文章编号:1671-4598(2023)03-0015-07 DOI:10.16526/j.cnki.11—4762/tp.2023.03.003 中图分类号:TP391 文献标识码:A

基于 Swin Transformer 的 YOLOv5 安全帽佩戴检测方法

郑楚伟，林 辉
(韶关学院 智能工程学院，广东 韶关 512005)

摘要：针对目前施工现场的安全帽检测方法存在遮挡目标检测难度大、误检漏检率高的问题，提出一种改进 YOLOv5 的安全帽检测方法；首先，使用 K-means++聚类算法重新设计匹配安全帽数据集的先验锚框尺寸；其次，使用 Swin Transformer 作为 YOLOv5 的骨干网络来提取特征，基于可移位窗口的 Multi-head 自注意力机制能建模不同空间位置特征之间的依赖关系，有效地捕获全局上下文信息，具有更好的特征提取能力；再次，提出 C3-Ghost 模块，基于 Ghost Bottleneck 对 YOLOv5 的 C3 模块进行改进，旨在通过低成本的操作生成更多有价值的冗余特征图，有效减少模型参数和计算复杂度；最后，基于双向特征金字塔网络跨尺度特征融合的结构优势提出新型跨尺度特征融合模块，更好地适应不同尺度的目标检测任务；实验结果表明，与原始 YOLOv5 相比，改进的 YOLOv5 在安全帽检测任务上的 mAP@.5:.95 指标提升了 2.3%，检测速度达到每秒 35.2 帧，满足复杂施工场景下安全帽佩戴检测的准确率和实时性要求。

关键词：安全帽佩戴检测；YOLOv5；Swin Transformer；Ghost；新型跨尺度特征融合；K-means++

YOLOv5 Helmet Wearing Detection Method Based on Swin Transformer

ZHENG Chuwei, LIN Hui
(College of Intelligent Engineering, Shaoguan University, Shaoguan 512005, China)

Abstract: Aiming at the problems of difficult detection of occluded objects, high false detection and missed detection rate in current helmet detection methods on construction sites, an improved YOLOv5 helmet detection method is proposed in this paper. Firstly, a K-means++ clustering algorithm is used to redesign the prior anchor box size to match the helmet dataset. Secondly, Swin Transformer is used as the backbone network of YOLOv5 to extract features. The multi-head self-attention mechanism based on shiftable windows can model the dependencies between different spatial location features, effectively capture the global context information, and have better the feature extraction capability. Thirdly, a C3-Ghost module is proposed to improve the C3 module of YOLOv5 based on Ghost Bottleneck, and generate more valuable redundant feature maps through low-cost operations, which effectively reduces model parameters and computational complexity. Fourthly, a new feature fusion module is proposed based on the structural advantages of cross-scale feature fusion in bidirectional feature pyramid network, which can better adapt to the target detection tasks of different scales. The experimental results show that compared with the original YOLOv5, the mAP@.5:.95 index of the improved YOLOv5 on the helmet detection task is improved by 2.3%, and the detection speed reaches 35.2 frames per second, which meets the accuracy and real-time requirement of the helmet wearing detection in complex construction scenarios.

Keywords: helmet wearing detection; YOLOv5; Swin Transformer; Ghost; new cross-scale feature fusion; K-means++

0 引言

在工地的作业现场，正确佩戴安全帽能有效地防止施工人员在生产过程中遭受坠落物体对头部的伤害。然而在实际生产活动中，尽管每个施工项目都明文要求人员一定要正确佩戴安全帽，但仍杜绝不了个别工人缺少自我安全防范意识，在施工现场不戴或者不规范佩戴安全帽的现象[1]。目前施工现场对安全帽佩戴情况的监控大多仍依赖人工监视[2]，这种方式存在成本高、耗时长、容易出错的不足。采用视频自动监控方法有利于实时监控施工现场人员的安全帽佩戴情况，对安全生产环节中的安全隐患进行实时评估。

目前已有学者对安全帽检测方法进行研究。刘晓慧等[3]采用肤色检测的方法定位人脸，再利用支持向量机(SVM)实现安全帽的识别；刘云波等[4]通过统计工人图像的上三分之一区域出现频率最高的像素点色度值并与安全帽颜色相匹配，以此来判断安全帽佩戴情况。但传统的目标检测需要通过手工设计特征，存在准确率低、不具备鲁

收稿日期:2022-07-09； 修回日期:2022-08-31。
基金项目:广东大学生科技创新培育专项资金资助项目(pdjh2022b0470)。
作者简介:郑楚伟(1999-),男,广东汕头人,大学本科,主要从事深度学习和图像处理方向的研究。
通讯作者:林 辉(1984-),男,广东乳源人,博士研究生,副教授,主要从事机器视觉理论及应用领域方向的研究。
引用格式:郑楚伟,林 辉.基于 Swin Transformer 的 YOLOv5 安全帽佩戴检测方法[J].计算机测量与控制,2023,31(3):15-21.

