文章编号:1671-4598(2023)09-0102-07 DOI:10.16526/j.cnki.11—4762/tp.2023.09.016 中图分类号:TP751;TP391 文献标识码:A

基于改进 YOLOv5 的遥感图像目标检测研究

李建新[1]，陈厚权[2]，范文龙[2]
(1. 保定市不动产登记中心，河北 保定 071051；
2. 河北大学 质量技术监督学院，河北 保定 071002)

摘要：针对遥感图像中背景复杂度高、目标尺寸多样所导致的目标检测精度低的问题，提出一种基于改进 YOLOv5 的遥感图像目标检测算法；采用 ConvNeXt 网络作为主干网络，结合了 CNN 的局部特性和 Transformer 的全局特性，克服了传统 CNN 在全局上下文信息处理和长距离依赖关系挖掘上的局限性，实现了对全局信息的有效捕获；引入 SimAM 注意力机制，在不增加网络参数的情况下推断出特征图的 3D 注意力权值，提高网络的稳定性以及抗干扰能力；采用 CFP 捕获全局长距离依赖关系以及遥感图像的局部关键区域信息，以及新颖的 SIoU loss 边界框定位损失函数和 Soft-SIoU-NMS 非极大值抑制方法，进一步提升了遥感图像实时检测的效果；在 RSOD 数据集上进行的测试，结果表明本算法相比于原网络提高了 10.6%的平均精度，达到了 94.2%。

关键词：遥感图像；目标检测；YOLOv5；SimAM；CFP

Research on Object Detection in Remote Sensing Images Based on YOLOv5

LI Jianxin[1], CHEN Houquan[2], FAN Wenlong[2]
(1. Baoding Real Estate Registration Center, Baoding 071051, China;
2. College of Quality and Technical Supervision, Hebei University, Baoding 071002, China)

Abstract: For the problems of low detection accuracy in complex backgrounds and diversity of target sizes in remote sensing images, an improved YOLOv5-based algorithm for remote sensing image object detection is proposed, The ConvNeXt network is adopted as the backbone network, CNN's local features and Transformer's global features are combined to overcome the limitations of processing global contextual information and effectively capturing long-range dependencies for traditional CNNs, and achieve the effective capture of global information. The SimAM attention mechanism is introduced to deduce the 3D attention weights of the feature maps without increasing network parameters, enhancing the network stability and anti-interference ability Concurrently, centralized feature pyramid (CFP) is used to capture the global long-range dependencies , local key region information in remote sensing images, novel SIoU loss function, and Soft-SIoU-NMS non-maximum suppression metho, and further improve the real-time detection performance of sensing images . Testing on the RSOD dataset, the results show that compared to the original network, the average precision of the proposed method improves 10.6%, reaches, 94.2%.

Keywords: remote sensing images; object detection; YOLOv5; SimAM; CFP

0 引言

遥感技术和人工智能的融合为全球监测、环境变迁、资源管理等提供了新的视角和工具。特别是，遥感图像目标检测在这个交叉领域中占据了重要的地位。然而遥感图像高纬度、大规模、复杂性和动态变化的特性，给精准的遥感图像目标检测带来了挑战。目前遥感图像目标检测算法大多来源于自然图像目标检测算法的改进，传统的目标检测算法需要通过手动提取感兴趣区域的特征，提取特征方式繁琐低效，且提取特征十分单一。如何有效解读这些图像，从中提取并识别出有价值的信息，成为了目前遥感图像目标检测领域研究的关键问题之一。

当前以卷积神经网络为基础的目标检测方法主要分为两类：双阶段和单阶段检测算法，其主要的区别在于是否有候选框的生成。双阶段算法如 R-CNN[1]、Fast R-CNN[2]、Faster R-CNN[3]等，通过先生成目标候选区域再进行分类的方式实现目标检测，检测精度高但速度慢。单阶段检测算法如 SSD[4]和 YOLO[5]系列，舍弃了候选框生成阶段，利用卷积神经网络直接对输入图像进行特征提取，然后进行目标分类和位置预测，有效地提高了检测算法的识别速度。在这些单阶段检测算法中，YOLOv5 算法以模型训练和预测快速、便于模型部署等优点而受到广泛关注。尽管如此，针对遥感图像，这种算法仍存在一些问题，例如特征提取不足，对复杂背景的适应性差，以及对小目标的检测性能不佳。

对此，文献[6]运用空洞残差卷积的思想提取浅层特

收稿日期:2023-07-07； 修回日期:2023-08-03。
作者简介:李建新(1970-),男,硕士,副高级工程师。
通讯作者:范文龙(1998-),男,硕士研究生。
引用格式:李建新,范文龙,陈厚权. 基于改进 YOLOv5 的遥感图像目标检测研究[J].计算机测量与控制,2023,31(9):102-108,115.

