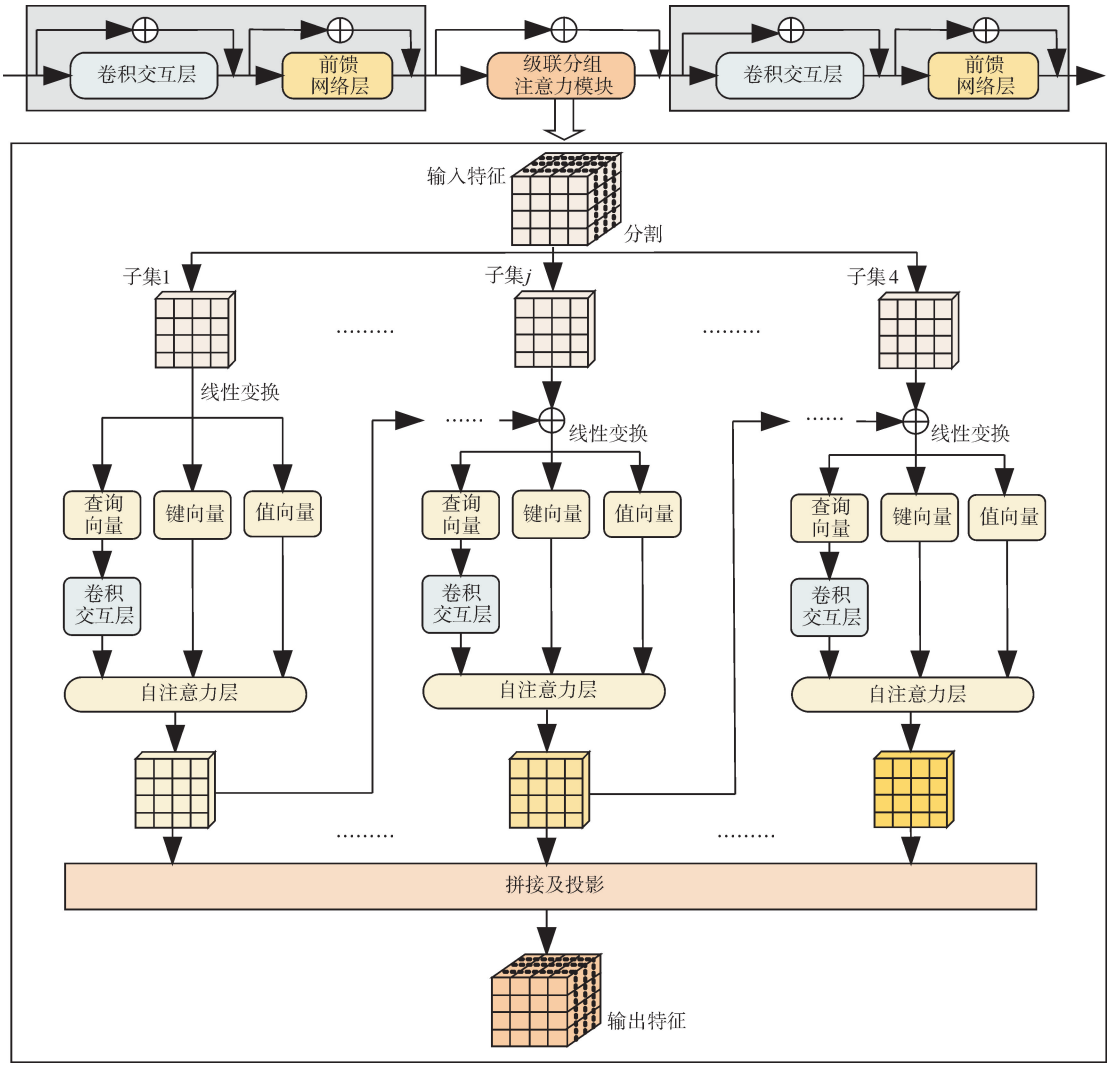

图2 高效视觉变换器网络子模块结构图

Fig. 2 Structure of efficient vision transformer network submodules

块还将上一个子集自注意力计算的输出结果$\boldsymbol{A}_{j-1}$级联到当前输入$\boldsymbol{X}_j$,即:

$$\bar{\boldsymbol{X}}_j = \boldsymbol{A}_{j-1} + \boldsymbol{X}_j, 1 < j \leqslant 4 \tag{4}$$