棒性等问题。

随着深度学习的发展，国内外已有大量学者使用基于卷积神经网络算法对安全帽检测进行了一系列研究。其中有先提取候选框再回归定位的两阶段算法，如 R-CNN（regions with convolutional neural network features）[5]、Fast R-CNN（fast region-based convolutional neural network）[6]和 Faster R-CNN（faster region-based convolutional neural network）[7]等网络和直接进行一阶段目标检测的 SSD（single shot MultiBox detector）[8]和 YOLO（you only look once）[9]系列算法。张玉涛等[10]使用轻量化的网络设计减小模型的计算量，使得模型达到 137 帧每秒的运行速度，但是总体的检测错误率达到 7.9%。张明媛等[11]使用 Faster RCNN 网络检测施工人员的安全帽佩戴情况，但未考虑检测效率的问题，无法实现实时检测。杨莉琼等[12]提出了一种基于机器学习的安全帽检测方法，使得每帧图像的检测时间小于 50 ms，满足时效性需求，但在检测图像中的小目标时准确率较低。孙国栋等[13]提出了一种通过融合自注意力机制来改进 Faster RCNN 的目标检测算法，具有较好的检测效果，但是模型的参数量和计算复杂度高。张锦等[14]在 YOLOv5 特征提取网络中引入多光谱通道注意力模块，使网络能够自主学习每个通道的权重，提升了模型的平均准确率，但网络模型参数量以及检测速率有待提升。

本文提出一种改进 YOLOv5 的安全帽检测方法，将 Swin Transformer 作为 YOLOv5 的骨干网络，使得模型能够更好地提取图像特征。同时，使用 K-means++聚类算法重新设计匹配安全帽数据集的先验锚框尺寸，基于 Ghost Bottleneck 对 YOLOv5 的 C3 模块进行改进从而减少模型参数，提出新型跨尺度特征融合模块，更好地适应不同尺度的目标检测任务。实验结果表明，改进的 YOLOv5 在安全帽检测任务上的 mAP@.5:.95 指标提升了 2.3%，检测速度达到每秒 35.2 帧，有效解决施工现场的安全帽检测方法存在遮挡目标检测难度大、误检漏检率高的问题，满足复杂施工场景下安全帽佩戴检测的准确率和实时性要求。

1 系统结构及原理

本文通过改进 YOLOv5 网络结构以解决安全帽检测过程存在遮挡目标检测难度大、误检漏检率高的问题。改进 YOLOv5 网络结构如图 1 所示，分为数据输入、骨干网络、颈部及预测部分。

数据输入部分使用自适应图像填充、Mosaic 数据增强来对数据进行处理，提升小目标的检测的精度。骨干网络部分使用 Swin Transformer 作为 YOLOv5 的主干特征提取网络。颈部部分借鉴双向特征金字塔网络跨尺度特征融合的结构优势，在 FPN + PANet 结构进行特征融合的基础上，增添了两条特征融合路线，用较少的成本使得同层级上的特征图能够共享彼此的语义信息；另一方面，为了减少模型参数，提出了基于 Ghost Bottleneck 对 YOLOv5 的 C3 模块进行改进的 C3-Ghost 模块。预测部分，如图 1 最右侧所示，从上到下依次是经过特征融合后得到原图的 4、8、16、32 倍下采样的特征图，在此基础上使用二元交叉熵损失函数计算置信度预测损失和分类预测损失，使用广义交并比（GIoU，generalized intersection over union）计算边界框的损失，同时采用非极大值抑制对多个目标锚框进行筛选来提高对目标识别的准确度。

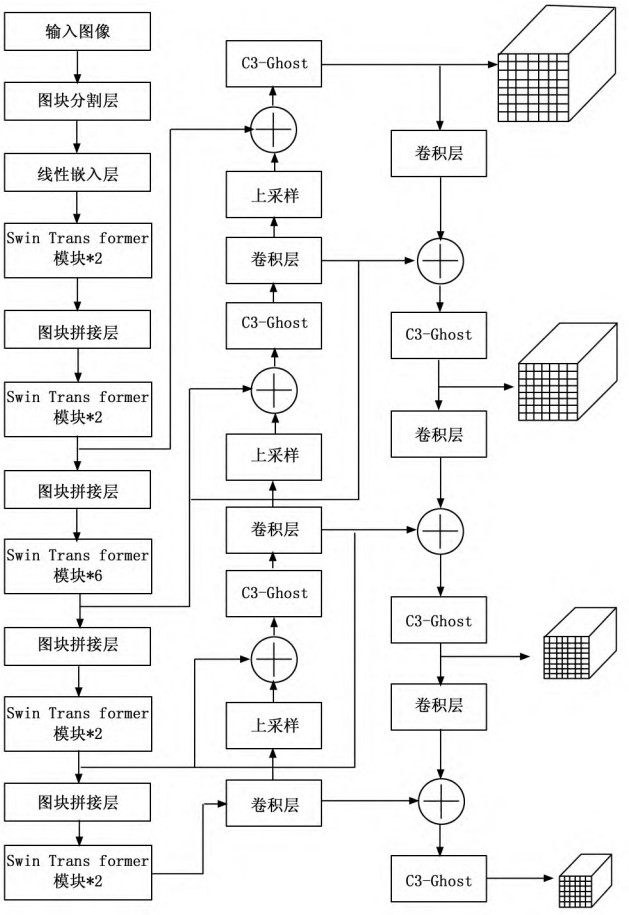

图 1 改进 YOLOv5 网络结构

2 改进 YOLOv5 网络模型

2.1 Swin Transformer 模块

Swin Transformer 是基于具有全局信息建模能力的 Transformer 来构建分层特征图，同时借鉴 locality 思想将自注意力计算限制在无重叠的窗口区域内并允许移动窗口进行特征交互，对图像大小具有线性计算复杂度[15]，可将其作为 YOLOv5 的骨干网络来更有效的提取图像特征。