征，随后与深层特征进行融合，有效提高了遥感图像中飞机的检测精度。姚艳清等[7]使用了一种双尺度特征融合模块，保证了图像特征的丰富性，以缓解深层信息的丢失问题，有效提高了多尺度遥感目标的检测能力。文献［8］提出了多阶段级联结构的遥感图像目标检测算法，在水平框和旋转框两个检测任务上均有提升。以上方法，虽然通过融合浅层特征和深层特征，保存了丰富的图像特征，但是对于目标尺度变化较大的图像易出现漏检的问题。文献［9］在 YOLOv5s 的骨干网络的卷积块中加入了一种即插即用的轻量级有效通道注意力（ECA，efficient channel attention）模块[10]，形成新的卷积有效通道注意力（CECA，convolutional efficient channel attention）模块，基于不降维的局部跨信道交互策略加强遥感目标的特征提取能力。此外，他们在多尺度特征融合的过程中引入具有 Swin Transformer[11]网络特性的 C3STR 模块和坐标注意力机制，以增强网络的局部感知能力，提高小尺度目标的检测精度。文献［12］在主干网络引入通道－全局注意力机制（CGAM，channel-global attention mechanism），以增强模型对不同尺度目标的特征提取能力和抑制冗余信息的干扰，解决了复杂背景的适应性差的问题。虽然通过添加注意力机制，提高了复杂背景的适应性以及小目标的检测性能，但对于云层阴影以及光照变化等不可抗因素的干扰，会产生严重的检测性能下降以及漏检等问题。

本研究提出了一种基于改进的 YOLOv5 模型的遥感图像目标检测算法，通过设计新的网络结构和优化策略来解决原始 YOLOv5 算法对遥感图像特征提取不足，对复杂背景的适应性差，以及对小目标的检测性能不佳等常见问题。并在遥感图像目标检测任务中取得了优秀的性能。根据在 RSOD 数据集上进行的测试结果显示，改进后的算法相比于原网络的检测效果在平均精度上提高了 10.6%，达到了 94.2%。

1　YOLOv5 算法

YOLOv5 网络结构分为输入端、主干、颈部和头部四部分。YOLOv5 在输入端采用了 Mosaic 数据增强，即将四张图片进行随机缩放、裁剪和排布并拼接在一起，可以大大丰富数据量较少的遥感图像数据集，同时进一步提升对小目标的检测性能。相较于 YOLOv3[15]和 YOLOv4[16]采用固定长宽比的锚框值，YOLOv5 中可以根据不同的数据集特点，自适应计算所需锚框的大小尺寸。相较于最新的 YOLOv7，YOLOv5 的训练和推理速度比 YOLOv7 快得多，并且具有较低的内存占用，这使得 YOLOv5 在移动设备或资源受限的应用场景中更具优势。输入网络之前，原始图片需要统一缩放到同一标准尺寸，YOLOv5 采用自适应图片缩放的方法，来为图像添加最少量的黑边，减少计算量并提升 YOLOv5 网络的推理速度。除了在输入端进行的优化，YOLOv5 基于 YOLOv4 网络在主干网络、颈部网络和损失函数部分又做了进一步的改进与提升。图 1 所示为 YOLOv5 的整体网络结构。需要注意的是，YOLOv5 的 V6.0 版本后网络第一层的 focus 模块替换成了的 6 * 6 的卷积层（conv），focus 模块原来的作用即为了实现无信息丢失的下采样。两个模块的作用是等效的，但是更换为 6 * 6 的卷积层会使得当前利用 GPU 进行检测网络计算时更加高效，更适合实际工程环境下进行部署使用。

