DOI:10.19652/j.cnki.femt.2305629

基于分部特征计算的轻量化非结构目标检测

金友祺[1] 赵 津[2] 刘 畅[2] 孙念怡[2]
(1.贵州大学现代制造技术教育部重点实验室 贵阳 550025;2.贵州大学机械工程学院 贵阳 550025)

摘 要:针对非结构化场景(工地、矿场)缺少特殊目标的通用数据集、复杂特征难以准确提取以及计算复杂度高的问题,构建了一个面向非结构场景的特殊目标检测数据集,并提一种轻量化目标检测模型 YOLO-PT,以极低的计算量达到了较高的检测精度。通过构建分部特征计算(partial feature calculation, PFC)模型减少特征冗余信息的计算,并引入了多头自注意力机制来增强复杂特征的提取精度,同时设计多通道金字塔结构对多尺度特征进行渐进式融合,提高复杂对象的识别精度。最后在非结构场景进行实验验证,结果表明,所提出方法仅在 4.3×10^6 的参数量下就达到了 53%的准确率,在精度、参数量以及浮点运算量上均优于其他方法。

关键词:非结构场景;多头注意力;目标检测;分部特征计算;数据集
中图分类号: TP391.41; TN957.52 **文献标识码**:A **国家标准学科分类代码**: 520.604

Efficient object detection method with partial calculation for unstructured scenes

Jin Youqi[1] Zhao Jin[2] Liu Chang[2] Sun Nianyi[2]
(1. Key Laboratory of Advanced Manufacturing Technology of the Ministry Education, Guizhou University, Guiyang 550025, China; 2. School of Mechanical Engineering, Guizhou University, Guiyang 550025, China)

Abstract: To address the challenges of the absence of shared datasets covering unique targets in unstructured scenes (such as construction sites and mining sites), the difficulty in precise extraction of complex features, and the high computational complexity, this paper creates a dedicated object detection dataset for unstructured scenes. We present a lightweight object detection model named YOLO-PT, which attains high detection accuracy while requiring minimal computational resources. We mitigate the computation of redundant feature information by developing a partial feature calculation (PFC) model. We also incorporate a multi-head self-attention mechanism to enhance the precision of complex feature extraction and design a multi-channel pyramid structure for the gradual fusion of multi-scale features, thereby improving the recognition accuracy of complex objects. Finally, experimental validation is conducted in unstructured scenarios. The results demonstrate that the method proposed achieves the accuracy of 53% with a mere 4.3×10^6 parameters, outperforming other methods in terms of accuracy, the number of parameters and floating-point operations.
Keywords: unstructured scene; multi-head self-attention; object detection; partial feature calculation; dataset

0 引 言

目标检测是计算机视觉的一项重要任务[1],已经广泛应用于各个领域[2],相关研究也已经取得了显著的成果,然而非结构场景目标检测的研究仍然较少。在面向非结构环境的作业中(如挖掘机、装载机作业),准确及时识别其他工程装备、物料、岩石和地下管道等特殊对象在避免工程事故、提高工作效率上具有重要意义。非结构场景指没有明显规律、复杂多变的场景,其检测目标具有如下特点:1)多样性和复杂性,非结构场景往往包含各种各样的检测对象,这些对象可能在尺寸、形状、颜色和材质等方面都能存在较大差异;2)不规则性,非结构场景中,目标的位置和排列通常是不规则的,例如杂乱的石块,以及分部不均匀的物料堆;3)背景干扰,非结构场景中的目标往往受到背景的干扰,这种背景可能包括其他物体(被土壤半遮盖的管道)、光照变化等;4)尺度变化,非结构场景中同一

收稿日期:2023-10-10

对象通常存在不同的尺度跟大小，以及远近的视角变换。相比与规整的街道等结构化场景，非结构场景的特点使得目标检测成为一项具有挑战性的任务，需要综合考虑多种因素来实现准确和可靠的目标检测。