Swin transformer 网络架构如图 2（a）所示。首先将输入的 [H，W，3] 图像传入图块分割层（patch partition），将每 4×4 相邻的像素分块为一个 Patch，并沿着通道方向展开，使得图像维度变成了 [$H/4$，$W/4$，48]，然后在通过线性嵌入层（linear embeding）对每个像素的通道数据做线性变换，使图像维度变成 [$H/4$，$W/4$，C]，同时将每个样本在特征维度进行归一化处理。

Swin transformer 构造图像的层次特征图是通过在每个

阶段之间使用图块拼接层（patch merging）对图像进行下采样，图块拼接层的实现类似于 YOLOv5 的 Focus 结构对图片进行切片操作，将间隔为 2 的相邻像素划分为一个个 Patch 后再进行通道拼接（concat）操作，使得特征图尺度减半。然后在通过一个标准化层（LN，layer normalization）进行归一化操作，最后通过一个全连接层将特征图的通道数线性变换为原来的一半。原特征图经过图块拼接层后，宽和高会减半，通道数翻倍，随着网络层次的加深，节点的感受野也在不断扩大。

构造出不同尺度的特征图后，将使用 Swin Transformer 模块提取图像的特征。Swin Transformer 通过将传统 Transformer 模块中的标准 Multi-head 自注意力模块（MSA，multi-head self attention）替换为由窗口 Multi-head 自注意力层（W-MSA，window multi-head Self Attention）和滑动窗口 Multi-head 自注意力层（SW-MSA，shifted window based multi-head self attention）交替组成的基于窗口的 Multi-head 自注意力模块，即将输入图像均匀的划分为互不重叠的窗口，将注意力计算限制在每个独立的窗口内。本文所设计的骨干网络中第一、二、四、五个阶段均包括两个 Swin transformer 模块，第三个阶段包括六个 Swin transformer 模块，并将第二、三、四、五个阶段输出的特征图进行融合，得到 4 个不同网格尺寸的输出特征图。Swin Transformer 模块的结构如图 2（b）所示，包括 4 个 LN 层、一个 W-MSA 层、一个 SW-MSA 层、两个多层感知器（MLP，multilayer perceptron）、4 个路径随机失活层（DropPath）和 4 个残差连接层。输入到该模块的特征 Z^{l-1} 先经过 LN 进行层归一化后，利用 W-MSA 层提取特征，再将残差操作得到 $\hat{Z}^{l}$ 特征使用 LN 层归一化，然后是一个中间带有 GELU 非线性激活函数的 2 层对通道维度进行线性变换的 MLP，再使用残差连接得到输出特征 Z^{l}，再将其输入到包含 SW-MSA 层的类似结构中。路径随机失活层的作用是将 Swin Transformer 模块的多分支路径随机失活的正则化策略，以此提高模型的泛化能力，防止过拟合[16]。同时采用残差连接结构，目的是解决神经网络中的退化问题[17]。

Swin Transformer 模块中的 W-MSA 层是使用从左上角像素开始的常规窗口划分策略。如图 2（c）所示，其包括窗口分割（window partition）模块、窗口重组（window reverse）模块和 MSA 模块。其中窗口分割模块用于将输入的特征图分割为多个 M×M 相邻像素的互不重叠的独立窗口；窗口重组模块用于对每个独立窗口的 Multi-head 自注意力特征进行还原拼接为完整的 Multi-head 自注意力特征图；MSA 模块用于对每个独立窗口分别进行 Multi-head 的缩放点积注意力计算，步骤包括：对每个独立窗口的图块向量在通道维度进行线性变换，使通道数增加两倍，同时在特征维度上分割为 h 个子空间（h 为注意力 head 的个数）；通过 h 个不同的参数矩阵 $\boldsymbol{W}^{Q}$、$\boldsymbol{W}^{K}$、$\boldsymbol{W}^{V}$ 分别在 h 个子空间中对每个像素的查询 Q（Query）、键 K（Key）和权重 V（Value）进行线性变换，并进行缩放点积注意力计算；将 h 个计算结果通过可学习的权重矩阵 W^{O} 进行拼接融合，以联合来自不同子空间中学习到的特征信息，得到 Multi-head 自注意力特征[18]。其中，第 i 个注意力头的缩放点积注意力计算结果 $head_i$ 的表达式如式（1）所示：

$$head_i = Attention(\boldsymbol{QW}_i^{Q}, \boldsymbol{KW}_i^{K}, \boldsymbol{VW}_i^{V}) \tag{1}$$

式（1）中，W_i^{Q}、W_i^{K}、W_i^{V} 分别为第 i 个参数矩阵 W^{Q}、W^{K}、W^{V}，$Attention$（）为归一化的缩放点积模型，其表达式如式（2）所示：

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = SoftMax(\boldsymbol{QK}^{T}/\sqrt{d}+\boldsymbol{B})\boldsymbol{V} \tag{2}$$