YOLO 系列目标检测算法首先将图像输入到输入端进

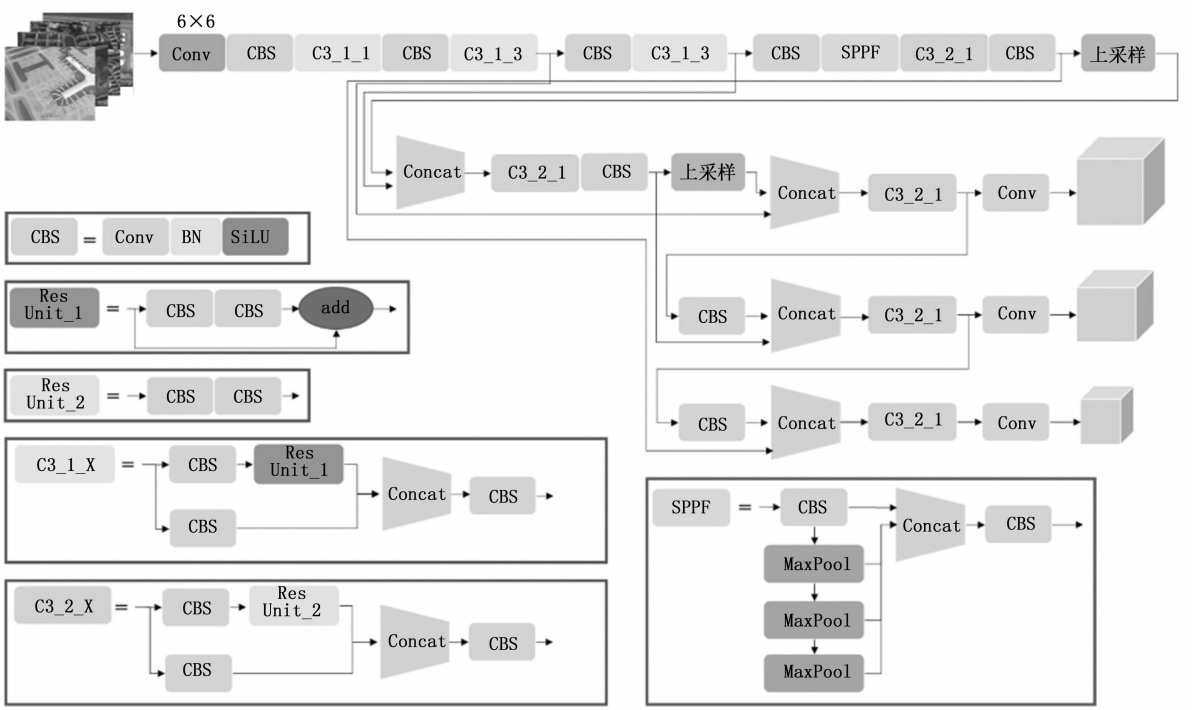

图 1　YOLOv5 的整体网络结构示意图

行马赛克（Mosaic）数据增强、自动拼接等预处理操作。同时将图像分成若干个区域，在每个区域的中心，聚类生成一系列设定初始长宽比的锚框。随后输入到骨干网络中，对图像进行三次连续下采样操作，生成三张不同分辨率的特征图，并通过特征融合模块对提取到的抽象语义信息和浅层特征信息进行重构与融合。进而，将特征融合模块输出的特征图输入到输出端进行预测，包括类别分类和目标边界框的回归预测。最后与真实标签框比较，计算差值来反向更新，进而迭代卷积神经网络参数。

2 改进 YOLOv5 算法

本文提出的改进 YOLOv5s 的遥感图像目标检测算法（ConvN-sim-yolo）的整体框架结构如图 2 所示。在骨干网络方面，使用 ConvNeXt 取代 Darknet53，提高捕获全局信息的能力。由于遥感图像存在云层阴影以及光照变化等不可抗因素的干扰，因此引入 SimAM 注意力机制，提高网络抗干扰能力。此外，引入 CFP 捕获全局长距离依赖关系以及遥感图像的局部关键区域信息，提高了遥感图像目标检测的精度。

2.1 主干网络使用 ConvNeXt

YOLOv5 主干网络依然沿袭使用 YOLO 系列常用的 Darknet53 网络，其借鉴了 ResNet 卷积神经网络的设计思想。2020 年以来，Transformer[17] 网络在各类图像识别领域表现优异，在图像分类等任务中超越传统卷积神经网络的性能。由于 CNN 网络中卷积操作仅能实现局部信息的捕获，全局信息的捕获受网络本身的结构限制无法实现。Transformer 则可以通过其固有的自注意力机制提取图像全局信息，扩大图像的感受野，获取更多的上下文信息，相较于 CNN 保留了更多空间信息。然而由于 Transformer 网络不具备 CNN 网络中的平移不变性、特征局部性等网络特点，只有在拥有大规模数据集进行网络训练时才能使得网络模型达到良好的检测效果。结合 RSOD 数据集其遥感图像数量少特点，设计使用基于 Transformer 风格的卷积神经主干网络——ConvNeXt。

ConvNeXt[18] 作为一种纯卷积网络，基于 ResNet50 网络，借鉴 Transformer 网络的设计思想从宏观设计，深度卷积，逆瓶颈化，大卷积核，微观设计这五个方面改进所得，三者的结构对比如图 2 所示。图 3（a）表示 Swin Transformer 模块结构，图 3（b）为 ResNet 模块，图 3（c）表示 ConvNeXt 模块，其中 d7×7 表示 7×7 大小的深度卷积。

1）宏观设计：首先改变阶段（stage）计算比率，如将 ResNet50 中残差块堆叠次数比值由从（3，4，6，3）更改为（3，3，9，3）与 Swin Transformer 网络中的堆叠比例相似。其次，ResNet50 的“stem cell”层包含一个大小为 7 * 7，步长（stride）为 2 的卷积层和一个最大池化层。参考 Transformer 网络使用的“patchify（修补）”策略，使用大小为 4 * 4，步长为 4 的“补丁”来替换 ResNet50 网络中的“stem cell”层。

2）深度卷积：此处借鉴 ResNeXt 网络中组卷积的思想，采用深度卷积替换 ResNet50 网络中的传统卷积层。深度卷积的操作与 Swin Transformer 网络中自注意力机制的加权求和类似，仅进行空间信息的交互，可降低网络的计算量 FLOPs。同时将网络宽度增加至与 Swin Transformer 相同的 96 通道数。