其中,$\bar{\boldsymbol{X}}_j$表示当前实际输入,该级联操作能够有效融合多层次特征信息,提升注意力图的多样性,还能确保各层次特征在网络中的传递,从而提高模型的特征表达能力。

1.2 优化颈部网络结构

为进一步提升模型有效捕获小目标的能力,本文使用空间到深度卷积替换原颈部网络中的下采样卷积,轻量化的同时还能保留更多的小物体细粒度信息。空间到深度卷积通过在空间维度上分割输入特征图,并沿通道维度重新排列和扩展,以保留更多细粒度信息。最后使用1×1卷积将转换后特征图的通道数恢复到原始维度,确保提取有效特征。

为进一步增强颈部网络的特征融合能力,本文在颈部网络末端添加了空间通道重组卷积。空间通道重组卷积由空间重构模块(spatial reconstruction unit,SRU)和通道重构模块(channel reconstruction unit,CRU)级联而成。输入特征$\boldsymbol{X}$的维度为$C \times H \times W$,其中C代表通道数,即特征图数量,H、W分别代表特征图的高和宽。SRU主要利用组归一化(group normalization,GN)操作的缩放因子$\boldsymbol{\gamma}$评估不同特征图的重要性,并基于$\boldsymbol{\gamma}$计算重构权重$\boldsymbol{\omega}_{\gamma}$,从而获得特征空间$U$下的精炼特征$\boldsymbol{X}_U$。GN操作将$\boldsymbol{X}$沿着通道维度划分为$C$组$[\boldsymbol{X}_1, \cdots, \boldsymbol{X}_g, \cdots, \boldsymbol{X}_C]$,并在每个组内进行归一化处理及仿射变换,具体计算过程如下:

$$G(\boldsymbol{X}_g) = \gamma_g \frac{\boldsymbol{X}_g - \mu_g}{\sqrt{\sigma_g^2 + \epsilon}} + \beta_g \tag{5}$$

其中:$G(\cdot)$代表GN操作;μ_g和σ_g分别为每组特征图$\boldsymbol{X}_g$中所有元素的均值和标准差;γ_g和β_g分别为

模型训练过程中学习得到的组归一化缩放和偏移参数；ϵ为一个很小的正数，用于避免分母为0，保证数值稳定。

重构权重$\boldsymbol{\omega}_\gamma$由$\boldsymbol{\gamma}$经归一化、S型函数映射和门控变换得到，首先将缩放因子$\boldsymbol{\gamma}=[\gamma_1,\gamma_2,\cdots,\gamma_C]$归一化，再通过S型函数映射到(0,1)范围内，得到混合重构权重$\boldsymbol{M}$。然后设定门控机制，将$\boldsymbol{M}$中权值大于0.5的元素设为1，其余权值保持不变，得到信息权重$\boldsymbol{\omega}_{\gamma,1}$；另外，将$\boldsymbol{M}$中权值小于0.5的元素设为0，其余权值保持不变，得到非信息权重$\boldsymbol{\omega}_{\gamma,2}$。具体计算过程如下：

$$\boldsymbol{\omega}_\gamma=\left[\boldsymbol{\omega}_{\gamma,1},\boldsymbol{\omega}_{\gamma,2}\right]=\Gamma\left(S\left(\frac{\gamma_i}{\sum_{i=1}^{C}\gamma_i}\right)\right) \tag{6}$$

其中，Γ表示门控变换操作，S表示S型函数映射操作。

将$\boldsymbol{\omega}_{\gamma,1},\boldsymbol{\omega}_{\gamma,2}$分别与输入特征$\boldsymbol{X}$相乘即可得到信息丰富和信息稀疏的空间加权特征$\boldsymbol{X}_1$、$\boldsymbol{X}_2$，通过交叉融合$\boldsymbol{X}_1$的子集$[\boldsymbol{X}_{1,1},\boldsymbol{X}_{1,2}]$和$\boldsymbol{X}_2$的子集$[\boldsymbol{X}_{2,1},\boldsymbol{X}_{2,2}]$，最终得到空间精炼特征$\boldsymbol{X}_U$。

$$\boldsymbol{X}_U=\boldsymbol{X}_1\oplus\boldsymbol{X}_2=\left(\boldsymbol{X}_{1,1}+\boldsymbol{X}_{2,2}\right)\oplus\left(\boldsymbol{X}_{1,2}+\boldsymbol{X}_{2,1}\right) \tag{7}$$