目前，基于深度学习的目标检测算法主要分为两类[3]，两阶段目标检测算法和一阶段目标检测算法。两阶段目标检测算法分粗定位和精细分类两个步骤，代表算法为 Faster R-CNN[4]，它们往往推理速度较慢，在实时性上有较大的劣势。一阶段目标检测算法直接在输出层对检测框类别和位置进行回归，代表算法包括 SSD[5]、YOLO 系列[6-7]等，这些算法实现了更快的推理速度，已经成为了目标检测领域的热门。除此之外，Vaswani 等[8]在 2017 年提出了基于多头自注意力机制(multi-head self-attention，MHSA)的 Transformer，有效捕获了序列的全局关系，并在自然语言处理中大放异彩。受此启发，Dosovitskiy 等[9]提出基于 MHSA 的视觉模型 ViT(vision transformer)，将图像分割成固定大小的图像补丁，进而转换为序列数据使用 Transformer 模型进行处理，并激发了后续一系列视觉 Transformer 的研究。这些方法利用注意力机制捕捉图像中的全局关系，在一些任务中取得了与卷积神经网络(convolutional neural networks，CNN)方法相媲美甚至更优越的性能，但如何降低计算复杂度仍然是面临的难题。在工程应用中，徐先峰等[10]针对在施工场所佩戴安全帽检测易受复杂环境干扰的问题，构建了复杂背景下的安全帽数据集进行研究。刘洺睿等[11]针对矿场无人机小目标检测不准确的问题，利用无人机低空遥感影像构建了矿用卡车数据集，进行了复杂环境小目标检测的研究。黄璐等[12]针对复杂场景下岩石形状多样的特点，构建了岩石、沙丘和沙地数据集，提出了一种新的地外环境下非结构化目标智能辨识方法。以上研究虽然针对复杂环境的目标检测进行了一些研究，但只进行了安全帽、石块这些单一目标的检测，没有考虑实际工程中非结构目标多样性的影响，泛用性不高。Wang 等[13]构建了建筑工地可回收材料的数据集，提出了基于 Transformer 的两阶段可回收材料检测模型，虽然实现了多个类别的识别，但是大量的计算使其在精度与效率之间难以平衡。

非结构场景目标检测仍然面临着诸多挑战，例如非结构场景数据集的缺失以及不同尺寸、不规则形状等复杂特征难以准确提取，特别是在环境相似度较高、类别形状复杂多变的情况下，以及存在目标遮挡(如半掩埋的地下管道)、形变、视角变化等情况时，现有的方法往往难以取得理想的效果。同时，推理速度与精度的平衡仍然是一大难题，虽然一些方法达到了较好的检测效果，但为了保持较高准确性牺牲了实时性能，且在边缘设备难以部署。为了解决以上问题，本文提出了基于多头注意力机制的分部计算的特征提取模型，兼顾了全局特征提取能力与计算复杂度，有效地捕获了通道和空间方面的特征。本文还设计了多通道金字塔结构，有效避免了非相邻层高级特征与低级特征融合间隙大的问题，提高了多尺度特征融合精度并增强了多维度相互作用的能力。同时构建了非结构复杂场景工程装备数据集，对本文所提的模型以及其他主流模型进行了验证。

1 分部特征计算模型

1.1 视觉 Transformer

与传统 CNN 直接处理图像网格数据不同，ViT 将图像的像素转化为序列，并利用多头自注意力机制来捕获图像中的空间和语义信息，主要由 Embedded Patches 和 Transformer Encoder 组成，如图 1 所示。

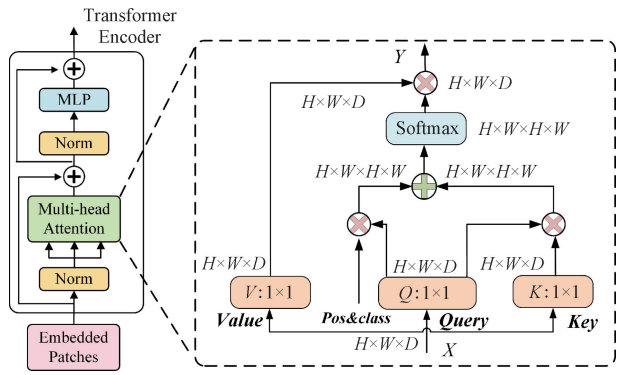

图 1 Vision Transformer 主要模块

Fig. 1 Vision Transformer main modules

对于标准的 Transformer，要求输入向量序列，而图像数据为{H，W，C}格式的三维矩阵，因此需要通过编码层对数据进行变换，通常将一张图像按给定大小分为多个维度相同的 Patches，然后通过线性映射将每个 Patches 映射到一维向量中，最后将一维向量作为 Transformer 编码器的输入将其进行自注意力计算，自注意力计算公式如下：

$$Attention = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}\cdot\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right) \tag{1}$$

$$\boldsymbol{Y} = Attention\cdot\boldsymbol{V} \tag{2}$$

式中：$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 为映射后的一维向量；d_k 为向量的维度。

在进行计算前需要添加类别向量(class token)和位置编码(position embedding)这两个可训练参数，类别向量承载全局语义信息，而位置编码便于模型学习理解每个位置的相对重要性和关系，使其在处理图像数据时能够融合全局信息。

1.2 分部特征计算模

在计算机视觉任务中，通常使用卷积层对输入的图像进行特征提取得到一系列特征图，每个特征图代表了输入图像在不同语义级别上的特征表示，然后根据特征图中的不同特征进行进一步的处理和分类。许多研究[14]发现特征图在不同通道之间具有高度的相似性，如图 2 所示。