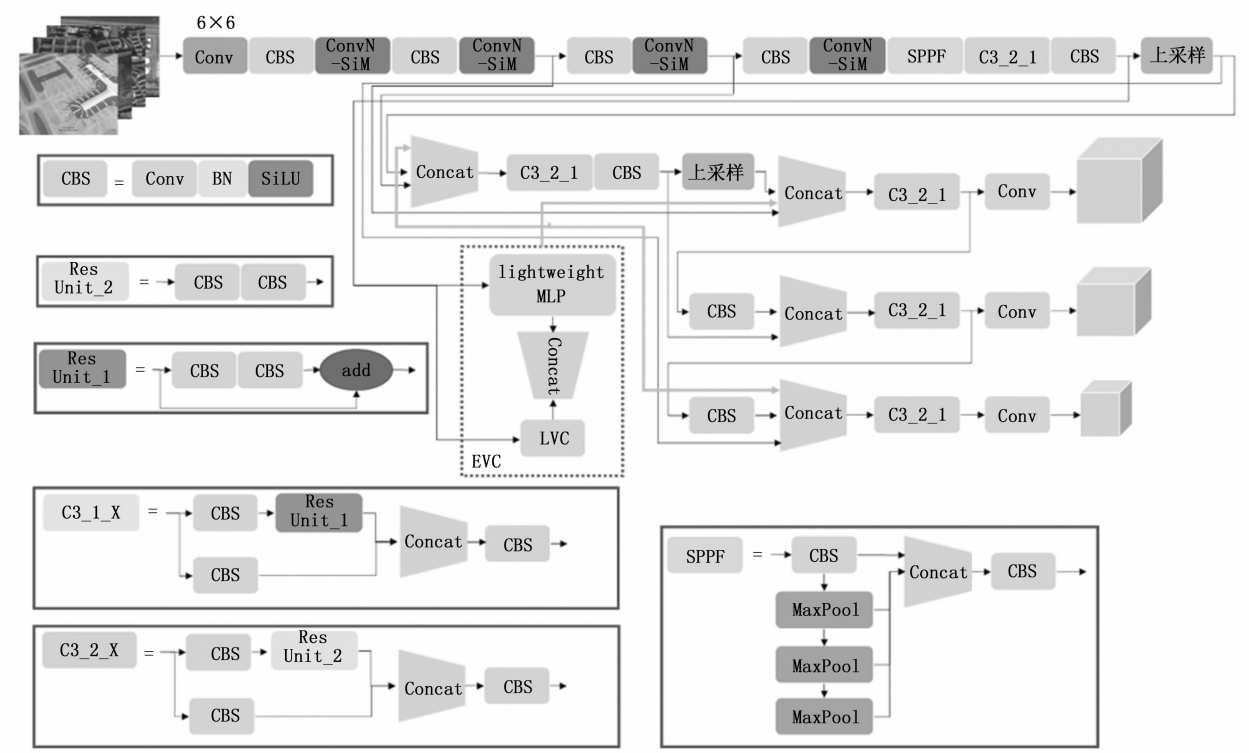

图 2 改进的 YOLOv5 的整体网络结构示意图

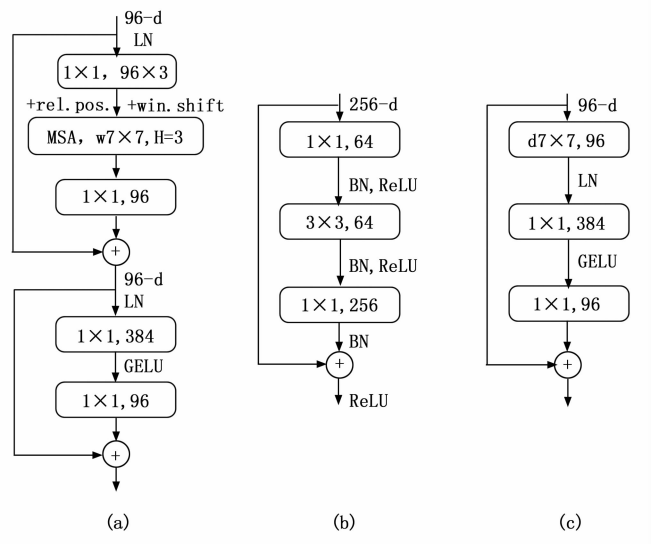

图 3　Swin Transformer、ResNet 和 ConvNeXt 模块结构对比示意图

3）逆瓶颈化：ConvNeXt 采用了 MobileNetV2 中设计的逆瓶颈结构，如图 4 所示，与 transformer 模块中 MLP 隐藏层是输入层的 4 倍宽的结构类似，减少整体网络的计算量 FLOPs，避免了降采样过程中小息肉特征信息的丢失，提升网络性能。

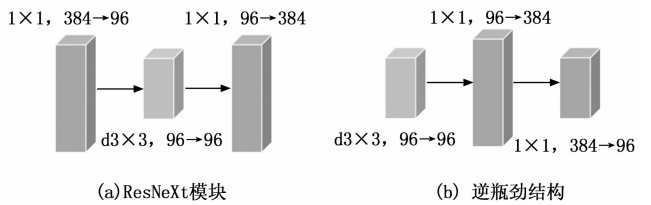

图 4　ConvNext 逆瓶颈化示意图

4）大卷积核：当前大多数卷积神经网络采用堆叠多个 3 * 3 卷积核来替代一个大尺寸卷积，使得感受野大小受限。将上述深度卷积的卷积核尺寸从 3 * 3 调整到与 Swin Transformer 的自注意力模块中 local 窗口大小相同的 7 * 7，利用大尺寸卷积核来增大感受野，获取更多的图像信息。

5）微观设计：基于 Transformer 网络的设计，ConvNeXt 替换 ReLU 激活函数为 GeLU 激活函数，并减少了其数量。使用更少的正则化层，并将 Batch Normalization（BN）正则化操作替换为 Layer Normalization（LN），使得模型更加稳定，减少梯度振荡。ConvNeXt 采用大小为 2 * 2，步距为 2 的卷积进行空间下采样，并在下采样操作之前以及全局池化以后增加 LN 正则化操作以维持训练的稳定性。