其中，$\oplus$代表沿通道维度拼接，+代表逐元素相加。

CRU则进一步精炼$\boldsymbol{X}_U$以减少通道冗余，确保特征通道表达的有效性。首先，将$\boldsymbol{X}_U$沿通道维度按一定比例划分为$\boldsymbol{X}_U^{\triangle}$和$\boldsymbol{X}_U^{\triangledown}$。其次，对于上层特征$\boldsymbol{X}_U^{\triangle}$，分别使用3×3高效组卷积和1×1点卷积提取高层抽象信息，并将卷积后得到的特征图相加，得到上层通道精炼特征；对下层特征$\boldsymbol{X}_U^{\triangledown}$使用1×1点卷积提取底层细节信息，并与原特征$\boldsymbol{X}_U^{\triangledown}$拼接，得到下层通道精炼特征。最后，利用全局平均池化生成自适应权重，加权融合上层和下层通道精炼特征。

1.3　引入动态检测头

为满足瓷砖表面瑕疵检测的高精度、强鲁棒性要求，本文在原检测头的基础上引入动态头机制，如图3所示，总输入特征$\boldsymbol{X}\in\mathbb{R}^{P\times O\times C}$由来自不同特征金字塔层级的子特征$[\boldsymbol{X}_1,\boldsymbol{X}_2,\boldsymbol{X}_3,\cdots,\boldsymbol{X}_P]$组成（$P$为特征金字塔层级的总数量；$O$为特征图的空间大小，即特征图高度$H_i$与宽度$W_i$的乘积）。动态头模块通过依次在空间、层级、通道维度上应用对应的注意力机制，使检测头能够更有效地融合不同层级的特征，从而更精准地聚焦检测目标，优化模型预测表现。

空间感知注意力模块$\Omega(\cdot)$首先基于输入特征$\boldsymbol{X}$生成偏移参数，再通过可变形卷积引入稀疏化特征，并将其应用于所有特征层进行加权平均，计算

