Computer Science and Application 计算机科学与应用, 2022, 12(6), 1518-1528
Published Online June 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.126151

Hans汉斯

改进YOLOX中特征融合结构的目标检测方法

杨　利[1]，李允臣[1]，王家宝[1*]，赵志杰[2]，李　阳[1]，苗　壮[1]

[1]陆军工程大学指挥控制工程学院，江苏　南京
[2]31700部队，辽宁　沈阳

收稿日期：2022年5月6日；录用日期：2022年6月6日；发布日期：2022年6月13日

摘　要

因无人机俯拍视角的特殊性，航拍目标在成像中呈现出小尺度/多尺度、外观相似度高、背景复杂干扰大等特点，导致航拍目标检测相对通用目标检测更具挑战和难度。为了解决该问题，针对通用目标检测中常用于融合多尺度特征的路径聚合网络(Path Aggregation Network, PANet)模块，本文提出一种改进PANet的多距离关联依赖MDAD (Multi-Distance Association Dependency)模块，该模块包含跨层连接和同层连接两种连接方式，通过密集的跨尺度交互融合增强不同尺度特征层的弱特征信息。同时，基于YOLOX框架和所提出的MDAD模块，构建了更加适合航拍多尺度复杂目标的检测方法。在公开的典型航拍目标检测数据集VisDroneDet上，实验验证了本文所提方法的有效性。所提模块可适用于在不同模型大小的主干网络上进行扩展，具有较好的实际应用价值。

关键词

目标检测，YOLOX，特征融合，路径聚合网络

Improving Object Detection Method of Feature Fusion Structure in YOLOX

Li Yang[1], Yunchen Li[1], Jiabao Wang[1*], Zhijie Zhao[2], Yang Li[1], Zhuang Miao[1]

[1]Command & Control Engineering College, Army Engineering University of PLA, Nanjing Jiangsu
[2]31700 Troop, Shenyang Liaoning

Received: May 6th, 2022; accepted: Jun. 6th, 2022; published: Jun. 13th, 2022

[*]通讯作者。

文章引用: 杨利, 李允臣, 王家宝, 赵志杰, 李阳, 苗壮. 改进 YOLOX 中特征融合结构的目标检测方法[J]. 计算机科学与应用, 2022, 12(6): 1518-1528. DOI: 10.12677/csa.2022.126151

Abstract

Due to the particular camera view of unmanned aerial vehicles, the captured objects show the characteristics of small-scale/multi-scale, high similar appearance, complex background and large interference in imaging, which makes it more challenging and difficult than general object detection. In order to solve this problem, based on the Path Aggregation Network (PANet) module, which is often used to fuse multi-scale features in general target detection networks, this paper proposes a Multi-Distance Association Dependency (MDAD) module by improving PANet, which includes two connection modes, Connection across Different Layers (CDL) and Connection on the Same Layer (CSL). The weak feature information of different scale feature layers is enhanced through intensive cross-scale interactive fusion. At the same time, based on the YOLOX framework and the proposed MDAD module, an object detection method suitable for aerial multi-scale complex objects is proposed. Experiments verified the effectiveness of the proposed MDAD module on the public aerial object detection dataset (VisDroneDet). The proposed module is suitable for expansion on backbone networks with different model sizes, and has good practical application value.

Keywords

Object Detection, YOLOX, Feature Fusion, Path Aggregation Network

1. 引言

目前无人机目标检测已广泛应用于战场情报侦察、重要目标鉴别、矿产资源勘探、灾情环境监测等军用、民用各个领域[1]-[6]，与通用目标检测相比，无人机目标检测因无人机俯拍视角的特殊性，目标在图像中呈现出小尺度/多尺度、外观相似度高、背景复杂干扰大等特点，很难对目标进行高精度的定位检测。图 1 展示了 VisDroneDet 数据中影响精度的主要因素：1) 目标呈现小尺度/多尺度：航拍大视场场景中目标多以小的尺寸呈现，且目标尺度会随航拍高度呈现一定的多尺度变化。2) 目标外观相似高：航拍高度越高，目标的尺度越小、像素越少，其不同类别目标间的差异难以体现，如小汽车、面包车、货运车等差异性会被弱化，区分不同的目标类别更难。3) 背景复杂干扰大：不同于道路监控、海面监测等单一背景的目标检测任务，无人机航拍场景存在着城市街区、深山老林、道路交错等各类复杂场景，这些场景中的光线、阴影、遮挡等现象会影响着目标外观、大小，并带来一定的噪声。上述因素容易导致目标检测器出现误检、漏检，降低目标检测模型性能。

近年来，基于卷积神经网络的目标检测模型[7]-[12]在通用目标检测数据集[13] [14] [15]上取得了长足的进步，不断刷新检测记录。其中 YOLO (You Only Look Once)一阶段目标检测方法经历了 YOLOv1 [16]、YOLOv2 [17]、YOLOv3 [7]、TinyYOLO、YOLOv4 [8]、YOLOv5、YOLObile [18]、YOLOF [19]的改进发展，因其时效性优越被广泛应用于实际工程项目中。2021 年旷视科技发表了该系列最新改进算法 YOLOX [20]，并取得了非常不错的检测效果。但是，针对无人机航拍目标检测任务，在通用目标检测数据集训练出来性能出众的检测模型会存在着跨域适配问题，需要研究新的适配航拍场景目标检测的特定检测器部件或模块来提升性能。