2.2　引入 SimAM 注意力机制

为了降低云层阴影、光照变化等复杂环境对检测任务的干扰，以提升网络的抗干扰能力，本研究在 ConvNext 模块中又增加了 SimAM 注意力机制。与现有常用的空间及通道注意力机制相比，SimAM 可实现在不增加 ConvNext 主干网络参数的同时推断出特征图 3D 注意力权值，以提升网络性能。图 5（a）所示为 ConvNext 模块结构，其由深度卷积（Deepwise conv）、层归一化（Layer Norm）、普通卷积和 GELU 激活函数组成。本研究将 SimAM 注意力机制添加在 ConvNeXt 模块中的深度卷积层之后，构成 ConvN-Sim 模块，如图 5（b）所示。

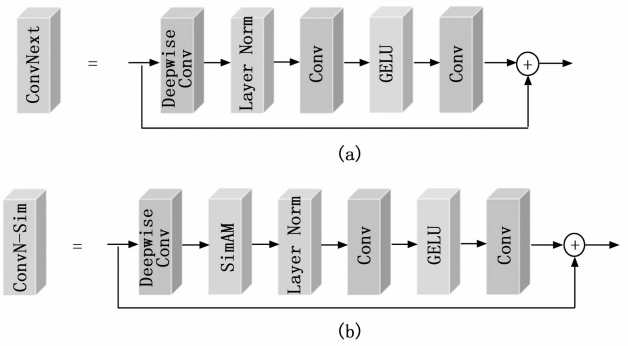

图 5　SimAM 注意力机制示意图

该注意力机制采用神经科学引导所得能量函数来计算注意力的权值，无须进行大量的工程性实验，最小能量计算如公式（1）所示。

$$e_t^* = \frac{4(\hat{\sigma}^2+\lambda)}{(t-\hat{\mu})^2+2\hat{\sigma}^2+2\lambda} \tag{1}$$

其中：t 代表输入特征 X 的目标神经元，$\hat{\mu}$ 和 $\hat{\sigma}^2$ 是除 t 外所有的神经元计算所得的平均值和方差，λ 的最佳取值为 0.000 1。由上述公式计算所得的值越小，表示能量低，神经元与周围神经元的差异大，可推导出该神经元的重要程度越高。依据各神经元的重要程度，对特征图进行增强，计算公式如（2）所示。

$$\widetilde{X} = sigmoid\ \frac{1}{E} \odot X \tag{2}$$

其中：$\widetilde{X}$ 表示增强后的特征图，是原始特征图 X 与 sigmoid 函数处理后的能量的逐元素乘积（$\odot$ 表示逐元素乘法）。X 表示输入特征图的原始值。E 代表每个神经元的能量，由式（1）计算得出，值越小表示神经元的重要程度越高。

2.3　采用集中特征金字塔结构

YOLOv5 所采用的 FPN（feature pyramid network）特征金字塔结构[19]，提出了一种自顶向下的层间特征交互方法。它可以为不同尺寸的目标提供相应尺度的特征表示，并提供上下文信息，融合多尺度特征信息以实现不同尺度下对不同大小目标的预测，最终提升检测网络的识别性能。然而当前计算机视觉中的特征金字塔方法研究重点在于层间特征交互而忽略了一些层内特征表示。由于卷积神经网络的固有特性，使得感受野大小受限，仅能捕获局部的上下文信息。本研究中，我们提出采用全局显式集中调节方案的集中特征金字塔（CFP，centralized feature pyramid），使用计算效率更高的轻量化多层感知机（MLP）来捕获全局长距离依赖关系，并行学习视觉中心（EVC，explicit visual center）机制捕获输入遥感图像的局部关键区域信息。

同时，由于深层特征常具有浅层特征所不具备的视觉集中特征表示[20]，为了提升层内特征调节的计算效率，针对常用的视觉特征金字塔结构，提出一种效率更高的自顶向下的全局集中调节（GCR，global centralized regulation）方法，从深层特征获取显示视觉中心特征来优化浅层特征，由此获得全面而具有差异化的图像特征表示。如图 6 所示为 YOLOv5 添加 CFP 模块（EVC+GCR）后的网络结构。

EVC 的计算公式如式（3）所示。

$$X = cat(MLP(X_{in}); LVC(X_{in})) \tag{3}$$

其中：X 为并行可学习视觉中心机制 EVC 的输出，X_{in} 为输入，cat（·）表示沿通道维度拼接特征图。MLP（X_{in}）和 LVC（X_{in}）分别表示轻量化多层感知机 MLP 和可学习视觉中心机制的输出特征。

轻量级 MLP 主要由深度卷积残差模块和通道 MLP 残差块组成，计算公式如下（4）和（5）所示。

$$\widetilde{X}_{in} = DConv(GN(X_{in})) + X_{in} \tag{4}$$

$$MLP(X_{in}) = CMLP(GN(\widetilde{X}_{in})) + \widetilde{X}_{in} \tag{5}$$

其中：$DConv$（·）表示大小为 1 * 1 的深度卷积核，$CMLP$（·）表示通道 MLP，$\widetilde{X}_{in}$表示基于深度卷积的模块输出，GN（·）表示群组归一化。

LVC 可学习视觉中心机制是带有特定字典的编码器，其计算过程如式（6）～（9）所示。

$$e_k = \sum_{i=1}^{N} \frac{e^{-S_k \| \check{x}_i - b_k \|^2}}{\sum_{j=1}^{K} e^{-S_k \| \check{x}_i - b_k \|^2}} (\check{x}_i - b_k) \tag{6}$$