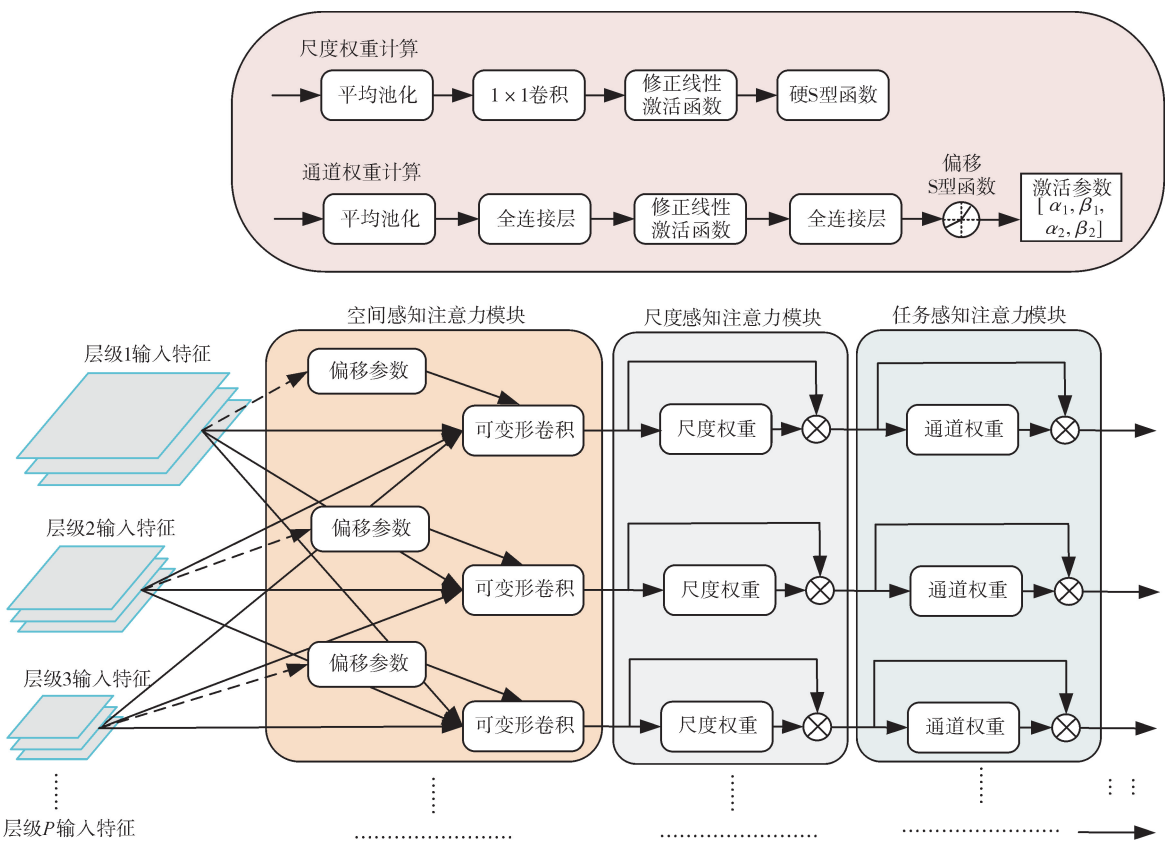

图3　动态头模块网络结构图

Fig. 3　Structure of dynamic head block network

过程可表示为：

$$\Omega(\boldsymbol{X})\cdot\boldsymbol{X}=\frac{1}{P}\sum_{p=1}^{P}\sum_{k=1}^{K}w_{p,k}\cdot\boldsymbol{X}\left(p;\ o_k+\Delta o_k;\ c\right)\cdot\Delta m_k \tag{8}$$

其中：p表示特征金字塔的层级索引，用于遍历不同的特征层；k表示稀疏采样位置的索引，用于标识在特定层级中进行稀疏采样的位置；K代表稀疏采样位置的总数量；$w_{p,k}$为用于特征层加权的自学习权重；o_k为稀疏采样位置的空间索引；Δo_k为o_k处的自学习空间偏移量，用于调整采样位置，使模型聚焦关键特征区域；$\boldsymbol{X}\left(p;\ o_k+\Delta o_k;\ c\right)$表示在层级$p$，第$c$层通道处，输入特征$\boldsymbol{X}$在$o_k+\Delta o_k$空间位置处的特征值；$\Delta m_k$为输入特征$\boldsymbol{X}$在$o_k$处的自学习重要性权重。

尺度感知注意力模块$\Phi(\cdot)$为不同的特征层生成加权融合系数，由自适应平均池化层、1×1卷积层、修正线性单元（rectified linear unit，ReLU）激活函数层和硬S型函数层组成，计算过程可表示为：

$$\Phi(\boldsymbol{X})\cdot\boldsymbol{X}=\tau\left(f\left(\left(\sum_{O,C}\boldsymbol{X}\right)\Big/(O\cdot C)\right)\right) \tag{9}$$

其中：$\left(\sum_{O,C}\boldsymbol{X}\right)\Big/(O\cdot C)$代表输入特征$\boldsymbol{X}$在空间及通道维度上的平均池化操作；$f(\cdot)$即线性映射，代表1×1卷积；$\tau(x)=\max(0,\ \min(1,(1+x)/2))$为硬S型函数。

任务感知注意力模块$\Psi(\cdot)$将输入特征$\boldsymbol{X}$通过自适应平均池化层和一系列全连接层生成初始激活参数，随后通过偏移的S型函数将初始激活参数归一化到［−1，1］的范围，从而得到两组特定的激活参数$(\alpha_1,\beta_1),(\alpha_2,\beta_2)$。最后，将特定激活参数应用于特征切片$\boldsymbol{X}_c$，即输入特征$\boldsymbol{X}$在通道层$c$处的特征张量。模块最终输出为动态激活结果的最大值，即：

$$\Psi(\boldsymbol{X})\cdot\boldsymbol{X}=\begin{cases}\alpha_1(\boldsymbol{X})\cdot\boldsymbol{X}_c+\beta_1(\boldsymbol{X}),\alpha_1(\boldsymbol{X})\cdot\boldsymbol{X}_c+\beta_1(\boldsymbol{X})\geqslant\alpha_2(\boldsymbol{X})\cdot\boldsymbol{X}_c+\beta_2(\boldsymbol{X})\\ \alpha_2(\boldsymbol{X})\cdot\boldsymbol{X}_c+\beta_2(\boldsymbol{X}),\alpha_2(\boldsymbol{X})\cdot\boldsymbol{X}_c+\beta_2(\boldsymbol{X})>\alpha_1(\boldsymbol{X})\cdot\boldsymbol{X}_c+\beta_1(\boldsymbol{X})\end{cases} \tag{10}$$