Figure 1. The main factors affecting the accuracy of images in the VisDroneDet dataset
图 1. VisDroneDet 数据集图像影响精度的主要因素

作为 YOLO 系列最新的、效果最佳的目标检测模型 YOLOX，其通过对主干网络的多个阶段输出特征进行融合实现了鲁棒的特征提取。其中，融合特征提取模块是基于特征金字塔网络(Feature Pyramid Network, FPN)改进的路径聚合网络(Path Aggregation Network, PANet) (如图 2(a)所示)，包含向上和向下两个方向的相邻阶段不同输出特征的融合操作。现有研究已经表明[21] [22] [23] [24]，主干网络不同阶段提取的特征是不同的，底层阶段输出的特征包含更多的纹理形状等信息，而高层阶段输出的特征包含更多的语义类别信息。YOLOX 的路径聚合网络实现了对相邻阶段特征的融合，但是缺乏像 HRNet [25]、DenseNet [26]网络类似的对非相邻阶段特征的跨层融合，会存在一定程度上的特征融合提取不充分问题。

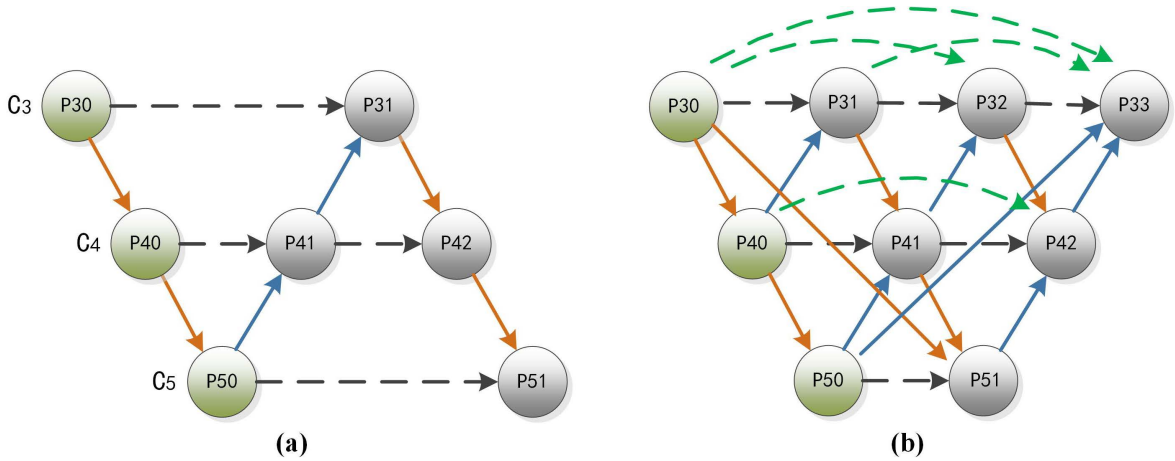

Figure 2. (a) Structure of PANet, (b) Structure of MDAD
图 2. (a) 路径聚合网络结构，(b) 多距离关联依赖结构

为了提升 YOLOX 中 PANet 模块对特征的融合提取能力，我们提出一种改进 PANet 的多距离关联依赖(Multi-Distance Association Dependency, MDAD)模块，该模块结构如图 2(b)所示。具体地，该模块包含跨层连接(Connection across Different Layers, CDL)和同层连接(Connection on the Same Layer, CSL)两种连接方式，其中 CDL 跨层连接对不同阶段特征层进行充分外部特征融合；CSL 同层连接对同阶段特征层进行二次内部特征融合，增强特征表达和定位精度。与 YOLOX 中的 PANet 模块相比，改进模块可以更好

地提取无人机航拍捕获的目标特征，提升模型的检测性能。

本文贡献及创新点如下：

1) 针对通用目标检测中常用于融合多尺度特征的路径聚合网络(Path Aggregation Network, PANet)模块，提出了一种改进 PANet 的多距离关联依赖 MDAD (Multi-Distance Association Dependency)模块。

2) 基于 YOLOX 框架和所提出的 MDAD 模块，构建了更加适合航拍多尺度复杂目标的目标检测方法。在公开的典型航拍目标检测数据集 VisDroneDet 上，实验验证了本文所提模块和方法的有效性。

2. 相关工作

2.1. 一阶段目标检测

2016 年，以速度快、实用强著称的 YOLOv1 [16]模型诞生，自此一阶段目标检测器得到了迅速发展和广泛应用。同年，Liu Wei 等人提出单发多框检测器(Single Shot Multibox Detector, SSD) [9]，该模型使用 VGG-16 作为主干网，新增加了 4 层不同尺寸的特征图，并利用不同尺寸的特征图进行不同尺度目标的检测，对多尺度目标尤其是小目标可以获得了更好的检测精度，但是由于不同层特征图之间没有进行融合，特征图包含的目标信息有限。2017 年，Tsung-Yi Lin 等人提出 FPN [27]，该模型使用性能更好的 ResNet-50 作为主干网，并将语义信息丰富的深层特征图进行 2 倍上采样后，与位置细节信息丰富的浅层特征图进行横向连接融合，从而得到不同尺度的特征金字塔，有效地增强了特征图的特征表示能力。同年，Redmon 等人提出 YOLOv3 [7]，该模型借鉴了残差网络 ResNet 的设计思想，设计了更加高效强大的主干网 Darknet-53，并借鉴了 FPN 的设计思想，对不同层次特征图进行了更加充分地融合，进一步提高了小目标的检测精度。