其中：$\check{x}_i$ 为第 i 个像素点，b_k 为第 k 个可学习视觉码，S_k 为第 k 个比例因子。$\| \check{x}_i - b_k \|$ 是每个像素相对于视觉码的位置。K 为视觉中心总数。

$$e = \sum_{k=1}^{K} \Phi(e_k) \tag{7}$$

e 表示整个图像相对 K 个视觉码字的完整信息。

$$Z = X_{in} \otimes (\delta(Conv_{1\times1}(e))) \tag{8}$$

X_{in}表示输入特征，δ 表示比例因子系数，Z 表示局部边角区域特征，⊗表示通道乘法。

$$LVC(X_{in}) = X_{in} \oplus Z \tag{9}$$

其中：⊕表示通道加法。

2.4 采用 SIoU 损失函数

传统的边界框定位损失函数依赖于预测框与真实框之间的重叠面积、中心点距离等几何因素，并未考虑目标真实框与目标预测框之间的方向这一要素，导致检测网络收敛效率低下。本研究使用了新的边界框定位损失函数－SIoU loss[21]，通过在损失函数中引入边界框回归之间的向量角度，与传统损失函数方法（例如 CIoU 损失）相比，在网络训练阶段可以实现更快的收敛，并在推理方面实现更优越的准确性。SIoU loss 计算公式如（10）～（13）所示。

$$SIoU\ Loss = 1 - SIoU = 1 - IoU + \frac{\Delta + \Omega}{2} \tag{10}$$

$$\Lambda = 1 - 2 * \sin^2\left(\arcsin(x) - \frac{\pi}{4}\right) \tag{11}$$

$$\Delta = \sum_{t=x,y} (1 - e^{-(2-\Lambda)\rho_t}) \tag{12}$$

$$\Omega = \sum_{t=w,h} (1 - e^{-\omega_t})^{\theta} \tag{13}$$

其中：Λ 表示角度损失（Angle cost）函数、Δ 为基于角度损失考虑下的距离损失（Distance cost）函数，Ω 表示形状损失（Shape cost）代价函数。ρ_t 表示预测框和真实框的中心点之间的距离。ω_t 表示预测框和真实框的宽度和高度的差异。θ 表示调整形状损失影响程度的参数。

2.5 采取非极大值抑制

非极大值抑制 NMS 常用于目标检测网络中，在网络预测的最后过滤掉多余候选框，找到目标的最佳检测位置。为了避免当前检测框与得分最高的检测框 IoU 大于阈值时，该检测框被直接置零，造成相邻两个重叠的目标被漏检的现象出现，且同时能够对框与框之间的位置关系进行合理化的评估，本研究将 NMS、Soft-NMS 和 SIoU 结合，构建新的非极大值抑制方法 Soft-SIoU-NMS。加权后的 NMS 能够更好地解决在遥感图像实时检测过程中，相邻检测目标互相遮挡的检测问题，提升目标检测网络的最终效果。Soft-SIoU-NMS 的计算公式如下（14）和（15）所示。

$$s_i = \begin{cases} s_i & SIoU(M, b_i) < N_t \\ s_i\, f(SIoU(M, b_i)) & SIoU(M, b_i) \geqslant N_t \end{cases} \tag{14}$$

$$f(SIoU(M, b_i)) = e^{-CIoU(M, b_i)^2/\sigma} \tag{15}$$

其中：s_i 表示当前检测框的得分，b_i 表示目标预测框，N_t 表示 SIoU 的阈值，M 表示得分最高的检测框，f（·）表示高斯衰减函数，σ 取值 0.5。

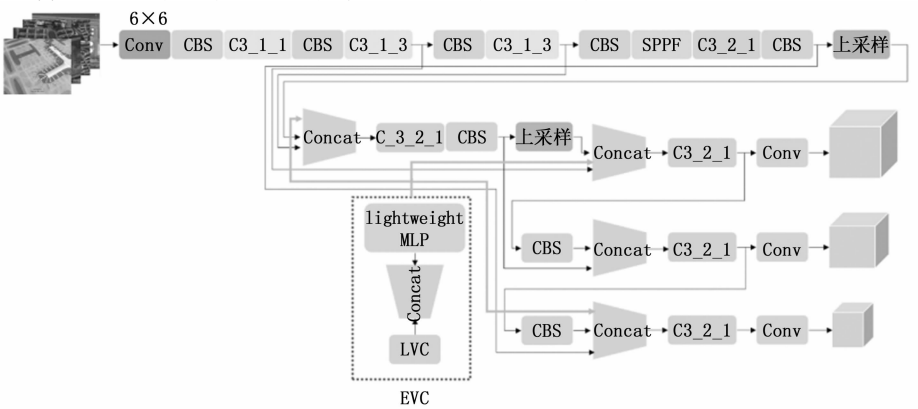

图 6 集中特征金字塔结构示意图

3 实验结果与分析

3.1 实验环境及参数设置

本实验深度学习框架为开源的 PyTorch 框架，PyTorch 是一个开源的 Python 机器学习库，是一个功能完备的框架，可用于构建深度学习模型，PyTorch 版本为 1.10.1。编程语言采用 Python 3.9.13，硬件设备配置为 Inter Core i7-7800X，使用的操作系统为 Ubuntu 18.04.5，GPU 为 NVIDIA GeForce RTX 2080Ti，CUDA 为 10.2。