通过上述空间、尺度及任务感知注意力模块的协同作用，动态检测头能更好地学习融合不同语义级别、形状和尺度的特征，以适应复杂瓷砖表面瑕疵检测任务。

2 结果及讨论

2.1 瓷砖表面缺陷数据集

实验采用阿里云天池官方提供的瓷砖瑕疵检测数据集进行训练，验证模型改进效果。该数据集包含5 388张存在表面瑕疵的产线实拍瓷砖图像，包括边异常、角异常、白色点、浅色块、深色点和光圈瑕疵共6种类型的瑕疵。但原图像尺寸过大，会增加模型的计算复杂度，导致模型训练难以收敛。且绝大部分瓷砖表面瑕疵的宽度/高度尺寸都小于原图像宽度/高度尺寸的1%，使得高精度检测难以实现。为解决上述问题，将原数据集中所有图像均以640×640像素切片，最终得到共19 960张图像。并按8∶1∶1的比例划分为训练集、测试集和验证集，重新建立数据集。

2.2 实验环境及评价指标

本文训练和测试所使用的GPU型号为英伟达RTX4090，显存为24 GB，CPU型号为英特尔至强E5-2650，操作系统为乌班图20.04。训练时，图片已经过切片预处理，输入大小为640×640像素，训练轮次为400轮，批大小为64，网络初始学习率为0.01。

本文采用平均精度、参数量、计算量、每秒处理图像帧数这4个常用检测指标来衡量模型的性能。其中，参数量指模型中可学习参数的总数；计算量以每秒浮点运算次数为度量，用于衡量模型推理过程中的计算复杂度；每秒处理图像帧数则反映了模型的实时推理速度。

2.3 消融实验

为评估本文所提改进策略对模型综合性能的影响，本文在瓷砖瑕疵检测数据集上进行了消融实验。实验以YOLOv8n为基准模型，依次评估了在基准模型上采用替换主干网络、引入动态检测头、优化颈部网络等改进策略的效果，并与结合所有改进策略的优化模型进行对比。实验结果如表1所示。

实验结果表明，各改进策略均显著提升了模型的平均精度。将基准模型YOLOv8n的主干网络替换为高效视觉变换器网络，模型的平均精度提升了4.2%。引入动态检测头将模型的平均精度提高了5.9%。为适应小目标检测需求，将输入动态检测头的所有特征图通道数统一为最小，故引入动态检测头后模型参数量和每秒浮点运算次数分别减少了16.7%和3.7%。此外，通过优化颈部网络，进一步提升了模型的特征融合能力，模型的平均精度提高了3.7%。综合所有改进策略后，模

表1　各改进策略对模型性能影响对比

Table 1　Comparison of the impact of improvement strategies on model performance

模型	平均精度 / %	参数量 / 10^6	每秒浮点运算次数 / 10^9	每秒处理图像帧数
基准模型	59.9	3.0	8.1	123
基准模型（替换主干网络）	64.1	4.0	9.4	65
基准模型（引入动态检测头）	65.8	2.5	7.8	68
基准模型（优化颈部网络）	63.6	4.0	9.7	77
本文模型	71.5	4.6	9.3	48

型的平均精度提升了11.6%。

为更直观地评估各增强策略对模型性能提升的贡献，本文选择了部分具有代表性的瓷砖瑕疵图像，并对比了不同增强策略应用前后在对应模块的输出特征图。其中，主干网络输出特征图为主干网络完成特征提取后得到的特征图。高效视觉变换器网络能更有效地聚焦关键局部特征，减少背景干扰。如图4所示，替换主干网络后的模型能更准确地区分出复杂背景中的瑕疵特征。颈部网络输出特征图是主干网络提取特征经过颈部网络融合后的综合表示。本文通过将普通卷积替换为空间到深度卷积，并在颈部网络末端添加空间通道重组卷积模块，进一步优化了颈部网络的特征融合能力。如图5所示，优化后的颈部网络能更好地聚焦目标的空间位置分布，并关注到了更多小目标。此外，动态检测头通过自适应调整注意力分布，可以更好地处理不同尺度和形状的多目标任务，如图6所示，引入动态检测头后，模型在处理多目标时能更精确地捕捉特征信息。

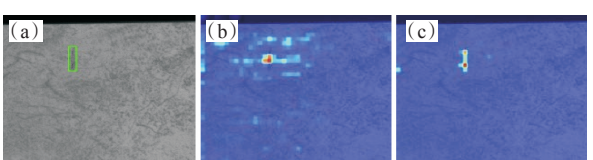

图4　（a）原始图像及瑕疵标注；替换主干网络前（b）和替换后（c）的输出特征图

Fig. 4　(a) Original image with defect annotations; Before replacing the backbone (b) and after replacing (c) output feature map