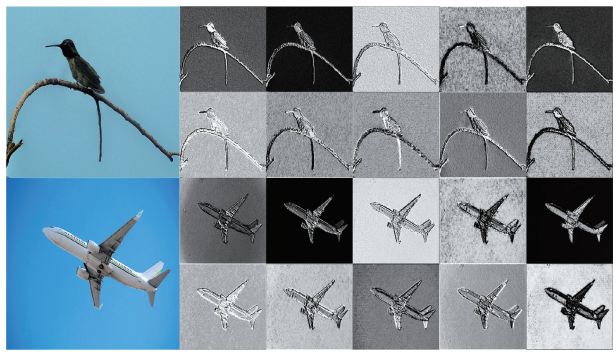

图 2 可视化特征图

Fig. 2 Visualized feature maps

在特征提取过程中，相似特征的重复计算是不必要的，然而少有学者研究如何以简单有效的方法减少这种冗余信息的重复处理。基于以上问题，本研究提出了一种分部特征计算(partial feature calculation，PFC)方法来减少这种重复特征的处理。如图 3 所示，将 Transformer 模块仅应用于一部分输入通道来进行空间特征提取，其余通道则直接与计算后的特征进行拼接，然后使用点卷积(point-wise convolution，PWConv)[15]处理不同通道的特征信息。本文保持通道数量不变而不是从特征图中删除它们，因为这些信息对于后续计算有用，在分部计算后设置 PWConv，将不同通道的特征进行线性组合，使得特征信息流过所有通道，从而融合不同层次、不同来源的信息。也可以用 CNN 代替 Transformer 模块实现极致的轻量化，但是精度上会出现一定的损失。

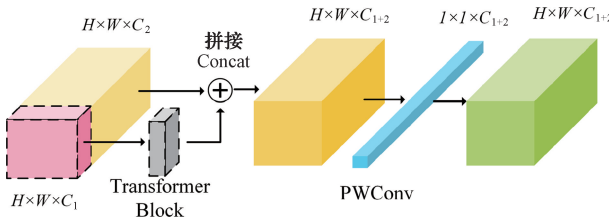

图 3 分部特征计算模块

Fig. 3 Partial feature calculation modules

参考 ResNet[16]的残差结构，本文设计了 Bottle-PFC 模块，来有效地保持了提取特征信息的一致性，结构如图 4 所示。该模块将输入分为 3 个分支，CBS 模块、局部计算模块以及一支残差结构，然后 3 个分支进行拼接，该结构在提取全局特征的同时有效减少参数量与浮点数，残差结构使网络能够在减少梯度消失与梯度爆炸的情况下到达更深的程度，自注意力机制的全局关联性又极大的提高特征提取精度，二者结合将在提高计算效率的同时极大地改善了特征表示。

2 多通道轻量化目标检测模型

2.1 多通道金字塔结构

在检测任务中，物体通常以不同的尺度出现在视野

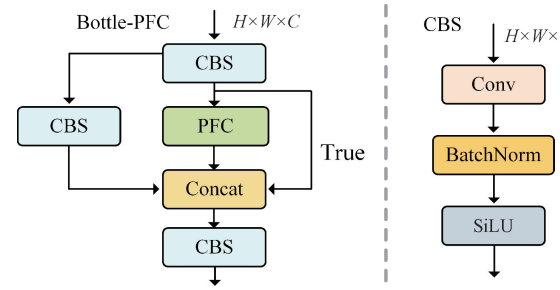

图 4 Bottle-PFC 模块和 CBS 模块

Fig. 4 Bottle-PFC module and CBS module

中，频繁的下采样使得部分小尺寸物体特征丢失从而导致整体精度下降[17]，为了有效捕获不同尺度的语义信息，特征金字塔(feature pyramid network，FPN)采用自顶向下的方式将高层特征传递到低层特征，实现不同层次特征的融合，并结合深层语义信息与浅层位置信息提高检测精度[18]。然而这些方法没有考虑非相邻特征信息差别较大的问题，削弱了非相邻水平的融合效果[19]。针对此问题，本文设计了多通道融合金字塔结构(multi-channel feature pyramid network，MFPN)，在骨干网络的自底向上特征提取过程中，渐进式的集成低级特征和高级特征。如图 5 所示，该结构先融合两个低级别邻近的特征，然后再融合高级别特征，有效避免非相邻级别间的较大语义间隙，同时高级特征与低级特征相互补充，提高特征提取的精度。为了对齐不同特征维度，采用双线性插值的方式进行上采样，设置不同的卷积核与步长实现相应的下采样。在特征融合之后继续使用若干个残差单元(ResNet block，Resb)来学习特征，每个残差单元由两个 CBS 结构(图 4)与 1 个快捷连接组成。