以上方法虽然在特征提取上通过融合多层特征实现了更强的特征表示，但是特征融合后的分类与定位任务中，分类和定位的检测头都是耦合在一起的，因为分类与定位任务不同，导致耦合任务存在冲突，会降低检测精度。2021 年，YOLOX [20]以 YOLOv3 SPP 版本为基础进行改进，将分类与定位的检测头解耦，分别计算定位损失、分类和置信度损失，并采用数据增强、无锚框设计和先进的标签分配方法 SimOTA (Simple Optimal Transport Assignment)等策略，在 MS COCO 数据集上达到了目前最优的性能。但是，YOLOX 中 PANet 模块仅对相邻阶段特征进行融合，缺乏像 HRNet [25]、DenseNet [26]网络类似的对非相邻阶段特征的跨层融合，会存在一定程度上的特征融合提取不充分问题。

2.2. 特征金字塔网络

目标检测是一项对目标尺度变化和空间位置信息都较敏感的计算机视觉任务，因此特征金字塔网络 FPN [27]及其改进方法 PANet [20]已广泛用于该领域以此提高多尺度目标检测性能。这些方法对不同阶段获取的特征图通过改变分辨率大小统一到相同尺度进行通道特征的拼接或合并，使得浅层细节、位置信息在深层表示的更丰富，深层语义信息在浅层表示的更完整，实现上下文彼此关联。然而，它们的特征融合仅仅体现在相邻特征层之间，忽视与其它特征层信息的交互依赖，降低语义差异效果受限。同时，在相同特征层之间也缺少如 DenseNet 这样的密集跳连融合，这也会带来特征位置偏差。

为了更好地处理多尺度特征，Tan Mingxing 等人在 EfficientDet 中提出的加权双向特征金字塔网络 BiFPN (Bidirectional Feature Pyramid Network) [28]，通过引入类似 Attention 权重，更好地平衡不同尺度的特征信息。Guo Caoxu 等人针对 FPN 融合前高层特征传递信息会随通道减少而丢失、不同特征层融合时会产生语义差异及融合后直接忽视某一特征层信息会带来融合特征不完整的问题，提出了 AugFPN (Augmented FPN)特征金字塔结构[29]，利用设计的一致性监控、剩余特征增强和软 RoI (Region of Interest)选择三个模块解决上述特征融合缺陷。Tsung-Yi Lin 等人采用递归 FPN 的方式提出了 Recursive-FPN

(Recursive Feature Pyramid Network) [30]，将传统 FPN 融合后的特征输出作为输入返回给 Backbone，进行二次循环特征融合。然而，这些改进的特征金字塔网络因为缺乏不同特征层之间的远距离关联，彼此还存在着语义差异。GiraffeDet [31]采用了轻 Backbone、重 Neck 的设计，虽然最大化地实现了特征的充分融合，但是同时也增加了过多计算成本。ASFF (Adaptively Spatial Feature Fusion) [32]的融合方法实现了每个特征层的输出节点特征都来自所有特征层输入节点的特征融合，不仅关联了相邻特征层间的语义信息，而且建立了不同特征层之间的远距离依赖，减少了语义差异，但因忽略横向跳连，使得特征融合不够充分。