式（2）中，$\boldsymbol{Q}$，$\boldsymbol{K}$，$\boldsymbol{V}\in\mathbb{R}^{M^2\times d}$，$\boldsymbol{QK}^{T}$ 是不同特征进行信息交互的过程，采用点积来计算不同特征之间的相似度；除以$\sqrt{d}$进行缩放操作能保证梯度的稳定性；同时，在每一个 $head_i$ 中添加可学习的相对位置编码 $\boldsymbol{B}\in\mathbb{R}^{M^2\times M^2}$。Multi-head 自注意力特征的拼接融合表达式如式（3）所示。

$$MultiHead(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = Concat(head_1,\cdots,head_h)\boldsymbol{W}^{o} \tag{3}$$

为了实现不重叠窗口之间的信息传递，可使用 SW-MSA 的方式重新计算窗口偏移之后的自注意力，让模型能够学习到跨窗口的信息。如图 2（c）所示，SW-MSA 中使用了循环移位（Cyclic Shift）的方法，即通过将特征图最上面的 $M/2$ 行（M 是每个划分窗口的尺寸）像素移动到最下面，再将最左边的 $M/2$ 列像素移动到最右边，再使用 W-MSA 划分窗口的方法将重组的特征图划分为不重叠的窗口；然后通过掩码机制，将每个窗口内来自不相邻区域的像素点之间的权重系数置为 0，隔离原特征图中不相邻区域的像素点之间无效的信息交流，以此将自注意力计算限制在每个子窗口内；最后再通过反向循环移位（reverse cyclic shift）操作还原特征图的相对位置。

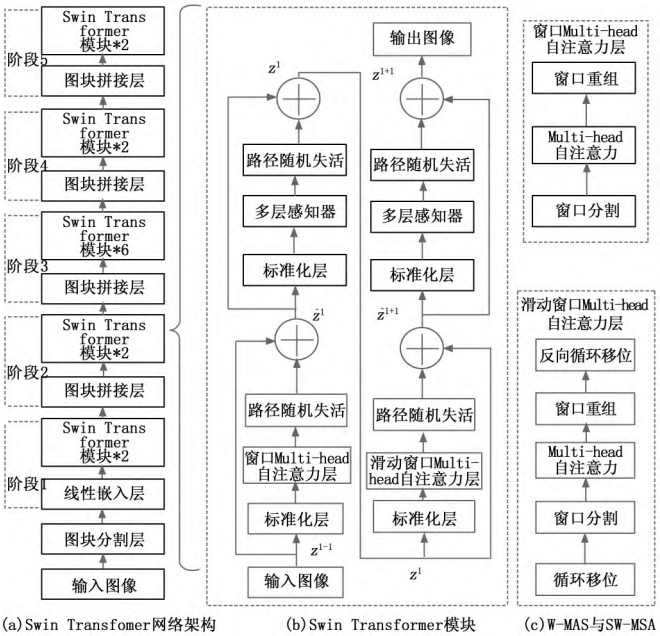

图 2　Swin Transformer 模块结构图

2.2　C3-Ghost 模块

在深层神经网络的特征映射中，丰富甚至冗余的信息

往往保证了对输入数据的全面理解。部分冗余的特征可能是深层神经网络有效的一个重要原因。Ghost 模块能以一种高效的方式获得这些冗余的特征[19]。本文提出 C3-Ghost 模块，如图 3 所示，基于 Ghost Bottleneck 对 YOLOv5 的 C3 模块进行改进，旨在通过低成本的操作生成更多有价值的冗余特征图，有效的提升网络性能。

Ghost 模块将普通卷积层拆分为两部分，首先使用若干个 1×1 的卷积核进行逐点卷积，生成输入特征的固有特征图，然后用逐层卷积进行一系列线性变换来高效地生成冗余特征图，再将冗余特征图和固有特征图进行拼接，得到和普通卷积结果具有相似作用的特征图，如图 3 所示。与普通卷积操作相比，Ghost 模块在不改变输出特征尺寸和维度的情况下能有效减少模型参数和计算复杂度。

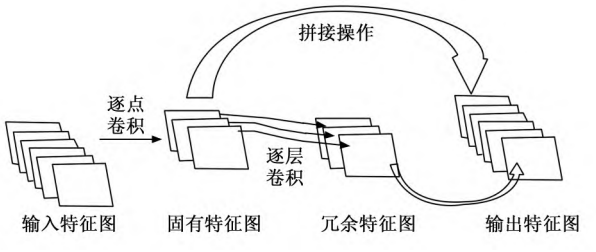

图 3 Ghost 模块

本文设计的 C3-Ghost 模块使用 Ghost Bottlenecks 结构来替代原始 C3 模块的 Bottlenecks 结构，Ghost Bottlenecks 的本质是用 Ghost 模块代替 Bottlenecks 结构里面的普通卷积。如图 4 所示，第一层 Ghost 模块用于增加通道数量，从而增加特征维度，第二层 Ghost 模块用于减少特征维度使其适配残差连接，将输入与输出相加。引入批量归一化（BN，batch normalization）尽可能保证每一层网络的输入具有相同的分布[20]，引入具有稀疏性的 ReLu 激活函数能避免反向传播的梯度消失现象，第二层 Ghost 模块后没有使用 ReLu 激活函数是因为 ReLU 负半轴存在的硬饱和置 0 会使其输出数据分布不为零均值而导致神经元失活，从而降低网络的性能[21]。