2.2 整体模型

本文所提模型主干由 4 个阶段组成，每个阶段由若干个 Bottle-PFC 模块组成，同时后面都有一个步长为 2 的 3×3 卷积用于下采样，实现多级特征提取。在第 4 阶段之后设置空间金字塔池化(spatial pyramid pooling-fast，SPPF)结构，对输入特征图进行分层池化和特征拼接，以适应不同大小的输入特征。本文使用 MFPN 作为颈部来实构建多尺度信息，加强非相邻特征层之间的信息互补，准确融合不同阶段的特征信息。最后，使用标准的 YOLO 格式检测头进行最后的检测框生成，其损失由分类损失、目标损失和回归损失 3 个部分组成。其中分类损失、目标损失均采用二元交叉熵损失(binary cross-entropy loss，BCEL)，考虑边界框的形状和位置信息，定位损失采用完整交并比损失(complete intersection over union，CIOU)以更好地反映目标的真实位置，该模型命名为 YOLO-PT，具体结构如图 6 所示。

为了在速度和准确性之间实现更好的权衡与更多的选择，将主干网络的多级特征通道深度和宽度进行压缩，提出了 YOLO-PT 的 3 种变体模型，即 YOLO-PT-N、

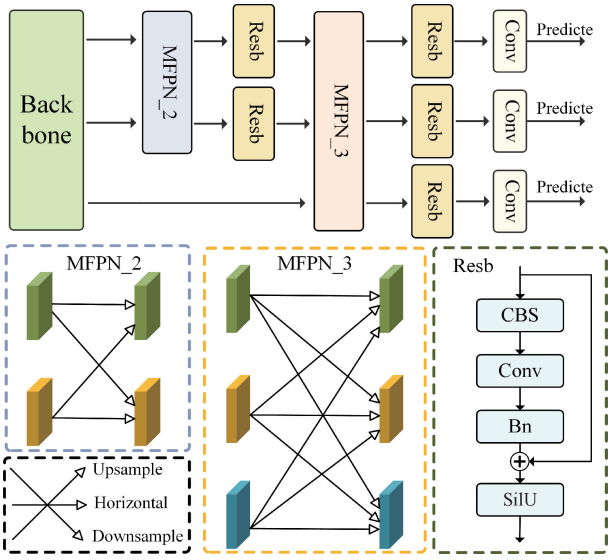

图 5　多通道融合金字塔结构

Fig. 5　Multi-channel feature pyramid structure

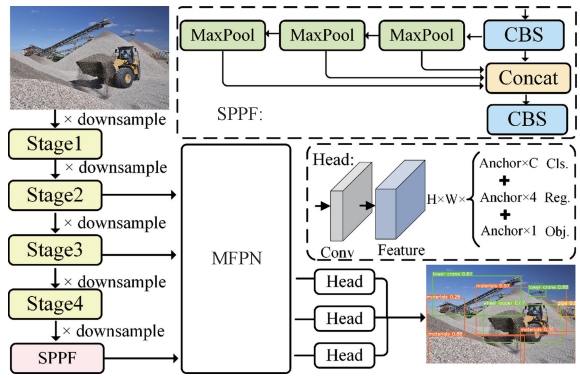

图 6　YOLO-PT 结构框图

Fig. 6　YOLO-PT structure

YOLO-PT-S、YOLO-PT-M。不同规模的配置如表 1 所示，其中 C 指特征图的通道数，L 指 Bottle-PFC 模块的叠加数量。

表 1　YOLO-PT 的简要配置

Table 1　Summary configuration for YOLO-PT

模型	输出尺寸	YOLO-PT-N	YOLO-PT-S	YOLO-PT-M
Stage 1	$\frac{H}{2}\times\frac{W}{2}\times C$	$C=32$，$L=1$	$C=64$，$L=1$	$C=96$，$L=1$
Stage 2	$\frac{H}{4}\times\frac{W}{4}\times C$	$C=64$，$L=2$	$C=128$，$L=2$	$C=192$，$L=4$
Stage 3	$\frac{H}{8}\times\frac{W}{8}\times C$	$C=128$，$L=3$	$C=256$，$L=3$	$C=384$，$L=6$
Stage 4	$\frac{H}{16}\times\frac{W}{16}\times C$	$C=256$，$L=1$	$C=512$，$L=1$	$C=768$，$L=2$
参数量/($\times10^6$)		1.1	4.3	10.3
浮点数/GFLOPs		3.4	10.7	26.2

3　模型验证与仿真结果分析

3.1　数据集设计

针对目前非结构工程环境数据集缺失的问题，本文集聚焦于工程装备、石块、地下管道等复杂检测对象构建了非结构工程场景数据集，该数据集由真实施工场景采集的视频图像与网络收集的相关图片组成，共 4 700 张分辨率为 1 920×1 080 pixels 的彩色图像，如图 7 所示。