3. 本文方法

3.1. 模型架构

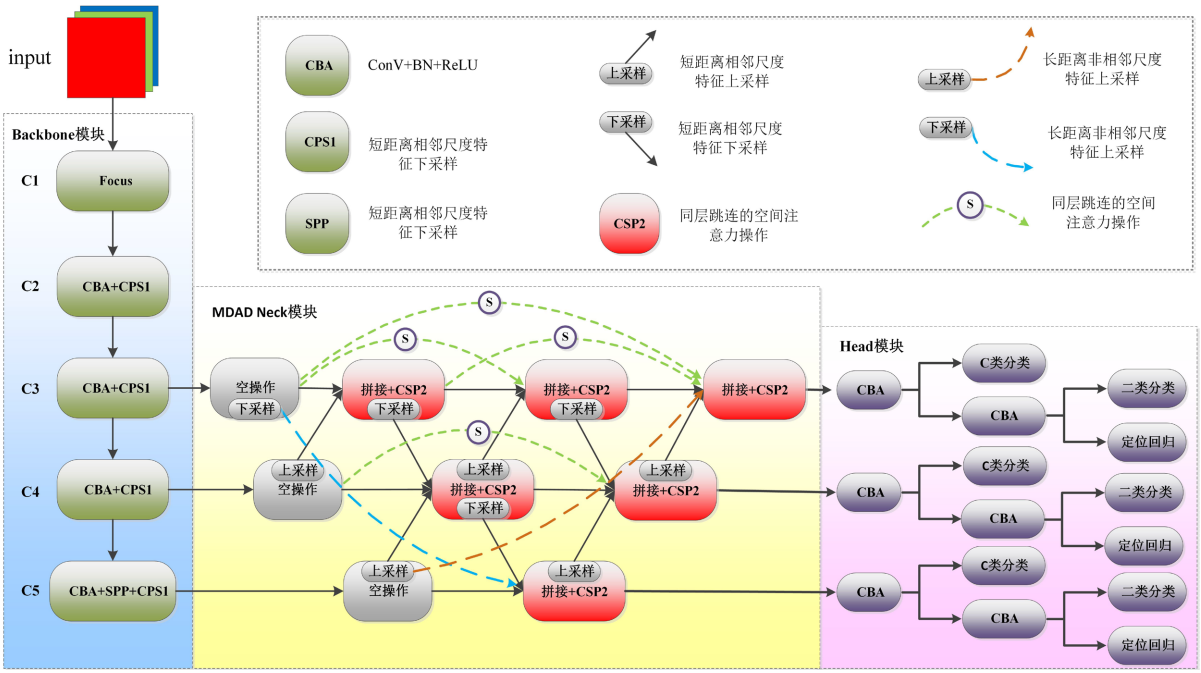

Figure 3. The architecture of the proposed model
图 3. 所提模型架构

图 3 展示了本文所提模型的整体架构，由用于特征提取的 Backbone 模块、用于特征融合的 MDAD Neck 模块、用于预测输出的 Head 模块三个部分构成。

Backbone 模块：采用和 YOLOX 一样的结构。对于一幅统一尺度的输入图像，主干网采用 Focus、CBS、CPS1 和 SPP 等卷积、下采样操作，在阶段 C3、C4、C5 输出特征，为后续特征融合并最终预测目标位置和类别提供多尺度特征信息。随着主干网深度的增加，输出特征图分辨率(尺寸)不断下降，低维细节空间特征会不断减少，高维抽象语义信息不断增加。借鉴 YOLOX，本文也采用阶段 C3、C4、C5 输出三个不同尺度特征图 $\{\mathbf{X}_{3,0},\mathbf{X}_{4,0},\mathbf{X}_{5,0}\}$ 作为后续特征融合及预测的主要特征。

MDAD Neck 模块：为了克服 YOLOX 中 PANet Neck 模块存在的层间和层内特征相互嵌入、彼此关联不足的问题，本文借鉴 HRNet [25]和 DenseNet [26]等稠密连接结构，提出了一种新的 MDAD Neck 模块，该模块通过同层跳连和跨层互联的连接方法，实现了层间和层内特征的充分融合。具体包括：一个双线性插值或转置卷积的上采样操作、CBS 卷积或空洞卷积的下采样操作、将不同阶段或同一阶段不同

层特征进行 Concat 拼接操作，以及用于定位目标的注意力机制操作。这些操作实现了对不同阶段、不同尺寸特征的融合，并对目标区域特征进行注意力增强，具体结构见 3.2 节。

Head 模块：与 YOLOX 类似，本文在 Head 模块中将预测目标的类别、位置和置信度进行解耦计算，不仅能提高模型检测精度，而且能加快模型训练的收敛速度。该 Head 模块对 MDAD 输出的三个不同尺度的特征分别进行解耦操作，每个解耦操作都是先通过 1 × 1 大小的通道卷积来压缩通道数、降低计算量，再分解成判断类别、判断前背景和定位目标三个分支操作。其中，判断类别、判断前背景属于分类任务，通过两个 3 × 3 和一个 1 × 1 卷积映射到目标类别数和有无目标 2 类别数，再利用 Sigmoid 操作转化为预测概率；定位目标属于回归操作，也是通过两个 3 × 3 和一个 1 × 1 卷积映射至所需回归的 4 个坐标值(中心点横坐标、纵坐标、宽、高)。训练过程中，Head 模块之后采用与 YOLOX 一样的损失函数。

3.2. MDAD 模块结构

图 3 中展示的 MDAD 模块结构，由跨层连接(Connection across Different Layers, CDL) (不同尺度大小特征层层间互连)和同层连接(Connection on the Same Layer, CSL) (相同尺度大小特征层层内互连)两种信息传递路径，与三种尺度 6 个融合特征输出节点构成。该模块输入为 Backbone 模块第 3、4、5 阶段输出的多尺度特征(C3、C4、C5)，输出为 MDAD 模块处理后的多尺度特征，特征个数和维度与输入相同。该模块能融合特征表示多尺度特征，具有更充分的特征提取能力。其中：

3.2.1. 跨层连接 CDL

跨层连接是纵向连接相邻和非相邻阶段不同尺度的特征，通过汇集多尺度特征实现对低层细节特征与高层语义特征的融合。具体又可分为短距离相邻尺度特征融合和长距离非相邻尺度特征融合。

短距离相邻尺度特征连接：采用双线性插值上采样操作和步长为 2 的普通卷积下采样操作来统一相邻阶段的特征尺度。双线性插值将深层特征图的宽、高放大为原来的 2 倍，通道数缩小 0.5 倍；步长为 2 的普通卷积操作将浅层特征图的宽、高缩小为原来的 0.5 倍，通道数增大 2 倍。以上上采样和下采样操作为不同尺度特征的交互连接提供了统一的尺度，形式化表示如下：

$$\hat{\mathbf{X}}_{i,j}^{U_2} = U_2\left(\mathbf{W}_{i,j}^{1} \otimes \mathbf{X}_{i,j}\right) \tag{1}$$

$$\hat{\mathbf{X}}_{i,j}^{D_2} = \mathbf{W}_{i,j}^{2} \otimes \left(\mathbf{W}_{i,j}^{1} \otimes \mathbf{X}_{i,j}\right) \tag{2}$$