2.3 新型跨尺度特征融合模块

在目标检测任务中，融合不同尺度的特征是提高性能的一个重要手段。目前已有的特征融合网络有 PANet[22]、FPN[23]、BiFPN[24]等。YOLOv5 使用 FPN＋PANet 结构将高层特征丰富的语义信息和低层特征丰富的细节信息相互融合，如图 5（a）所示。

考虑到加权双向特征金字塔网络（BiFPN，bi-directional feature pyramid network）[24]的结构优势，本文提出将 BiFPN 的思想应用到 YOLOv5 的多尺度特征融合部分，通过添加横向跳跃连接，即在处于同一层级的原始输入和输出节点之间添加一条新的融合路线，如图 5（b）所示，在原始 YOLOv5 特征跨尺度融合模块基础上添加沿着两条虚线的特征融合路线，用较少的成本使得同层级上的特征图能够共享彼此的语义信息，加强特征融合以提高模型精度。

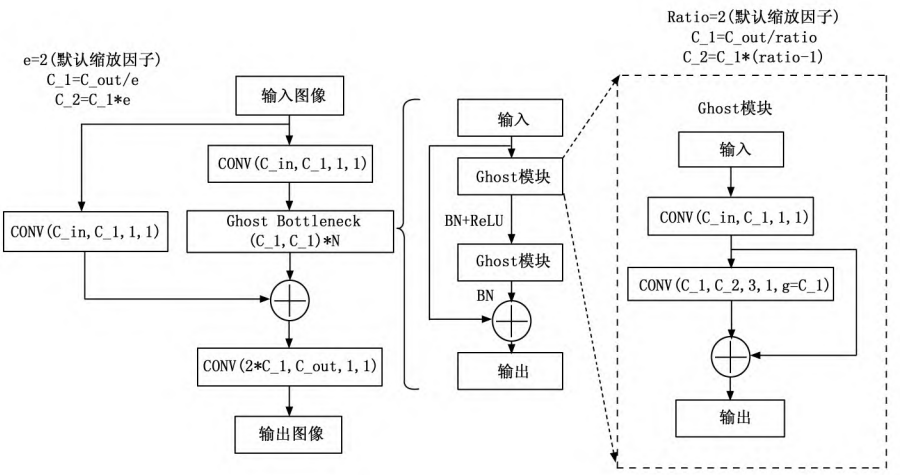

图 4 C3-Ghost 模块

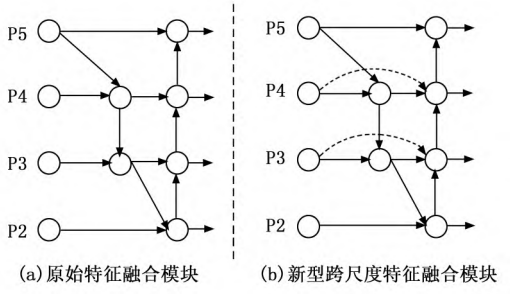

图 5 不同特征融合模块的对比

2.4 改进先验锚框尺寸

在模型训练中，先验锚框尺寸越接近真实边界框，模型将会越容易收敛，其预测边界框也会更加接近真实边界框。原始 YOLOv5 模型中预设了匹配 COCO 数据集的锚框，但本文使用的安全帽数据集的边界框具有类型单一、边界框尺寸比较集中的特点，预设的锚点不能合理地直接应用。因此本文提出使用 K-means＋＋对安全帽数据集的边界框进行聚类分析，找到 12 个聚类中心的边界框作为先验锚框参数的值，并将其匹配到相应的特征检测层，使模型能够更快收敛。

由于卷积神经网络具有平移等变性[25]，因而只需要通过 K-means＋＋对边界框的宽高进行聚类，不用考虑边界框位置的影响。首先通过轮盘赌算法依据概率大小选择初始聚类中心，然后依次计算每个边界框与初始聚类中心的距离，按照距离大小对边界框进行划分聚类，再更新聚类中心，直到在连续迭代中聚类中心位置稳定。在距离度量上，采用式（4）代替标准 K-means＋＋中使用的欧氏距离：

$$Loss = min\sum_{i=1}^{n}\sum_{j=1}^{k}\left[1-\frac{Box_i \cap Center_j}{Box_i \cup Center_j}\right] \tag{4}$$

其中：Box_i 为第 i 个真实边界框的面积，$Center_j$ 为第 j 个聚类中心的面积，n 为真实边界框总数，k 为聚类中心个数，本文 YOLOv5 有 4 个不同尺度的特征检测层，每层分配 3 个先验锚框，故 $k=12$。

由于安全帽边界框尺寸较为集中，为了更好的发挥 YOLOv5 算法的多尺度目标检测能力，将对 K-means++ 聚类所得的先验锚框尺寸进行线性变换操作[26]，如式（5）所示。

$$\begin{cases} x'_{\min} = \alpha \cdot x_{\min} \\ x'_{\max} = \beta \cdot x_{\max} \\ x'_i = \dfrac{(x_i - x_{\min})}{(x_{\max} - x_{\min})} \cdot (x'_{\max} - x'_{\min}) + x'_{\min} \\ y'_i = y_i \cdot \dfrac{x'_i}{x_i} \end{cases} \tag{5}$$

