doi: 10.12052/gdutxb.230044

基于YOLOv5的轻量化无人机航拍小目标检测算法

李雪森[1]，谭北海[2]，余　荣[1]，薛先斌[1]

（1. 广东工业大学 自动化学院, 广东　广州　510006；2. 广东工业大学 集成电路学院, 广东　广州　510006）

摘要: 针对无人机航拍视角下图像目标特征尺寸小且存在背景复杂、分布密集的问题，提出了一种基于YOLOv5的轻量化无人机航拍小目标检测改进算法GA-YOLO。该算法改进了Mosaic数据增强方法和网络整体结构，并增加了微小物体检测头，同时设计了轻量化的全局注意力模块和并行结构的空间通道注意力机制模块，提高了网络的全局特征提取能力和训练过程中卷积通道之间的竞争和合作关系。以4.0版本的YOLOv5s为基准，在公开无人机航拍数据集VisDrone2019-DET上实验，结果表明，改进后的模型相较于原模型，参数量下降了48%，计算量下降了26%，而mAP@0.5提高了4.9个百分点，mAP@0.5:0.95提高了3.3个百分点，有效地提高了无人机空中视角下对密集型小目标的检测能力。

关键词: 无人机航拍；YOLOv5s；小目标检测；数据增强；注意力机制

中图分类号: TP391.41　　**文献标志码:** A　　**文章编号:** 1007–7162(2024)03–0071–10

Small Target Detection Algorithm for Lightweight UAV Aerial Photography Based on YOLOv5

Li Xue-sen[1], Tan Bei-hai[2], Yu Rong[1], Xue Xian-bin[1]

(1. School of Automation, Guangdong University of Technology, Guangzhou 510006, China; 2. School of Integrated Circuits, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: A lightweight unmanned aerial vehicle (UAV) aerial photography small target detection algorithm GA-YOLO based on YOLOv5 is proposed to address the problem of small target feature size, complex background, and dense distribution in images from the perspective of UAV aerial photography. This algorithm improves the Mosaic data augmentation method and overall network structure, and adds a small object detection head. At the same time, a lightweight global attention module and a parallel spatial channel attention mechanism module are designed to enhance the network's global feature extraction ability and the competition and cooperation between convolutional channels during the training process. Based on the 4.0 version of YOLOv5s, experiments were conducted on the publicly available drone aerial photography dataset VisDrone2019-DET. The results showed that the improved model reduced the number of parameters by 48% and the computational complexity by 26% compared to the original model, and mAP@0.5 improved by 4.9 percentage points, mAP@0.5 0.95 increased by 3.3 percentage points, effectively enhancing the detection capability of unmanned aerial vehicles for dense small targets from an aerial perspective.

Key words: UAV aerial photography; YOLOv5s; small target detection; data enhancement; attention mechanism

近年来，基于深度学习的目标检测技术在图像识别、定位和追踪方面取得了非常大的进步，涌现出许多优秀的算法，这使得无人机可以与目标检测技术结合在一起，在军事、民用领域发挥出重要作

收稿日期： 2023-03-04

基金项目： 国家自然科学基金资助项目(61971148)；国家自然科学基金资助项目(U22A2054)；广东省基础与应用基础研究基金联合基金重点项目(2019B1515120036)；广西自然科学基金重点项目(2018GXNSFDA281013)

作者简介： 李雪森(1997-)，男，硕士研究生，主要研究方向为深度学习，E-mail：18325945913@163.com

通信作者： 谭北海(1980-)，男，副教授，博士，硕士生导师，主要研究方向为AI算法及芯片设计、人工智能、深度学习，E-mail：bhtan@gdut.edu.cn

用[1-2]。然而，与自然场景图像不同的是，无人机视角下的图像数据是非常复杂的，由于无人机飞行的高度较高、拍摄采集图像时为俯视角度且飞行的过程中会造成高度和拍摄角度的变化，导致其提取的图像数据中的目标物体尺寸过小，同一物体的尺寸变化剧烈，不容易被正确检测，因此设计一种适用于无人机视角的小目标检测模型以满足实际应用的需求，是一项具有重要意义和挑战性的研究课题。