其中，$\mathbf{X}_{i,j}$ 表示第 i 阶段第 j 个层的输出特征，$\otimes$ 表示卷积操作，$\mathbf{W}_{i,j}^{1}$ 表示大小为 1 × 1、步长为 1 的卷积操作参数，$U_2(\cdot)$ 表示双线性插值上采样，卷积用于改变特征通道数、上采样用于改变特征高度和宽度，目的是将第 i 阶段的特征 $\hat{\mathbf{X}}_{i,j}^{U_2}$ 与第 $i-1$ 阶段特征 $\mathbf{X}_{i-1,j}$ 尺度统一；$\mathbf{W}_{i,j}^{2}$ 表示大小为 3 × 3、步长为 2 的卷积操作参数，用于改变特征高度和宽度，目的是将第 i 阶段的特征 $\hat{\mathbf{X}}_{i,j}^{D_2}$ 与第 $i+1$ 阶段特征 $\mathbf{X}_{i+1,j}$ 尺度统一。

长距离非相邻尺度特征连接：采用转置卷积上采样操作和空洞卷积下采样操作来统一不同阶段的特征尺度。转置卷积将深层特征图的宽、高放大为原来的 4 倍，通道数减少原来的 0.25 倍，其相对于双线性插值信息冗余更小、映射能力更强。空洞卷积(膨胀率依次为 r = 1、r = 2、r = 3)操作将浅层特征图的宽、高缩小为原来的 0.25 倍，通道数增大 4 倍，其相对于普通卷积能够获得更大的感受野。形式化表示如下：

$$\hat{\mathbf{X}}_{i,j}^{U_4} = U_4^{\mathrm{T}}\left(\mathbf{W}_{i,j}^{3} \otimes \mathbf{X}_{i,j}\right) \tag{3}$$

$$\hat{\mathbf{X}}_{i,j}^{D_4} = \mathbf{W}_{i,j}^{4} \otimes \left(\mathbf{W}_{i,j}^{3} \otimes \mathbf{X}_{i,j}\right) \tag{4}$$

其中，$\mathbf{X}_{i,j}$ 表示第 i 阶段第 j 个层的输出特征，$\otimes$ 表示卷积操作，$\mathbf{W}_{i,j}^{3}$ 表示大小为 1 × 1、步长为 1 的卷

积操作参数，$U_4^{\mathrm{T}}(\cdot)$表示转置卷积上采样，卷积用于改变特征通道数、上采样用于改变特征高度和宽度，目的是将第 i 阶段的特征 $\hat{\mathbf{X}}_{i,j}^{U_4}$ 与第 $i-2$ 阶段特征 $\mathbf{X}_{i-2,j}$ 尺度统一；$\mathbf{W}_{i,j}^4$ 表示大小为 3 × 3 空洞卷积操作参数，用于改变特征高度和宽度，目的是将第 i 阶段的特征 $\hat{\mathbf{X}}_{i,j}^{D_4}$ 与第 $i+2$ 阶段特征 $\mathbf{X}_{i+2,j}$ 尺度统一。

3.2.2. 同层连接 CSL

同层连接横向串连或跳连同尺度特征，汇聚同一尺度不同处理过程的特征进行融合。具体有可分为短距离串连和长距离跳连。对于同层连接的每个节点，除了接收不同阶段传递来的特征，还接收同阶段相邻节点特征和非相邻节点特征，所有特征以拼接方式组合后，采用卷积块注意力模块(Convolutional Block Attention Module, CBAM) [33]实现对不同的多组特征进行充分融合，以及对目标主要区域进行注意力加权，增强对目标区域特征的表示能力，提升目标定位性能。

3.2.3. 特征汇聚融合

MDAD 模块特征汇聚融合过程具体如下：

假设 $\mathbf{X}_{i,j}$ 表示第 i (i = 3、4、5)阶段第 j 个操作节点输出特征(当 j = 0 时，$\mathbf{X}_{i,j}$ 表示由主干网络输出的特征；当 $j>0$ 时，$\mathbf{X}_{i,j}$ 表示 MDAD 模块融合后输出特征)，则该特征计算过程为

$$\mathbf{X}_{i,j}=\mathbf{W}_{i,j}^5\otimes\left[\tilde{\mathbf{X}}_{i,j},\delta_{j=2}\left(S_A\left(\mathbf{X}_{i,j-2}\right)\right),\delta_{j=3}\left(S_A\left(\mathbf{X}_{i,j-3}\right)\right)\right]\quad j>0 \tag{5}$$

其中，$S_A(\cdot)$表示空间域注意力机制[33]，$\delta_{j=t}(\cdot)$表示该项仅在 $j=t$ 时存在，$\otimes$表示卷积操作，$[\,]$表示按通道拼接操作，$\mathbf{W}_{i,j}^5$ 表示大小为 1 × 1、步长为 1 的卷积操作参数，用于改变特征通道数，$\tilde{\mathbf{X}}_{i,j}$ 计算过程如下：

$$\tilde{\mathbf{X}}_{i,j}=C_A\left(\left[\mathbf{X}_{i,j-1},\hat{\mathbf{X}}_{i+1,j}^{U_2},\delta_{j=3}\left(\hat{\mathbf{X}}_{i+2,j-1}^{U_4}\right)\right]\right)\quad i=3,j>0 \tag{6}$$

$$\tilde{\mathbf{X}}_{i,j}=C_A\left(\left[\mathbf{X}_{i,j-1},\hat{\mathbf{X}}_{i+1,j}^{U_2},\hat{\mathbf{X}}_{i-1,j}^{D_2}\right]\right)\quad i=4,j>0 \tag{7}$$

$$\tilde{\mathbf{X}}_{i,j}=C_A\left(\left[\mathbf{X}_{i,j-1},\hat{\mathbf{X}}_{i-1,j}^{D_2},\hat{\mathbf{X}}_{i-2,j-1}^{D_4}\right]\right)\quad i=5,j>0 \tag{8}$$