以 α、β 为缩放因子，将 (x_i, y_i) 等比例线性变换为 (x'_i, y'_i)，其中，$\alpha=0.8$，$\beta=1.5$，表示将锚框的宽的最小值变为原先的 0.8 倍，最大值变为 1.5 倍，并且保持宽高比例不变。

本文使用上述算法所设计的先验锚框尺寸如表 1 所示。最佳可能召回率（BPR，best possible recalls）是衡量先验锚框和真实边界框匹配程度的指标，定义为一个检测器最多能召回的真实边界框数量与真实边界框总数之比[27]。在训练期间，如果一个真实边界框被分配给至少一个先验锚框，则认为该真实边界框被召回。BRP 的最大值为 1 并且越接近越好。在本文改进的先验锚框下计算所得的 BPR 值为 0.999，表明本文所设的先验锚框和安全帽数据集的真实边界框具有很好的匹配程度。

表 1　改进先验锚框尺寸　　像素

特征图尺寸	先验锚框尺寸		
小	(5,7)	(6,8)	(9,11)
中等	(12,14)	(15,18)	(20,24)
大	(26,31)	(35,41)	(49,56)
最大	(69,78)	(105,122)	(190,225)

3　实验结果与分析

3.1　实验环境

本实验使用 Pytorch 1.9.1 深度学习框架、python3.6 环境，操作系统为 Window10，CPU 型号为英特尔 Core i9-9820X、128 GB 内存、24 GB 显存的 NVIDIA TITAN RTX 显卡的设备上完成训练，NVIDIA 驱动版本为 456.71，并行计算架构 CUDA 版本为 10.0.130，深度学习加速库 CUDNN 版本为 7.6.5。

在训练过程中，设置最大迭代次数为 200 个 epochs，采用 SGD 优化器，动量因子为 0.937，权重衰减系数为 0.000 5，初始学习率为 0.01，初始阶段使用 warmup 预热学习率，前 3 个 epochs 采用一维线性插值调整学习率，随后使用余弦退火算法更新学习率。

3.2　数据集

为了验证本文改进网络的优越性，实验使用了开源安全帽佩戴检测数据集（SHWD，safety helmet wearing detect dataset）进行验证。SHWD 提供 7 581 张图像，其中包括 9 044 个佩戴安全帽的正样本人物头像和 111 514 个未佩戴安全帽的负样本人物头像。本文以训练集：测试集＝9：1的比例划分数据集，其中训练集有 2 916 张正样本图像和 3 905 张负样本图像，测试集有 325 张正样本图像和 435 张负样本图像。

3.3　评价指标

在目标检测领域常采用精确度（P，precision）、召回率（R，recall）、平均精确度（AP，average precision）、平均精确度均值（mAP，mean Average Precision）、每秒传输帧数（FPS，frames per second）指标来评估模型性能。计算公式如式（6）～（9）所示。

$$P = \frac{TP}{TP + FP} \tag{6}$$

$$R = \frac{TP}{TP + FN} \tag{7}$$

$$AP = \int_0^1 P(R)d(R) \tag{8}$$

$$mAP = \frac{\sum_{i=1}^{n} AP(i)}{n} \tag{9}$$

式（9）中，n 表示类别总数。AP 用于衡量模型对某一类别的平均精度，mAP 是所有类别的平均 AP 值。其中，真正例（TP，true positive）、假正例（FP，false positive）、真反例（TN，true negative）和假反例（FN，false negative）的混淆矩阵如表 2 所示。

表 2　各指标的混淆矩阵

实际情况	检测结果	
	正样本	负样本
正样本	真正例	假反例
负样本	假正例	真反例

3.4　实验结果与分析

本实验使用具有 4 个检测尺度的 YOLOv5 作为基准模型，其是在标准模型的基础上将更低层级的特征图引入到特征融合网络中，使得特征融合网络能更有效地捕获更丰富的细节信息，从而提高小目标检测精度。实验结果如表 3 所示，其中 mAP@.5 是 IoU 阈值为 0.5 的情况下每一类别的 AP 值的平均；mAP@.5：.95 表示 IoU 阈值从 0.5 开始、以 0.05 的步长增长到 0.95 所对应的平均 mAP。表 3 中，“C3-Ghost”是指使用本文所设计的 C3-Ghost 模块替代基准 YOLOv5 特征融合网络中的 C3 模块；“新型跨尺度特征融合”是指使用本文提出的新型跨尺度特征融合模块替代基准 YOLOv5 的特征融合网络；“Swin Transformer”是指使用 Swin Transformer 替代基准 YOLOv5 的骨干网络的模型。