当前，基于深度学习的无人机航拍图像目标检测已成为研究热点，主流的目标检测算法根据是否生成候选区域可分为“一阶段”和“两阶段”两类，其中“一阶段”代表算法主要有SSD系列[2-4]、YOLO系列[5-8]、Retina-Net[9]等，这类算法不需要事先生成候选区域，而是直接采用原始特征来对目标进行类别和位置检测，因此检测速度非常快，但难以保证较高的检测精度；“二阶段”代表算法主要有Fast-RCNN[10]、Faster-RCNN[11]、SPP-Net[12]等，这类算法的基本原理是先进行候选区域生成，再通过卷积网络进行分类和定位回归预测，这种处理方式检测精度较高，但是速度稍慢。为了将目标检测技术以性能最大化的方式应用在无人机航拍图像处理任务中，国内外众多学者提出了多种优秀的算法。例如，LI P等[13]提出了一种基于MobileNet轻量化网络和YOLOv3目标检测算法的组合网络，该方法通过引入一种通道注意力机制，实现了对多尺度遥感目标检测精度的提升，但在检测速度上并不占优势。T LI等[14]在YOLOv4的基础之上引入一种超轻量子空间注意力机制，用于为特征图的每个子空间导出不同的注意力特征图，从而增强了网络的多尺度特征表示能力，改善了由于遮挡而导致的漏检目标的情况，但该方法计算量过大且训练成本较高。WANG M等[15]提出了一种基于上下文场景的注意力融合网络，通过对原始目标特征和场景上下文信息的交互，实现了精度和速度上的双重提升，但是该方法并没有优化遮挡性目标的检测准确性问题。

综上，对于无人机航拍目标检测的高复杂性以及目前已有研究还存在的一些急需解决的问题，本文从多个处理角度入手，设计了一种基于YOLOv5的轻量化无人机航拍小目标检测算法，主要贡献为

(1) 针对无人机采集的图像数据中目标尺寸较小且排列密集的特点，本文对Mosaic数据增强方法进行改进，改善了其对于小目标的数据增强恶化情况，使输入网络中的目标有效特征变多。

(2) 为使网络在降低模型复杂度同时保持较高的检测精度，本文使用轻量化网络ShffleNet中部分模块重构YOLOv5的骨干网络和特征融合网络，并在原有的检测尺度基础上增加一个微小物体检测头。

(3) 为提升不同特征通道之间的竞争合作关系和对局部特征的关注程度，通过对CBAM注意力机制进行改进，设计了一种并行结构的注意力模块。

(4) 为提高网络的全局信息提取能力，借鉴ViT网络中的Transformer结构和ShuffleNet中的“混洗”机制的特点，设计了一种轻量化的全局注意力模块(ShuffleViT)，通过将该模块嵌入到骨干网络中，有效提高了网络的全局信息提取能力。

1 本文方法

GA-YOLO(Global attention YOLO)算法的主要思想有两个方面，一是通过改进Mosaic数据增强算法提高网络的输入有效特征，二是通过轻量化处理和加入改进注意力模块使网络在保持较高推理速度的同时提高对小目标的检测能力。其网络结构框架如图1所示。其中，输入端使用本文改进的Mosaic数据增强算法进行数据预处理；主干网络部分使用嵌入全局注意力模块(ShuffleViT) 的Shu-ffleNet网络组成；检测头部分扩展了尺度，增加一个微小物体检测头，并嵌入本文提出的并行注意力模块。

1.1 Mosaic数据增强改进

Mosaic数据增强方法的原理是将4张缩放到固定尺寸的图像进行翻转、色域调整等变换后，以掩码图层上的随机点为中心将4张图像进行拼接，再将其缩放为指定的输入特征尺寸大小，最终输入到神经网络中进行学习。这种方法能够增加数据的多样性、丰富图像的背景，并增加小目标的数量、提高神经网络检测小目标的能力。但在处理航拍图像这种原始图像尺寸较大且目标特征尺寸较小的数据时，Mosaic可能会恶化小目标的数据增强效果。如图2所示，假设网络输入特征尺寸为640×640，对于尺寸为1 280×1 280的原始图像数据进行等比例缩放并使用Mosaic处理后，图像尺寸将缩小为原尺寸的1/16，这会导致图像数据中小目标的特征信息大量减小，甚至可能会丢失有效特征。因此，在处理原始图像尺寸较大的数据集时，有必要对输入特征尺寸缩放比例和数据增强方法进行优化，以避免小目标特征信息的过度丢失。