其中，$C_A(\cdot)$表示 CBAM 注意力机制。当 $j>0$ 时，特征 $\mathbf{X}_{i,j}$ 由不同阶段、不同层特征融合后经 CBAM 注意力机制生成得到。

4. 实验

4.1. 数据集及参数设置

数据集：实验采用 VisDroneDet2019 数据集进行训练和测试。该数据集由无人机高空对街区、交通等场景拍摄回去的图像组成，包括 12 个类别标签(行人、人、自行车、小汽车、敞篷车、货车、三轮车、人力三轮车、公共汽车、摩托车，外加一个其他类和忽略区域)，8629 张图像，其中 6471 张用于训练，548 张用于验证，1610 张用于测试。单张图像宽度在 2000 像素左右，目标总数达到 2,600,000 个。该数据集图像成像于昼夜全时段，绝大多数图像以复杂场景为背景且密集遮挡情况严重，单张图像目标类别和个数较多，如图 1。

参数设置：训练时，训练总迭代次数为 100 个 epoch，并前 2 个 epoch 采用热启动 Warmup，在 85 个 epoch 后终止 Moasic 增强。使用随机梯度下降(Stochastic Gradient Descent, SGD)优化器和余弦退火衰减学习机制调整学习率，初始学习率为 0.01，SGD 的动量设置为 0.9，权重衰减为 0.0005。使用 2 块英伟达 2080Ti GPU 计算，Batchsize 大小设为 16。测试时，Batchsize 大小设为 64。同时，使用了 Mosaic [8]、

Mixup [20]数据增强方法，其中 Mosaic 方法通过 4 张图像的随机拼接再缩放至相同输入尺寸大小，不仅增加了小目标的数量，也加大了背景的复杂程度。类似的，Mixup 方法通过叠加不同图像方式模拟了遮挡目标的效果。为了验证本文所提方法的有效性，将 YOLOX 模型作为基线模型，其中与本文方法的差异主要是 Neck 部分采用了不同的模块。

4.2. 实验结果与分析

定量实验结果分析

本文所提结构被称为 MDAD 结构(如图 2(b))，它是在基准模型的 PANet 结构[34] (如图 2(a))基础上提出的。为了有效论证本文所提方法的有效性，我们同时还构建了另外两种不同的 Neck 结构(如图 4(a)、图 4(b))。其中，图 4(a)展示了在 C3、C5 层中间添加两个特征输出节点，采用更加全面复杂的连接方式使每个特征输出节点与其它节点建立相互依赖关系，这种密集连接的结构被称为 FCP (Fully Connected parallel)结构；图 4(b)展示了将 C5 层的 P52 节点移到 C3 层中，形成倒梯形结构来加强浅层特征的表达能力，这种结构我们称为 IT (Inverted Trapezoid)结构。

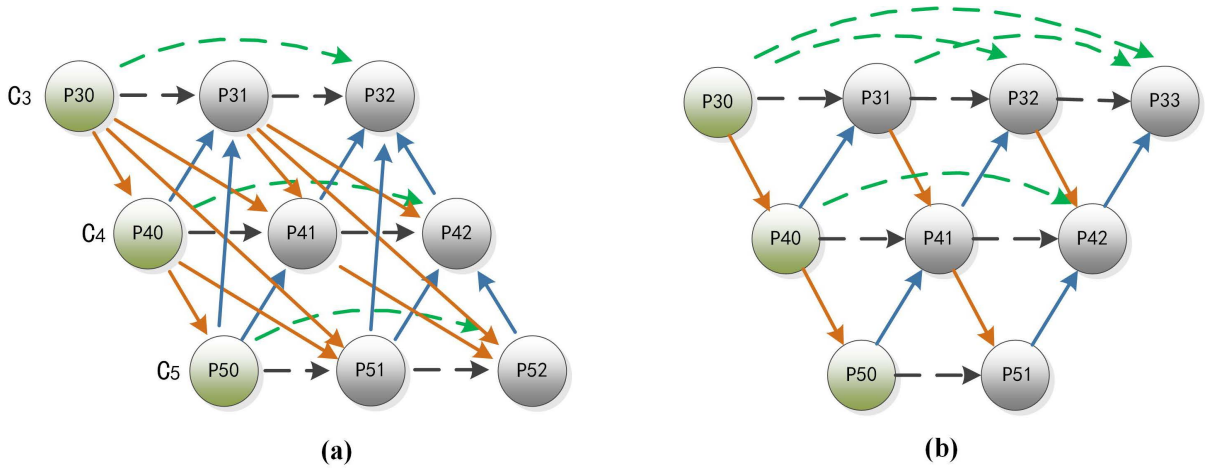

Figure 4. Schematic diagram of comparative analysis of different feature fusion Neck structures
图 4. 不同特征融合 Neck 结构对比分析示意图