由表 3 可知，YOLOv5 基准模型的参数量为 7.17×10^6，YOLOv5＋C3-Ghost 模型参数量为 6.14×10^6，在保持 mAP@.5：.95 值几乎不变的情况下，YOLOv5＋C3-Ghost 的参数量相比于基准 YOLOv5 减少了 14.4%，证明本文所设计的 C3-Ghost 模块能有效减少模型参数。YOLOv5＋新型跨尺度特征融合在 mAP@.5：.95 指标上相比基准 YOLOv5 提高了 0.6 个百分点。YOLOv5＋Swin Transformer 相比基准 YOLOv5 模型，在 mAP@.5：.95 指标上提升了 2.1 个百分点。YOLOv5＋Swin Transformer＋C3-Ghost 在 mAP@.5：.95 指标上相比 YOLOv5 基准模型提升了 1.9 个百分点。表 3 中“本文改进算法”是本文提出的改进 YOLOv5 网络模型，其具备 Swin Transformer 强大的特征提取能力外，既有 C3-Ghost 模块带来的轻便性，又有新型跨尺度特征融合加强特征融合带来的高准确率，从表 3 中可以看出本文所提出的改进 YOLOv5 网络模型相比基准 YOLOv5 网络模型，在 mAP@.5：.95 指标上提升了 2.3 个百分点，较基准模型具有显著提升。

为了验证本文改进算法的有效性，本文使用以 resnet50 为骨干网络的两阶段目标检测网络 Faster RCNN 和以 vgg 为骨干网络的单阶段目标检测网络 SSD 进行对比。从表 3 可以看出，本文所使用的改进 YOLOv5 模型在安全帽检测任务上的 mAP@.5：.95 值 55.2%远高于 Faster RCNN 的 mAP@.5：.95 值 37.2%以及 SSD 的 mAP@.5：.95 值 37.2%。

值得说明的是，YOLOv5 单独融合 Swin Transformer 模块和融合 Swin Transformer＋C3-Ghost 模块的算法相比于本文改进算法，在 mAP@.5：.95 指标上分别减少了 0.2%和 0.4%。本文认为，Swin Transformer 模块作为骨干网络能有效提取图像初步抽象的、细节性的特征，而本文设计的新型跨尺度特征融合模块能更有效的融合来自 Swin Transformer 浅层特征丰富的纹理信息和深层特征丰富的语义信息，使得网络更深层的颈部以及预测模块的特征图具有更丰富的高级语义信息，从而提高目标检测精度。

FPS 指标用于评估模型每秒处理的图像帧数。本文在施工现场监控中截取一段视频进行检测，如表 3 所示，可以看出，本文改进算法检测速度达到每秒 35.2 帧，能够达到实时检测的效果。由于 Swin Transformer 网络较高的计算复杂度而导致其 FPS 低于基准 YOLOv5，但仍高于 Faster RCNN 和 SSD 每秒能处理的图像帧数。

为了更直观地展示出改进 YOLOv5 模型的优势，本文使用改进前后的 YOLOv5 模型进行检测，如图 6 所示，其中正确佩戴安全帽的施工人员上方显示“hat”标签，未佩戴安全帽的施工人员上方显示“no-hat”标签。图 6（a）中，原始 YOLOv5 模型漏检了一个被施工设备遮挡的未佩戴安全帽的施工人员，而改进后的 YOLOv5 模型则可以正确检测出来，如图 6（b）所示。图 6（c）中，原始 YOLOv5 模型错误地将控制施工设备的圆形控制器判断为未佩戴安全帽的施工人员，而改进后的 YOLOv5 模型则可以得到正确的结果，如图 6（d）所示。图 6（e）是在弱光照下的检测效果，可以看出原始 YOLOv5 模型漏检了一个佩戴安全帽的施工人员，而改进后的 YOLOv5 模型则具有较好的表现，如图 6（f）所示，能正确检测出光照不充足的图像中的目标。由上述检测对比可知，本文改进后的 YOLOv5 模型能有效解决施工现场安全帽佩戴检测存在的遮挡目标检测难度大、误检漏检率高的问题，满足复杂施工

(a)遮挡目标漏检效果图 (b)漏检改进效果图

(c)目标误检效果图 (d)误检改进效果图

(e)弱光照下漏检的效果图 (f)弱光照下漏检的改进效果图

图 6 不同场景下原始模型和本文改进模型的检测结果对比

表 3 多种模型实验结果对比

模型	P/%	R/%	mAP@.5/%	mAP@.5：.95/%	参数量/M	FPS
YOLOv5	90.4	86.2	91.7	52.9	7.17	121.9
YOLOv5＋C3-Ghost	90.1	85.7	91.8	52.9	6.14	102.0
YOLOv5＋新型跨尺度特征融合	91.5	85.8	92.2	53.5	7.24	123.4
YOLOv5＋新型跨尺度特征融合＋C3-Ghost	90.0	86.1	91.8	53.0	6.24	125.0
YOLOv5＋Swin Transformer	88.9	83.4	90.0	55.0	42.80	35.4
YOLOv5＋Swin Transformer＋C3-Ghost	88.3	83.4	89.6	54.8	41.70	35.5
Faster RCNN	54.8	73.1	67.8	37.2	108.1	10.0
SSD	83.1	44.9	61.8	37.2	90.58	32.1
本文改进算法	89.3	83.0	90.0	55.2	41.90	35.2

场景下安全帽佩戴检测的准确率要求。

4　结束语

针对目前施工现场的安全帽检测方法存在遮挡目标检测难度大、误检漏检率高的问题，本文提出一种改进 YOLOv5 的安全帽佩戴检测方法。通过将 Swin Transformer 作为 YOLOv5 的骨干网络，能够有效结合 Swin Transformer 强大的特征提取能力和 YOLOv5 单阶段目标检测算法高效推理速度的优势。实验结果表明本文改进算法在安全帽检测任务上的 mAP@.5：.95 指标提升了 2.3%，每秒检测图片的帧数达到 35.2 帧，能够达到实时检测的效果，满足复杂施工场景下安全帽佩戴检测的准确率和实时性要求。下一步的工作是继续研究如何在保持精准度的情况下，减少网络模型参数量以及提升检测速率。

参考文献：

[1] 常　欣，刘鑫萌. 建筑施工人员不合理佩戴安全帽事故树分析[J]. 吉林建筑大学学报，2018，35（6）：65-69.