针对以上Mosaic存在的问题，本文提出一种简单有效的优化解决方法：在Mosaic数据增强处理过

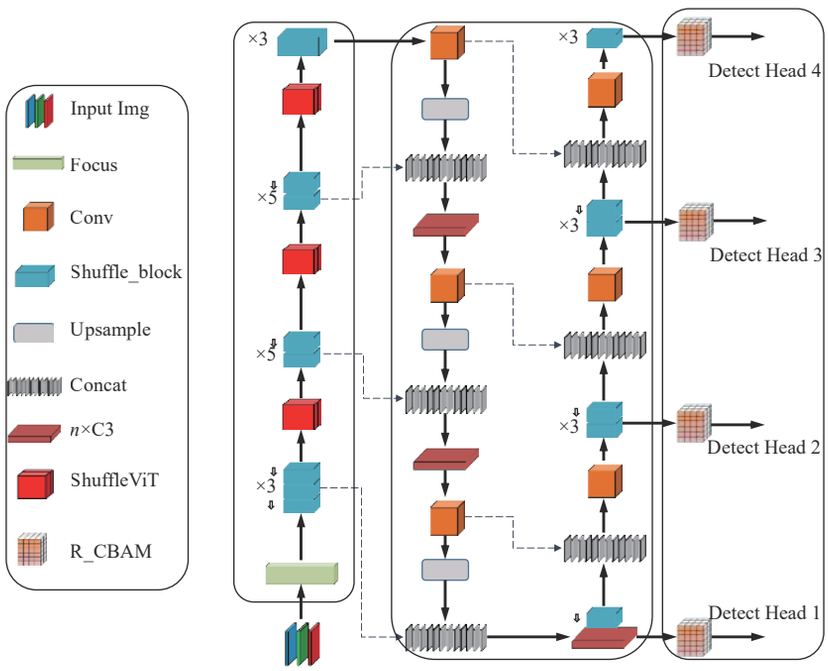


图 1　GA-YOLO整体结构

Fig.1　GA-YOLO overall structure

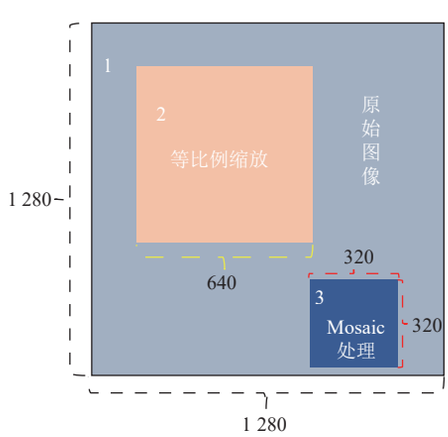


图 2　Mosaic图像缩放比例

Fig.2　Mosaic image scaling

程中引入一个图像尺寸放大系数α，根据α缩放因子调节送入Mosaic处理的图像尺寸，从而提高输入神经网络的小目标有效特征。具体来讲，α的作用是调节输入Mosiac的图像尺寸，并不改变输入网络训练特征的大小。例如，以640×640标准尺寸作为输入特征(即$\omega_{\text{img_size}}$ = [640, 640])，令α=1.31，此时输入Mosaic的图像尺寸将被调整至840×840(如图3所示)。利用Resize操作裁剪部分子图，舍弃部分原始图像数据，虽然减少了部分原始图像的特征输入，损失了部分背景特征信息，但却使输入到网络中的小目标有效特征信息增多，使得网络能够学习到更多小目标的特征，从而提升网络的检测性能。改进后的模型称之为α-Mosaic。定义以$\omega_{\text{img_size}}$表示原始输入特征大小，h、w表示原始图像的高和宽，h^m、w^m表示Mosaic中掩码层的高和宽，$\boldsymbol{T}_{\text{img}}^{h'\times w'}$表示经过初步尺寸调整过后的图像数据，$\boldsymbol{T}^{h\times w}$表示经过Mosaic处理之后的输出图像特征。则$\alpha$-Mosaic的推理过程可如式(1)~(3)所示。