实验结果如表 1 所示。其中，评测指标 AP、AR 均采用数据集标准评测指标[2]。

Table 1. Experimental results of the improved method on the VisDroneDet2019 dataset
表 1. 改进方法在 VisDroneDet2019 数据集上实验结果

Method	AP [%]	AP50 [%]	AP75 [%]	AR1 [%]	AR10 [%]	AR100 [%]	AR500 [%]
YOLOX (PANet)	22.24	40.29	22.02	0.06	0.72	8.13	32.62
YOLOX (FCP)	21.74	39.17	21.57	0.06	0.71	7.49	31.90
YOLOX (IT)	22.26	40.16	22.04	0.06	**0.73**	**8.51**	32.73
YOLOX (MDAD)	**22.54**	**40.67**	**22.41**	**0.06**	0.70	8.46	**33.10**

由表 1 可以发现，

1) 所提方法 YOLOX (MDAD)的 AP、AP50、AP75 精度分别为 22.54%、40/67%和 22.41%，较 YOLOX (PANet)的 22.24%、40.29%和 22.02%的精度有一定的提升。除了 AR10 指标外，YOLOX (MDAD)的 AR

指标均超过了 YOLOX (PANet)，验证了本文所提方法的有效性。

2) 对比 YOLOX (FCP)和 YOLOX (IT)，YOLOX (IT)具有明显的性能优势，因为 IT 相对于 FCP 将 C5 层的最后一个特征输出节点移到了 C3 层，目的是增加浅层特征的表达能力，同时也去掉了冗余相邻层间的下采样融合；将所提方法与 YOLOX (IT)对比，可以发现所提方法在 AR10 和 AR100 上稍低于 YOLO (IT)，但是在精度上有者较大的优势；

3) 对比 YOLOX (PANet)和 YOLOX (FCP)，可以发现 YOLOX (PANet)要更好，说明并不是任意增加连接就可以提升检测性能，需要进行一定的选择设计；同时，对比 YOLOX (IT)和 YOLOX (PANet)，可以发现 YOLOX (IT)相对更好，说明倒梯形结构相比于原网络结构性能更好。

4) 对比不同的 Neck 结构还可以发现，航拍图像目标检测的效果不是取决于特征输出节点数量的多少(PANet 的节点数量为 4 个，FCP、IT、MDAD 的节点数量相同均为 6 个)，而是与结构中特征输出节点的位置有关，倒梯形结构要比平行连接结构好。

4.3. 定性可视化效果分析

为了更好地展示本文方法的有效性及不足，本文对图 1 展示的 6 幅测试图像进行预测，结果如图 5 所示，其中左上角的图像中存在大量密集的目标，且绝大部分都被有效检测到，但是，也可以发现中下方红色圆框位置存在多个骑车人的目标未被检测到。其他 5 幅图像总体上检测目标都较准，也存在个别目标漏检的问题。

Figure 5. Detection results of the proposed method
图 5. 所提方法检测结果

5. 结论

针对航拍目标检测中存在小尺度/多尺度、外观相似度高、背景复杂干扰大等挑战和问题，本文基于

路径聚合网络结构模块，提出一种新的多距离关联依赖 MDAD 结构模块。同时，基于所提出的 MDAD 结构模块，改进 YOLOX 更加适合航拍目标检测。在公开的典型航拍目标检测数据集 VisDroneDet 上，实验验证了本文所提方法的有效性。

参考文献

[1] Gu, J., Su, T., Wang, Q., Du, X. and Guizani, M. (2018) Multiple Moving Targets Surveillance Based on a Cooperative Network for Multi-UAV. *IEEE Communications Magazine*, **56**, 82-89. https://doi.org/10.1109/MCOM.2018.1700422

[2] Zhu, P., Wen, L., Du, D., Bian, X., Ling, H., Hu, Q., *et al*. (2018) VisDrone-DET2018: The Vision Meets Drone Object Detection in Image Challenge Results. *European Conference on Computer Vision* (*ECCV*) *Workshops*, Munich, 8-14 September 2018, 437-468. https://doi.org/10.1007/978-3-030-11021-5_27

[3] Hird, J.N., Montaghi, A., Mcdermid, G.J., Kariyeva, J., Moorman, B.J., Nielsen, S.E., *et al*. (2017) Use of Unmanned Aerial Vehicles for Monitoring Recovery of Forest Vegetation on Petroleum Well Sites. *Remote Sensing*, **9**, Article No. 413. https://doi.org/10.3390/rs9050413

[4] Pajares, G. (2015) Overview and Current Status of Remote Sensing Applications Based on Unmanned Aerial Vehicles (UAVs). *Photogrammetric Engineering and Remote Sensing*, **81**, 281-329. https://doi.org/10.14358/PERS.81.4.281

[5] Kellenberger B., Volpi M. and Tuia D. (2017) Fast Animal Detection in UAV Images Using Convolutional Neural Networks. *IEEE International Geoscience and Remote Sensing Symposium* (*IGARSS*), Worth, 23-28 July 2017, 866-869. https://doi.org/10.1109/IGARSS.2017.8127090

[6] Shao, Z., Li, C., Li, D., Altan, O., Zhang, L. and Ding, L. (2020) An Accurate Matching Method for Projecting Vector Data into Surveillance Video to Monitor and Protect Cultivated Land. *ISPRS International Journal of Geo-Information*, **9**, Article No. 448. https://doi.org/10.3390/ijgi9070448

[7] Redmon, J. and Farhadi, A. (2018) YOLOv3: An Incremental Improvement.