本文使用开源软件 Labelme 进行检测框的标注并生成相应的标签文件。不同于现有的其他数据集，本文主要对挖掘机、倾倒卡车、轮式装载机、地下管道、物料堆、岩石和塔吊共 7 个非结构工程场景特殊检测对象进行标注，为非结构工程场景(建筑工地、矿区等)的检测与规划等任务提供有效支撑。

图 7　非结构场景数据集部分图例

Fig. 7　Partial illustrations for unstructured scenario dataset

3.2 训练策略与实验设置评估

本文采用随机梯度下降(stochastic gradient descent, SGD)作为优化函数,设置初始学习率为0.01,每轮训练后衰减0.01,动量系数设置为0.9,权重衰减系数为0.000 5。并在一张NVIDIA RTX 3060Ti和4张NVIDIA RTX 2080Ti上进行训练。在训练过程中,采用随机旋转、平移、翻转等操作进行数据增强,每项操作以0.5的概率应用。设置训练集(包含4 200张图像)进行训练,使用value集合(包含300张图像)进行验证,并在test集合(包含200张图像)进行测试。所有配置的模型均在不使用其他大规模数据集训练权重的情况下训练300轮。

在测试评估方面,使用平均精度均值(mAP)作为主要指标。mAP是所有类别平均精度(AP)的均值,AP是精确率-召回率曲线所包围的区域的面积,其值越大网络模型的整体性能就越好。精确率、召回率和平均精度的具体公式如下:

$$P = \frac{TP}{TP + FP} \tag{3}$$

$$R = \frac{TP}{TP + FN} \tag{4}$$

$$AP = \int_0^1 P(R)\mathrm{d}R \tag{5}$$

式中:TP(真阳性)表示模型检测后的样本与标签样本匹配;FP(假阳性)表示模型检测后的样本不包括在标签样本中;FN(假阴性)表示未检测到标签样本。

3.3 主要结果分析

进行模型性能测试与对比实验,本文所提模型在参数量、浮点运算量、mAP指标上均取得了优异的性能表现,结果如表2所示。

表2 在非结构数据集上与YOLOv5的表现对比

Table 2 Performance comparison with YOLOv5 on unstructured dataset

模型	挖掘机	倾倒卡车	物料堆	管道	石块	塔吊	轮式装载机	参数量/($\times10^6$)	浮点数/GFLOPs	mAP@0.5	mAP
YOLO-PT-N	0.281	0.638	0.441	0.430	0.309	0.241	0.822	1.1	3.4	0.726	0.450
YOLOv5n	0.315	0.611	0.387	0.399	0.274	0.277	0.796	1.8	4.2	0.673	0.437
YOLO-PT-S	0.373	0.674	0.490	0.507	0.370	0.432	0.862	4.3	10.7	0.763	0.530
YOLOv5s	0.422	0.657	0.462	0.453	0.329	0.344	0.850	7.0	15.8	0.731	0.502
YOLO-PT-M	0.446	0.677	0.505	0.517	0.394	0.360	0.882	10.3	26.2	0.793	0.541
YOLOv5m	0.440	0.675	0.487	0.495	0.347	0.372	0.885	21.2	49.0	0.733	0.531

由表2可得,YOLO-PT-M比YOLOv5m在检测精度高1%的同时参数量减少了51%,浮点数也减少了46.5%,计算复杂度减少1/2的同时实现了性能上的超越。对于不同的检测种类,面对较为常规的挖掘机、倾倒卡车等工程机械装备时,YOLOv5与YOLO-PT在检测精度上不相上下;但面对物料堆、管道、石块等尺寸多变、与背景极其相似的非结构复杂检测对象时,YOLO-PT要明显优于YOLOv5,3种不同规格的模型在精度上均有较大的领先。是因为石块等非结构检测对象特征更加复杂,相比于YOLOv5,YOLO-PT的长距离建模能力能够更好地进行上下文信息提取,即便是不规则的石块与半掩埋的管道也能精确识别。在训练方面,如图8所示,YOLO-PT-M收敛速度也更快更稳定。