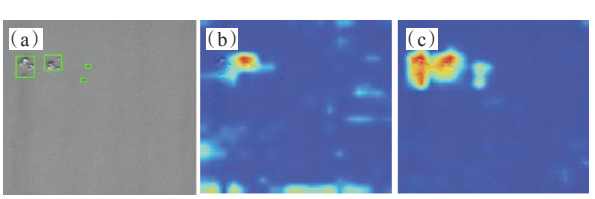

图5　（a）原始图像及瑕疵标注；优化颈部网络前（b）和优化后（c）的输出特征图

Fig. 5　(a) Original image with defect annotations; Before optimizing neck network (b) and after replacing (c) output feature map

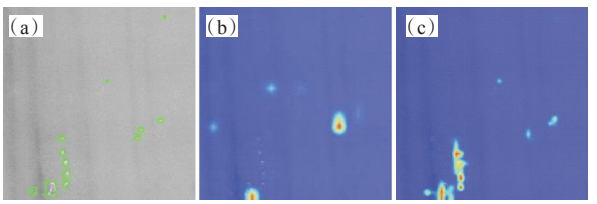

图6（a）原始图像及瑕疵标注；添加动态头前（b）和添加后（c）的输出特征图

Fig. 6　(a) Original image with defect annotations; Before adding dynamic head (b) and after replacing (c) output feature map

2.4　对比实验

为评估改进后算法的性能，本文进行了对比实验，将改进后的算法与多种经典的单阶段目标检测模型进行了比较。本次实验选择SSD、YOLOv5n、YOLOv5s、YOLOv6n[16]、YOLOv6s[16]、YOLOv7-tiny[17]及实时检测变换器[18]（real time detection transformer，RT-DETR）模型进行对比试验。所有模型均在相同实验条件下训练。如表2所示，本文提出的模型在瓷砖瑕疵检测任务中达到了71.5%的最佳平均精度。其检测精度分别比YOLOv5n、YOLOv5s、YOLOv6n、YOLOv6s、YOLOv7-tiny、YOLOv8n、YOLOv8s高出了14.1%、11.7%、16.0%、9.2%、19%、11.6%、3.4%。虽然本文模型参数量和每秒浮点运算次数略高于YOLOv5n、YOLOv6n等YOLO超轻量级模型，但在精度、参数量及计算量上综合表现均最优。与同为视觉变换器架构的RT-DETR模型相比，本文模型在各方面也均更优。综上，本文模型在提升检测精度的同时仍能满足实时检测需求。

图7和图8展示了在测试集上不同算法的检测结果对比。模型检测结果包括检测框和对应标签。检测框表示模型检测到的瑕疵目标在图像中的具体位置，每个检测框的标签包含了预测的瑕疵类别和相应的置信概率。从图7和图8中可以看出，本文提出的模型DViT-YOLO在应对复杂场景，如同时存在多个瑕疵或背景较为复杂时，检测表现更佳，漏检及误检率最低，且预测目标类别的置信度最高。

表2 各模型性能评估指标结果比较

Table 2 Comparison of performance evaluation indicators for different models

模型	平均精度 / %	参数量 / 10^6	每秒浮点运算次数 / 10^9	每秒处理图像帧数
SSD[3]	39.8	91.6	116.3	52
YOLOv5n	57.4	2.5	7.1	114
YOLOv5s	59.8	9.1	23.8	105
YOLOv6n[16]	55.5	4.2	11.8	116
YOLOv6s[16]	62.3	16.3	44.0	122
YOLOv7-tiny[17]	52.5	6.0	13.2	98
YOLOv8n	59.9	3.0	8.1	123
YOLOv8s	68.1	11.1	28.4	110
RT-DETR[18]	33.8	32.0	103.5	32
本文模型	71.5	4.6	9.3	48