[8] Bochkovskiy, A., Wang, C.-Y. and Liao, H.-Y.M. (2020) YOLOv4: Optimal Speed and Accuracy of Object Detection.

[9] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., *et al*. (2016) SSD: Single Shot MultiBox Detector. *European Conference on Computer Vision* (*ECCV*), Amsterdam, 11-14 October 2016, 21-37. https://doi.org/10.1007/978-3-319-46448-0_2

[10] Lin, G., Milan, A., Shen, C. and Reid, I. (2017) RefineNet: Multi-path Refinement Networks for High-Resolution Semantic Segmentation. *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 5168-5177. https://doi.org/10.1109/CVPR.2017.549

[11] Ren, S., He, K., Girshick, R.B. and Sun, J. (2015) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031

[12] He, K., Gkioxari, G., Dollár, P. and Girshick, R. (2020) Mask R-CNN. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **42**, 386-397. https://doi.org/10.1109/TPAMI.2018.2844175

[13] Porat, B. and Friedlander, B. (1990) A Frequency Domain Algorithm for Multiframe Detection and Estimation of Dim Targets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **12**, 398-401. https://doi.org/10.1109/34.50625

[14] Everingham, M., Eslami, S. M.A., Gool, L.V., Williams, C.K.I., Winn, J. and Zisserman, A. (2014) The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision*, **111**, 98-136. https://doi.org/10.1007/s11263-014-0733-5

[15] Xie, L., Liu, Y., Jin, L. and Xie, Z. (2019) DeRPN: Taking a Further Step toward More General Object Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**, 9046-9053. https://doi.org/10.1609/aaai.v33i01.33019046

[16] Redmon, J., Divvala, S.K., Girshick, R.B. and Farhadi, A. (2016) You Only Look Once: Unified, Real-Time Object Detection. *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 779-788. https://doi.org/10.1109/CVPR.2016.91

[17] Redmon, J. and Farhadi, A. (2017) YOLO9000: Better, Faster, Stronger. *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 6517-6525. https://doi.org/10.1109/CVPR.2017.690

[18] Cai, Y., Li, H., Yuan, G., Niu, W., Li, Y., Tang, X., *et al*. (2021) YOLObile: Real-Time Object Detection on Mobile Devices via Compression-Compilation Co-Design. *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**, 955-963.

[19] Chen, Q., Wang, Y., Yang, T., Zhang, X., Cheng, J. and Sun, J. (2021) You Only Look One-level Feature. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 13034-13043.

https://doi.org/10.1109/CVPR46437.2021.01284

[20] Ge, Z., Liu, S., Wang, F., Li, Z. and Sun, J. (2021) YOLOX: Exceeding YOLO Series in 2021.

[21] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) ImageNet Classification with Deep Convolutional Neural Networks. *Communications of the ACM*, **60**, 84-90. https://doi.org/10.1145/3065386

[22] Simonyan, K. and Zisserman, A. (2015) Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR, abs/1409.1556.

[23] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., *et al*. (2015) Going Deeper with Convolutions. *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, 7-12 June 2015, 1-9. https://doi.org/10.1109/CVPR.2015.7298594

[24] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90

[25] Sun, K., Xiao, B., Liu, D. and Wang, J. (2019) Deep High-Resolution Representation Learning for Human Pose Estimation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 5686-5696. https://doi.org/10.1109/CVPR.2019.00584

[26] Huang, G., Liu, Z., Weinberger, K.Q. and Van Der Maaten, L. (2017) Densely Connected Convolutional Networks. *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 2261-2269. https://doi.org/10.1109/CVPR.2017.243

[27] Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B. and Belongie, S. (2017) Feature Pyramid Networks for Object Detection. *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 936-944. https://doi.org/10.1109/CVPR.2017.106

[28] Tan, M., Pang, R. and Le, Q.V. (2020) EfficientDet: Scalable and Efficient Object Detection. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 10778-10787. https://doi.org/10.1109/CVPR42600.2020.01079

[29] Guo, C., Fan, B., Zhang, Q., Xiang, S. and Pan, C. (2020) AugFPN: Improving Multi-Scale Feature Learning for Object Detection. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 12592-12601. https://doi.org/10.1109/CVPR42600.2020.01261

[30] Qiao, S., Chen, L.-C. and Yuille, A.L. (2021) DetectoRS: Detecting Objects with Recursive Feature Pyramid and Switchable Atrous Convolution. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 10208-10219. https://doi.org/10.1109/CVPR46437.2021.01008

[31] Jiang, Y., Tan, Z., Wang, J., Sun, X., Lin, M. and Li, H. (2022) GiraffeDet: A Heavy-Neck Paradigm for Object Detection. ArXiv, abs/2202.04256.

[32] Liu, S., Huang, D. and Wang, Y. (2019) Learning Spatial Fusion for Single-Shot Object Detection.

[33] Woo, S., Park, J., Lee, J.-Y. and Kweon, I.S. (2018) CBAM: Convolutional Block Attention Module. *European Conference on Computer Vision* (*ECCV*), Munich, 8-14 September 2018, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

[34] Liu, S., Qi, L., Qin, H., *et al*. (2018) Path Aggregation Network for Instance Segmentation. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Salt Lake City, 18-23 June 2018, 8759-8768. https://doi.org/10.1109/CVPR.2018.00913