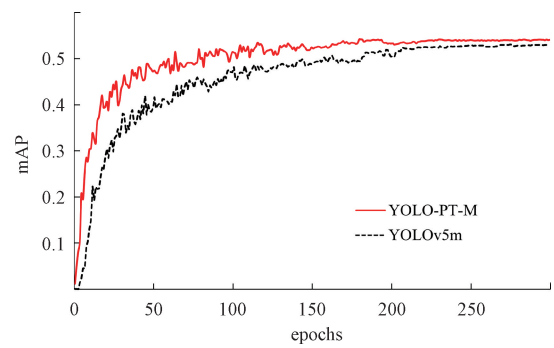

图8 mAP曲线

Fig. 8 mAP curve

在召回率方面,YOLO-PT也均高于YOLOv5,这意味着YOLO-PT能够更全面、更准确地检测出非结构环境中的复杂对象,不同检测对象召回率对比结果如图9所示。

此外,YOLO-PT-S在参数量、浮点数比YOLOv5m分别低80%、78%的情况下达到了相同的精度,且效率更高;YOLO-PT-N模型仅在1.1 M参数量的情况下也能达到不错的效果,低参数量、低浮点数使其非常方便部署于边缘设备,应用于多种计算资源受限的场合。

3.4 消融实验

为了更好地理解YOLO-PT,设计了消融实验来验证不同模块对检测性能的影响,包括模型参数、浮点运算量、Bottle-PFC模块与多通道金字塔模块等。

模型叠加,YOLO-PT采用5个阶段的下采样将图像尺度由(640,640)变为(40,40),设计5种不同的堆叠结构

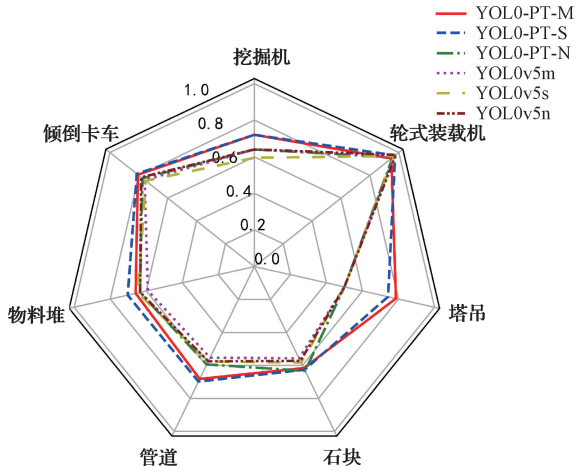

图 9　不同检测对象召回率对比

Fig. 9　Comparison of recall rates for different objects

进行实验：(4 × Bottle-PFC)、(3 × Bottle-PFC + 1 × C3Net)、(2 × Bottle-PFC + 2 × C3Net)、(1 × Bottle-PFC + 3 ×C3Net)和(4 × C3Net)(C3Net：YOLOv5 主干网络模块)，结果如表 3 所示。

表 3　不同数量的 Bottle-PFC 模块对精度的影响

Table 3　The influence of Bottle-PFC modules on precision

模型	参数量/($\times10^6$)	浮点数/GFLOPs	mAP
YOLOv5n(4 ×C3Net)	1.9	4.5	0.437
YOLOv5n+1×Bottle-PFC	1.3	3.6	0.477
YOLOv5n+2×Bottle-PFC	1.3	3.4	0.471
YOLOv5n+3×Bottle-PFC	1.2	3.1	0.467
YOLOv5n+4×Bottle-PFC	1.1	3.0	0.450

多通道融合金字塔结构，分别使用传统的 FPN 与所提出的 MFPN 进行多尺度特征融合。实验表明，MFPN 相比于传统的 FPN，在减少参数量的同时实现了性能的明显提升，结果如表 4 所示。

表 4　多通道融合金字塔结构对精度的影响

Table 4　The influence of multi-channel feature pyramid module on precision

模型	参数量/($\times10^6$)	浮点数/GFLOPs	mAP
YOLO-PT-N+FPN	1.54	3.8	0.445
YOLO-PT-N+MFPN	1.1	3.4	0.450
YOLO-PT-S+FPN	7.0	15.2	0.497
YOLO-PT-S+MFPN	4.3	15.8	0.530
YOLO-PT-M + FPN	16.6	37.4	0.520
YOLO-PT-M+MFPN	10.3	49.0	0.541

3.5　推理性能分析

本文以 YOLOv5 为基准模型，从真实的非结构场景中截取了 200 帧图像数据进行测试，通过计算每帧的平均推理速度进行比较。虽然多头自注意力机制需要大量的浮点运算来构建全局特征信息，但 YOLO-PT 仅使用了 1/4 的通道进行 Transformer 全局编码，极大的降低了计算量，内存交换频率也更低，因此与 YOLOv5 相比，YOLO-PT 的网络更加高效。但考虑到其他与推理相关的因素，例如存储器访问、并行度和平台特性等，参数量与浮点数并不能充分说明网络效率一定高。因此进行计算复杂度与推理性能实验直观展示推理效率，结果如表 5 所示。