[2] 王忠玉. 智能视频监控下的安全帽佩戴检测系统的设计与实现[D]. 北京：北京邮电大学，2018.

[3] 刘晓慧，叶西宁. 肤色检测和 Hu 矩在安全帽识别中的应用[J]. 华东理工大学学报（自然科学版），2014，40（3）：365-370.

[4] 刘云波，黄　华. 施工现场安全帽佩戴情况监控技术研究[J]. 电子科技，2015，28（4）：69-72.

[5] GIRSHICK R，DONAHUE J，DARRELL T，et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C] // CVPR 2014：Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. 2014：580-587.

[6] GIRSHICK R. Fast R-CNN [C] // proceedings of the IEEE international conference on computer vision，015：1440-1448.

[7] REN S，HE K，GIRSHICK R，et al. Faster R-CNN：Towards Real-Time Object Detection with Region Proposal Networks [J]. Advances in neural information processing systems，2015，28：91-99.

[8] LIU W，ANGUELOV D，ERHAN D，et al. SSD：SSD：single shot multiBox detector [C] // Proceedings of the 2016 European Conference on Computer Vision，2016：21-37.

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

[10] 张玉涛，张梦凡，史学强，等. 人员安全帽佩戴轻量化检测方法研究 [J]. 安全与环境学报：2023，23（2）：474-480.

[11] 张明媛，曹志颖，赵雪峰，等. 基于深度学习的建筑工人安全帽佩戴识别研究 [J]. 安全与环境学报，2019（2）：535-541.

[12] 杨莉琼，蔡利强，古　松. 基于机器学习方法的安全帽佩戴行为检测 [J]. 中国安全生产科学技术，2019，15（10）：152-157.

[13] 孙国栋，李　超，张　航. 融合自注意力机制的安全帽佩戴检测方法 [J/OL]. 计算机工程与应用：1-7 [2021-10-12]. http://kns.cnki.net/kcms/detail/11.2127.TP.20210621.1819.010.html.

[14] 张　锦，屈佩琪，孙　程，等. 基于改进 YOLOv5 的安全帽佩戴检测方法 [J/OL]. 计算机应用：1-11 [2022-02-10]. http://kns.cnki.net/kcms/detail/51.1307.TP.20210908.1727.002.html.

[15] LIU Z，LIN Y T，CAO Y，et al. Swin Transformer：Hierarchical Vision Transformer using Shifted Windows [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision，2021：10012-10022.

[16] LARSSON G，MAIRE M，SHAKHNAROVICH G. FractalNet：Ultra-Deep Neural Networks without Residuals [C] // 2016.

[17] HE K，ZHANG X，REN S，et al. Deep Residual Learning for Image Recognition [J]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2016：770-778.

[18] VASWANI A，SHAZEER N，PARMAR N，et al. Attention Is All You Need [C] // Advances in Neural Information Processing Systems，2017：5998-6008.

[19] HAN K，WANG Y，TIAN Q，et al. GhostNet：More Features From Cheap Operations [C] // 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)，2020：1577-1586.

[20] IOFFE S，SZEGEDY C. Batch Normalization：Accelerating Deep Network Training by Reducing Internal Covariate Shift [C] // JMLR.org. JMLR.org，2015.

[21] 张　焕，张　庆，于纪言. 激活函数的发展综述及其性质分析 [J/OL]. 西华大学学报（自然科学版）：1-10 [2022-02-12]. http://kns.cnki.net/kcms/detail/51.1686.N.20210708.1617.004.html.

[22] LIU S，QI L，QIN H，et al. Path Aggregation Network for Instance Segmentation [C] // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR），2018：8759-8768.

[23] LIN T Y，DOLLAR P，GIRSHICK R，et al. Feature Pyramid Networks for Object Detection [C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)，2017：936-944.

[24] TAN M，PANG R，LE Q V. EfficientDet：Scalable and Efficient Object Detection [C] // 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR），2020：10778-10787.

[25] 李俊英. 深度卷积神经网络的旋转等变性研究 [D]. 杭州：浙江大学，2019.

[26] 王品学，张绍兵，成　苗，等. 基于可变形卷积和自适应空间特征融合的硬币表面缺陷检测算法 [J/OL]. 计算机应用：1-9 [2022-02-12]. http://kns.cnki.net/kcms/detail/51.1307.tp.20210413.1607.002.html.

[27] TIAN Z，SHEN C，CHEN H，et al. FCOS：Fully Convolutional One-Stage Object Detection [C] // 2019 IEEE/CVF International Conference on Computer Vision（ICCV），2019：9626-9635.