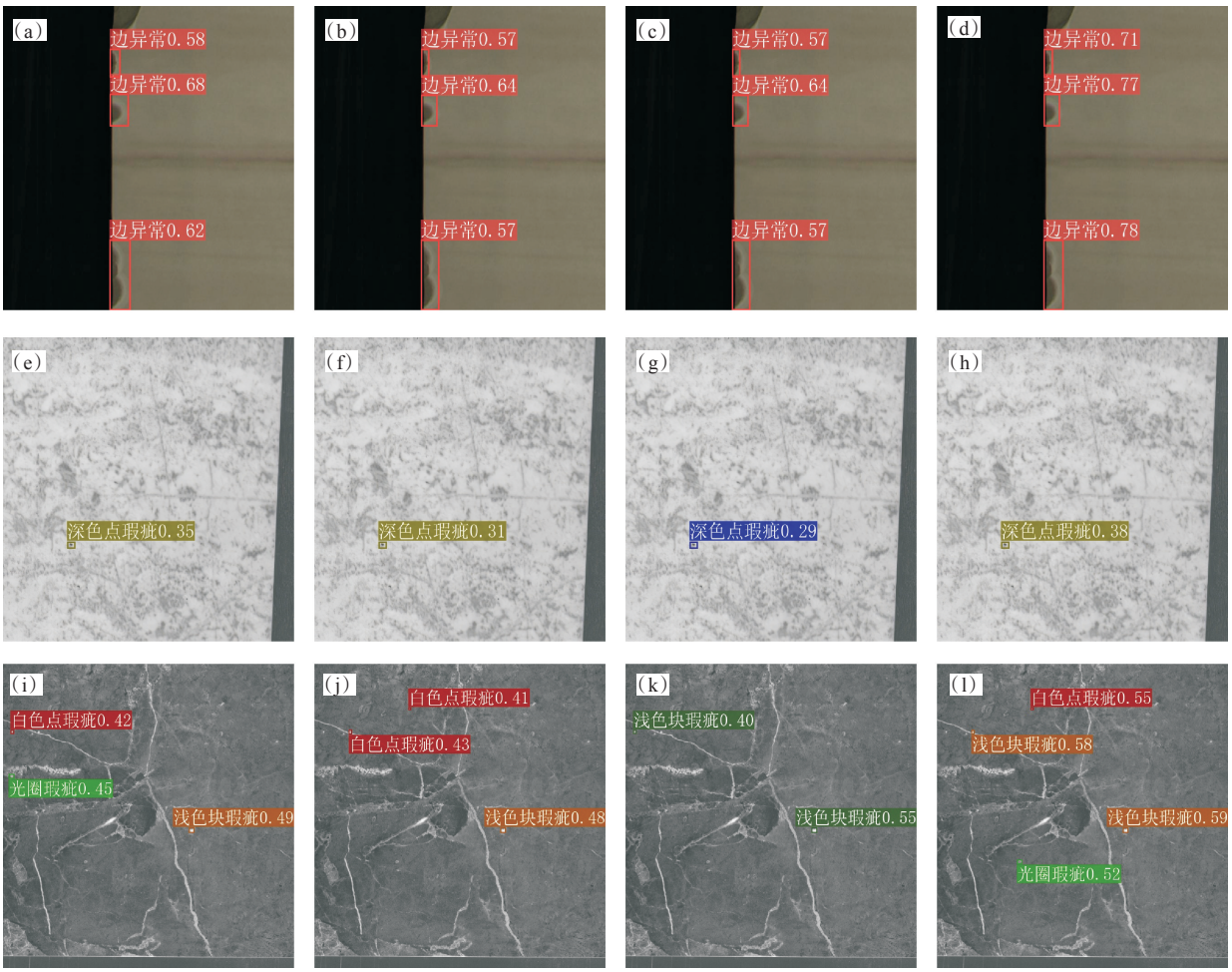

图7 YOLO超轻量级系列模型检测结果比较：(a、e、i)YOLOv5n，(b、f、j)YOLOv6n，(c、g、k)YOLOv7-tiny，(d、h、l)YOLOv8n

Fig. 7 Comparison of detection results among YOLO ultra-lightweight series models: (a, e, i) YOLOv5n, (b, f, j) YOLOv6n, (c, g, k) YOLOv7-tiny, (d, h, l) YOLOv8n