在训练过程中，采用随机梯度下降算法（SGD，stochastic gradient descent）训练 200 epoch。初始学习率设置为 0.01，并采用步长衰减的策略来降低学习率，在每个 epoch 后，将学习率降低 10%以避免模型在后期训练过程中陷入局部最优解。实验表明此学习率在保证模型在初期快速收敛的同时，且不会造成梯度爆炸或消失。基于硬件配置和模型的复杂性综合考虑，将批量大小（BatchSize）设置为 32，工作线程数（num _ workers）设置为 8。在该设置环境下模型能够在硬件上稳定运行，同时获得合理的训练速度。研究发现该设置环境使得模型在遥感图像目标检测任务上达到最佳性能。

3.2　数据集

本文所使用的数据集为 RSOD 遥感数据集。该数据集于 2015 年由武汉大学发布用于遥感图像目标检测的标准数据集，共有 976 张图片，6 950 个实例，包括飞机 446 张图片 4 993 个实例、油罐 165 张图片 1 586 个实例、立交桥 176 张图片 180 个实例和操场 189 张图片 191 个实例。RSOD 数据集是 PASCAL VOC 格式作为规范，为满足 yolo 训练的数据集格式，将 PASCAL VOC 格式转为 yolo 格式。从中选取 546 张图片作为训练集，137 张图片作为验证集，剩余的 293 张图片作为测试集。

3.3　评价指标

为了评估 ConvN-Sim-YOLO 网络的遥感图像目标检测的性能，引入常用于深度学习测试网络性能评价的六个重要指标，包含精确度（Precision）、召回率（Recall）、平均正确率（AP）和平均类别 AP（mean Average Precision，mAP）。这五者计算如公式（16）～（19）所示。

$$Precision = \frac{TP}{TP + FP} \tag{16}$$

精确度表示预测为阳性的样本中真正为阳性样本的比例。其中 TP 表示真阳性，即将阳性预测为阳性的数量；FP 表示假阳性，即将阴性预测为阳性的数量。

$$Recall = \frac{TP}{TP + FN} \tag{17}$$

召回率则表示预测正确的阳性样本占全部阳性样本的比例。式中 TP 表示真阳性，即将真阳性预测为阳性的数量；FN 表示假阴性，即将真阳性预测为阴性的数量。

$$AP = \frac{1}{11} \sum_{r \in (0, 0.1, \cdots, 1)} p_{\text{interp}}(r) \tag{18}$$

此平均正确率（AP）值的计算为 11 点计算方法，每个点处取该点右侧最大精确率，然后结合召回率＝［0，0.1，0.2，0.3，0.4，0.5，0.6，0.7，0.8，0.9，1.0］，绘制出 P-R 曲线，并通过（19）公式相加求平均值。

$$mAP = \frac{\sum_{j=1}^{class_{num}} AP_j}{class_{num}} \tag{19}$$

其中：$class\_num$ 代表类别总数，本研究实验取值为 4，AP_j 代表第 j 个类别的平均正确率。mAP 表示各类别 AP 的均值，描述网络对所有类别检测的最终效果。mAP@0.5 表示将 IoU 设置为 0.5 时，每一类图片的 AP 值再求平均，mAP@0.5：0.95 表示 IoU 在区间［0.5，0.95］内取值，步距间隔 0.05 计算一个 mAP 值，再将这些 mAP 值总和求平均。

3.4　消融实验

为验证本文改进的 ConvNeXt 主干网络、SimAM 注意力机制、CFP 结构、非极大值抑制方法（SIoU 损失函数、NMS 非极大值抑制）的有效性，进行消融实验，评估各个改进模块对本文检测算法的影响。消融实验以原始的 YOLOv5s 实验结果作为基准，实验数据如表 1 所示。

表 1　消融实验结果

改进网络结构	精确度%	召回率%	mAP
YOLOv5	83.6	84.5	83.6
YOLOv5＋ConvNext	87.6	89.8	90.5
YOLOv5＋Convext＋Sim	88.1	87.6	91.5
YOLOv5＋ConvNext＋Sim＋CFP	91.0	91.2	93.7
YOLOv5＋ConvNext＋Sim＋CFP＋Soft-SIoU-NMS	91.9	92.5	94.2

由表 1 可知，原始 YOLOv5 在 RSOD 数据集上可获得 83.6%的精确度，84.5%的召回率以及 83.6 的 mAP 的结果，逐步增加改进的四个模块后检测各指标基本都有提升，表明各个模块都有助于遥感图像目标检测任务，也验证了优化特征捕获能力、特征表达能力和加强抗干扰能力的出发点的合理性。进一步，首先将 YOLOv5 的主干网络 Darknet53 替换成 ConvNeXt 后精确度从 83.6%提升至 87.6%，召回率提升了 5.3%以及 mAP 值提升了 6.9，证明了改进后模型会捕获更多的全局空间信息。其次引入 SimAM 注意力机制后精确度提高了 0.5%，召回率略有所下降，mAP 值提升 1%，证明 SimAM 注意力机制提升网络在检测任务中的抗干扰能力。随后引入 CFP 精确率和召回率进一步提升 2.9%和 3.6%，mAP 值显著提升至 93.7%，证明网络可获取全局上下文信息。此外将原始的 NMS 替换为 Soft-SIoU-NMS 后精确度达到 91.9%，召回率达 92.5%，mAP 值提升了 0.5，证明此处改进，解决了目标互相遮挡的检测问题。最后，当集成四个改进模块时可将遥感图像目标检测的精确度提升至 91.9%，召回率提升至 92.5%，mAP 提升至 94.2，有效验证了所提出的遥感图像目标检测方法的有效性。