表 5　计算复杂度与推理性能

Table 5　Computational complexity and inference performance

模型	参数量/($\times10^6$)	浮点数/GFLOPs	时间/ms
YOLO-PT-N	1.1	3.4	2.1
YOLO-PT-S	4.3	10.7	3.9
YOLO-PT-M	10.3	26.2	7.6
YOLOv5n	1.8	4.2	5.8
YOLOv5s	7.0	15.8	6.9
YOLOv5m	21.2	49.0	10.8

3.6　可视化实验与分析

本文在真实环境中采集的数据上进行测试，并给出了一系列可视化图，如图 10 所示，相比于其他 YOLO 模型，YOLO-PT 在非结构场景的目标检测上预测的更加准确，主要原因如下：1)与其他常规的结构化场景相比，非结构场景检测对象形状更加不规则，且分布位置不均匀，注意力机制可以更好提取全局特征；2)非结构场景中物体的尺度变化更大，往往需要同时处理多个尺度的同一物体，且物体遮挡问题更加严重，导致部分物体特征丢失，所提出的 MFPN 能够从底层获取更多信息，并在非相邻通道渐进式融合，能够更好地融合不同尺度、不同细节的特征，提升检测精度。

在一些极具挑战性(如光线昏暗、目标被掩埋)的场景，YOLO-PT 也表现出极强的泛化能力。如图 11 所示，YOLOv5 无法识别被掩埋的管道与黑暗中的石块，而 YOLO-PT 能够准确识别被水泥遮盖 80%以上的地下管道，并且在低光环境、黑暗环境下均体现了极强的检测能力。

最后使用 Grad-CAM[20]生成热力图来评估图像哪个部分吸引检测器的注意力以及各区域对检测分类的贡献程度，如图 12 所示。由图 12 可得，对于多段管道，YOLOv5 虽然聚焦到了大致的位置，但包含了大量的冗余特征，无法准确分辨横竖管道的区域；YOLO-PT 则明

图 10 非结构场景下的目标检测结果

Fig. 10 Object detection results in unstructured scenes

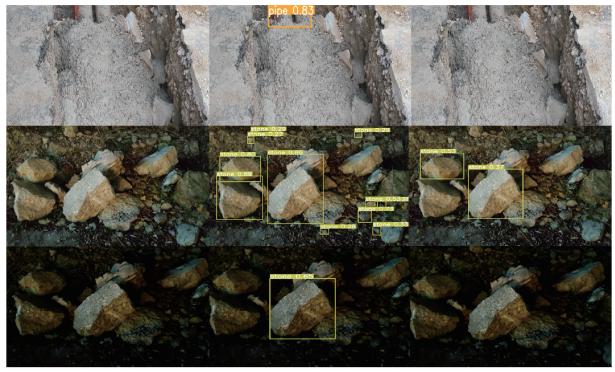

图 11 遮盖与低光条件下的检测结果

Fig. 11 Detection results under shading and low light conditions

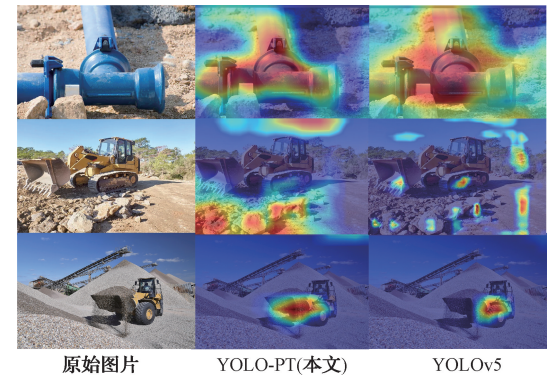

图 12 Grad-CAM 类激活可视化图

Fig. 12 Grad-CAM class-discriminative localization visualization

显表现的更好，准确聚焦了横竖两段管道的特征。此外，面对较为明显的检测对象（如挖掘机），YOLO-PT 与 YOLOv5 均能实现较好的效果，但面对尺度变化较大、背景高度相似的对象（石块），YOLOv5 无法聚焦到密集且相似的石块，YOLO-PT 则表现出较强的类响应能力与多尺度特征表示能力，能够准确聚焦到多尺度和与背景高度相似的石块。

4 结 论

本文针对目前对于非结构场景目标检测数据集缺乏、模型检测效果不佳等问题，构建了非结构场景目标检测数据集，同时研究不同维度图像特征的信息，从减少相似特征计算的角度出发提出了分部特征计算模块，配合 Transformer 自注意力机制与点卷积实现提高特征提取精度的同时降低计算复杂度，并设计了多通道金字塔结构实现对复杂不规则、尺度变化较大特征的准确融合。随后，本文提出了一种面向非结构场景的高性能实时检测算法 YOLO-PT，具有极少的参数量与计算复杂度，在保持高精度的同时实现了极致的轻量化。最后与其他检测器在检测精度、参数量、浮点数、推理速度等方面进行实验对比，结果表明 YOLO-PT 在各方面均优于工业主流的检测器。

考虑到非结构场景数据难以获取、标注困难等问题，在未来的工作中，将聚焦于仅使用小部分标注数据的弱监督或无监督目标检测研究，继续深入解决目标检测在非结构化场景应用泛用性不足的问题。

参 考 文 献

[1] 张顺，龚怡宏，王进军. 深度卷积神经网络的发展及其在计算机视觉领域的应用[J]. 计算机学报，2019，42(3)：453-482.
ZHANG SH，GONG Y H，WANG J J. The development of deep convolution neural network and lts applications on computer vision [J]. Chinese Journal of Computers，2019，42(3)：453-482.