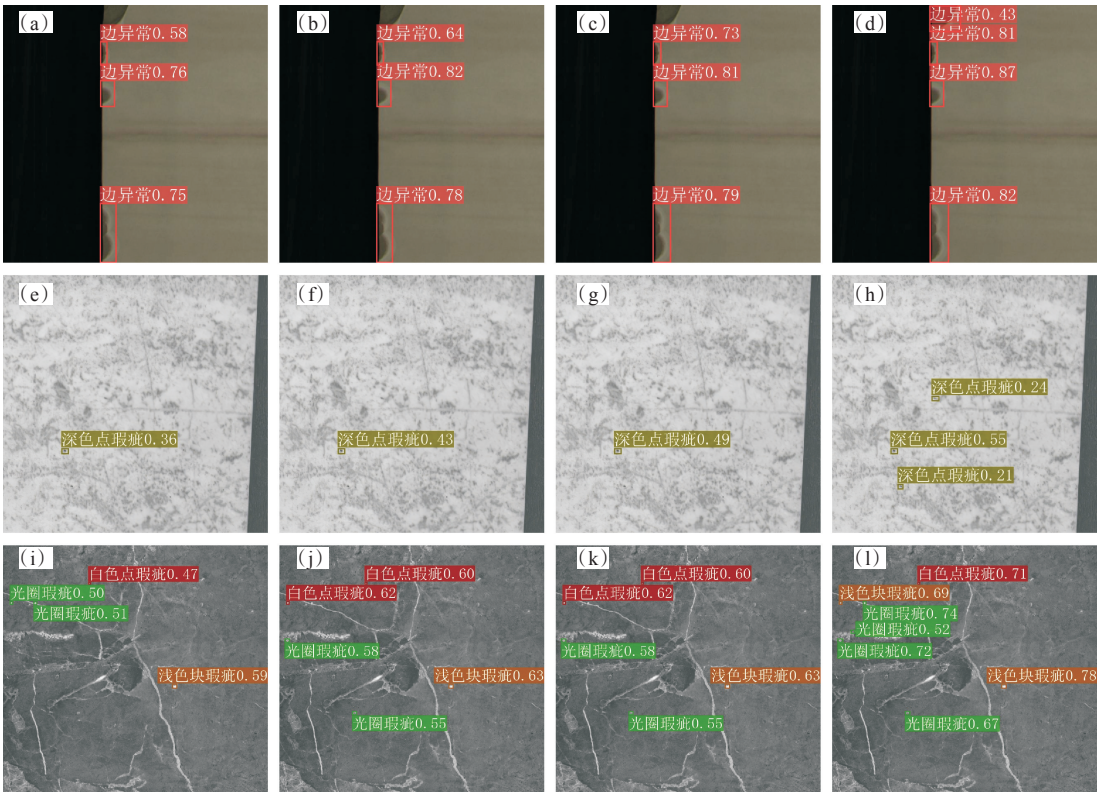

图8　YOLO轻量级系列模型与本文模型检测结果比较：(a、e、i)YOLOv5s，(b、f、j)YOLOv6s，(c、g、k)YOLOv8s，(d、h、l)**本文模型**

Fig. 8　Comparison of detection results between YOLO lightweight series models and our model: (a, e, i) YOLOv5s, (b, f, j) YOLOv6s, (c, g, k) YOLOv8s, (d, h, l) our model

3　结　论

本文提出了一种高性能、轻量级的瓷砖表面瑕疵检测模型DViT-YOLO，针对瓷砖表面瑕疵面积小、形状不规则、尺度多变、与复杂花纹背景难以区分等特点，进行了替换主干网络、使用空间到深度卷积替换普通下采样卷积以保留更多细粒度信息、颈部网络末端添加空间通道重组卷积模块优化特征表达、引入动态检测头等多方面的改进。

改进后模型的平均检测精度远优于原YOLOv8n模型及其他经典单阶段目标检测算法，并在保持与YOLOv8n相近的计算复杂度下，检测精度超越了更大规模的YOLOv8s模型。在瓷砖表面微小瑕疵检测、密集瑕疵检测及复杂背景下的瓷砖表面瑕疵检测等挑战性任务中表现出色，显著降低了漏检率和误检率。此外，该模型参数量较小，可以很好地满足工业场景中的边缘设备部署需求。每秒处理图像帧数为48，能够满足基本的实时检测需求，但仍有较大的优化空间。未来的工作将致力于进一步优化模型结构，使用模型剪枝、量化等技术，提升模型检测速度，使其能更好地落地实际应用场景。

参考文献

[1] WANG K, LI Z R, WANG X. Concatenated network fusion algorithm (CNFA) based on deep learning: improving the detection accuracy of surface defects for ceramic tile[J]. Applied Sciences, 2022, 12(3): 1249.

[2] CAO T L, SONG K C, XU L K, et al. Balanced multi-scale target score network for ceramic tile surface defect detection [J]. Measurement, 2024, 224: 113914.

[3] KUMAR A, ZHANG Z J, LYU H B. Object detection in real time based on improved single shot multi-box detector algorithm[J]. EURASIP Journal on Wireless Communications and Networking, 2020, 2020: 204.

[4] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 779-788.

[5] REN S Q, HE K M, 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, 2017, 39

(下转第555页)