3.5　不同算法对比实验

为验证本文提出的改进的 YOLOv5 目标检测算法相比于其他主流算法具有更好的目标检测能力，将提出的算法与现有主流算法进行对比实验。选取六种模型包括典型的目标检测模型如 SSD、Faster-RCNN、YOLOv3 等以及最近基于 YOLOv5 改进的 Swin-YOLOv5s[9] 进行对比实验。实验采用相同的遥感图像数据集 RSOD，实验结果如表 2 所示。

由表 2 可知，改进的方法在 RSOD 取得了最优的 mAP 值结果。与原始的 YOLOv5s 相比，虽然对于操场类别的准确率有所下降，但对于其他类别的准确率以及 mAP 值都有

表 2 不同算法在 RSOD 数据集的检测结果对比

网络模型	各类别准确率%				mAP
	飞机	油桶	立交桥	操场	
SSD	52.1	96.6	56.7	100	76.4
Faster R-CNN	63.1	84.1	76.9	97.8	80.5
YOLOv3	62.2	95.1	70.4	98.6	81.6
YOLOv4	81.3	98.1	71.7	100	87.8
YOLOv5	89.7	79.4	71.4	94	83.6
Swin-YOLOv5s	90.4	85.8	81.5	97.9	88.9
ConvN-sim-yolo	92.8	95	92.6	87.1	94.2

大幅度的提升。此外，与最近基于 YOLOv5 改进的 Swin-YOLOv5s 相比，本文改进方法 mAP 值达到 94.2，Swin-YOLOv5s 只有 88.9。虽然，本文算法和其他算法相比，检测操场目标的精度有所下降，但大部分类别的目标检测精度都有所提高，总体来看本文提出的改进方法能够有效提升遥感图像中的目标检测精度，在精度性能方面优势明显。

本文对 RSOD 数据集上的测试结果进行了可视化展示，如图 7 所示。图 7（a，b，c）为改进前基础的 yolov5 网络检测效果，图 7（d，e，f）为本文提出的 ConvN-sim-yolo 网络检测效果。通过比较图 7（a）和图 7（d），检测结果表明 Conv-sim-yolo 对于遥感图像小目标的检测具有更高的精度。进一步比较图 7（b）和图（e），图中有飞机处于阴暗或光照和阴暗交界处，而基础的 yolov5 对于处于云层阴影和光照变化的飞机检测效果并不好，而 Conv-sim-yolo 能够很好的检测此类飞机，展现出 Conv-sim-yolo 在抗干扰方面的优越性。图 7（c）和图 7（f）的检测结果表明，Conv-sim-yolo 能够察觉不易被发现的小目标，减少了漏检率，进而提升了遥感图像目标检测的精度值。

(a) YOLOv5s检测结果1 (b) YOLOv5s检测结果2

(c) YOLOv5s检测结果 (d) ConvN-sim-yolo检测结果1

(e) ConvN-sim-yolo检测结果2 (f) ConvN-sim-yolo检测结果3

图 7 在 RSOD 数据集上的可视化检测结果

4 结束语

针对遥感图像目标检测中存在云层阴影和光照变化干扰以及小目标漏检的问题，本文提出了 ConvN-sim-yolo 算法。首先，使用基于 Transformer 风格的卷积神经主干网络--ConvNeXt，捕获遥感图像全局信息，充分提取遥感图像丰富的特征。其次，为了应对云层阴影和光照变化的干扰，在 ConvNext 模块中加入 SimAM 注意力机制，推断特征图的 3D 注意力权值，提高了网络的稳定性和抗干扰能力。同时提出采用全局显式集中调节方案的集中特征金字塔（CFP，centralized feature pyramid），使用计算效率更高的轻量化多层感知机（MLP）来捕获全局长距离依赖关系，并行学习视觉中心（EVC，explicit visual center）机制捕获输入遥感图像图像的局部关键区域信息，降低了漏检的概率，展现了优越的目标检测的性能。最后经过实验对比，本文算法相比于原始的 YOLOv5s，平均检测准确率在 RSOD 数据集上提升了 10.6%，由此表明本文算法在遥感图像目标检测领域上改善了小目标漏检的问题，并对于更复杂的干扰环境依然具有良好的目标检测性能。但是，改进的 yolov5 算法在目标检测能力上还有一定的局限性，未来工作将继续优化网络以提升算法的目标检测能力。

参考文献：

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

[2] GIRSHICK R. Fast R-CNN [C] //2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015: 1440 - 1448.

[3] REN S, HE K, GIRSHICK R, SUN J. 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.

[4] LIU W, et al. SSD: Single Shot MultiBox Detector [C] // LEIBE B, MATAS J, SEBE N, WELLING M, eds. Computer Vision-ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9905. Cham: Springer, 2016.

[5] 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. Las Vegas: IEEE, 2016: 779 - 788.

（下转第 115 页）