[2] 周晓彦，王珂，李凌燕. 基于深度学习的目标检测算法综述[J]. 电子测量技术，2017，40(11)：89-93.

ZHOU X Y, WANG K, LI L Y. Review of object detection based on deep learning [J]. Electronic Measurement Technology,2017,40(11):89-93.

[3] 侯学良,单腾飞,薛靖国.深度学习的目标检测典型算法及其应用现状分析[J].国外电子测量技术,2022,41(6):165-174.
HOU X L, SHAN T F, XUE J G. Analysis of typical target detection algorithm based on deep learningand its application status [J]. Foreign Electronic Measurement Technology, 2022, 41(6): 165-174.

[4] 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(6):1137-1149.

[5] LIU W, DRAGOMIR A, DUMITRU E, et al. SSD: Single shot multibox detector [C]. European Conference on Computer Vision (ECCV), 2016: 21-37.

[6] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016:779-788.

[7] WANG C Y, BOCHKOVSKIY A, LIAO H Y. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),2023: 7464-7475.

[8] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30: 6000-6010.

[9] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. ArXiv Preprint ArXiv: 2010.11929, 2020.

[10] 徐先峰,王轲,马志雄,等.基于改进YOLOv4颈部优化网络的安全帽佩戴检测方法[J].重庆大学学报,2023,46(12):43-54.
XU X F, WANG K, MA ZH X, et al. Helmet wearing detection method based on lmproved YOLOv4 neck optimized network [J]. Journal of Chongqing University, 2023,46(12):43-54.

[11] 刘洺睿,车奔,董洪波,等.露天煤矿采场无人机遥感图像小目标检测[J].煤田地质与勘探,2023,51(11):132-140.
LIU M R, CHE B, DONG H B, et al. Detection of small objects in open-pit coal mine stopes using UAV remote sensing images [J]. Coal Geology & Exploration, 2023,51(11):132-140.

[12] 黄璐,毛晓艳,杜航,等.基于深度学习网络的星表非结构化岩石目标辨识方法研究[J].空间控制技术与应用,2021,47(6):27-33.
HUANG L, MAO X Y, DU H, et al. On star catalog unstructured rock target ldentification method basedon deep learning network[J]. Aerospace Control and Application, 2021,47(6):27-33.

[13] WANG X, HAN WI, MO S C, et al. Transformer-based automated segmentation of recycling materials for semantic understanding in construction [J]. Automation in Construction, 2023, 154: 104983.

[14] HAN K, WANG Y H, TIAN Q, et al. Ghostnet: More features from cheap operations[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020: 1577-1586.

[15] HOWARD A, SANDLER M, CHU G,et al. Searching for mobilenetv3 [C]. IEEE/CVF International Conference on Computer Vision (ICCV), 2019: 1314-1324.

[16] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2016: 770-778.

[17] 徐晓光,李海.多尺度特征在YOLO算法中的应用研究[J].电子测量与仪器学报,2021,35(6):96-101.
XU X G, LI H. Application research of multi-scale features in YOLO slgorithm[J]. Journal of Electronic Measurement and Instrumentation, 2021, 35(6): 96-101.

[18] 李晖晖,周康鹏,韩太初.基于CReLU和FPN改进的SSD舰船目标检测[J].仪器仪表学报,2020,41(4):183-190.
LI H H, ZHOU K P, HAN T CH. Ship object detection based on SSD improved with CReLU and FPN[J]. Chinese Journal of Scientific Instrument,

2020,41(4):183-190.

[19] XIE J, PANG Y W, NIE J, et al. Latent feature pyramid network for object detection[J]. IEEE Transactions on Multimedia, 2023, 25:2153-2163.

[20] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization[C]. IEEE International Conference on Computer Vision (ICCV), 2017:618-626.

作 者 简 介

金友祺,硕士研究生,主要研究方向为自动驾驶环境感知。
E-mail: gs.yqjin21@gzu.edu.cn

赵津(通信作者),教授,博士生导师,主要研究方向为领域为智能网联汽车。
E-mail: zhaoj@gzu.edu.cn