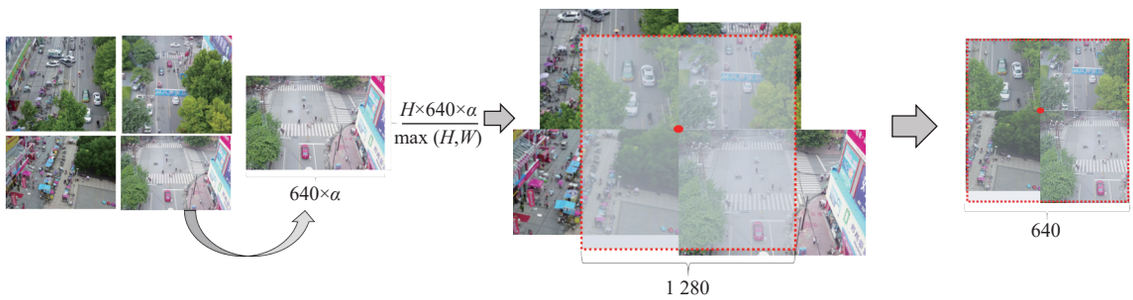


图 3　α-Mosaic实现过程

Fig.3　α-Mosaic Implementation process

$$\gamma = \frac{\omega_{\text{img\_size}} \times \alpha}{\max(h, w)} \tag{1}$$

$$h',\ w' = h \times \gamma,\ w \times \gamma \tag{2}$$

$$\boldsymbol{T}^{h \times w} = \text{Resize}_{h \times w}(\text{Mosaic}^{4}_{h^m \times w^m}(\boldsymbol{T}^{h' \times w'}_{\text{img}})) \tag{3}$$

式中：γ为对原始图像的尺寸调整系数，h'、w'为调整之后的图像高宽，max为取最大值函数，$\text{Resize}_{h \times w}$为裁剪到指定尺寸的函数，$\text{Mosaic}^{4}_{h^m \times w^m}$为Mosaic中对4张图像数据的拼接处理。

为证明α-Mosaic的改进有效性，本文在VisDrone2019-DET目标检测数据集上，以YOLOv5s为基线算法、640×640为输入特征尺寸、mAP@0.5、mAP@0.5:0.95和召回率为评价指标，通过改变α值设置了多组实验。以下选取6组在VisDrone2019-DET目标检测数据集上的实验结果进行分析探究，具体数据如表1和图4所示。

表1 不同α值的影响

Table 1 Influence of different α values

序号	α	mAP@0.5/%	mAP@0.5:0.95/%
1	1	31.3	16.1
2	1.187 5	31.7	16.3
3	1.312 5	32.5	17
4	1.406	32.4	16.9
5	1.5	33.1	17.4
6	1.594	32.6	17

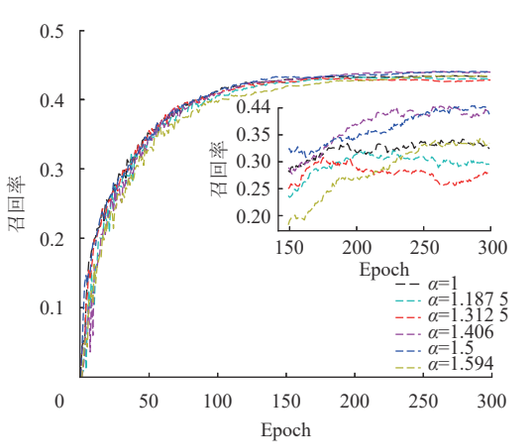


图4 不同α值对网络召回率的影响

Fig.4 Influence of differentαvalues on network recall

实验结果表明当α大小处在1.3至1.5之间时，可使精度达到最优值。从表1中数据可以看到，随着α值的增大，网络的检测精度在逐渐上升，说明α的存在可以有效提升检测网络对于小目标特征的学习能力。但当α值过大时，网络的精度反而会下降，这是因为过大的α值会使输入网络的有效特征信息损失太多，进而影响到了网络的特征提取能力。从图4中可以看出α的存在对网络召回率的影响，可见适当地增大α的值，可以提高网络的召回率，从而降低网络的漏检率。

为验证α-Mosaic在不同算法模型上的有效性，以下设置6组实验，固定α值为1.5，以mAP@0.5为衡量指标，分别在YOLOv5s、YOLOv4-Tiny、YOLOv3-Tiny 3个轻量化的模型输入端使用Mosaic和本文所提的α-Mosaic数据增强方法作对比，实验结果如图5所示。从图5中可以看出，在所有模型中，使用本文所提α-Mosaic数据增强方法的检测精度都要优于使用Mosaic数据增强方法的检测精度，这进一步证明了α-Mosaic数据增强方法的优越性。

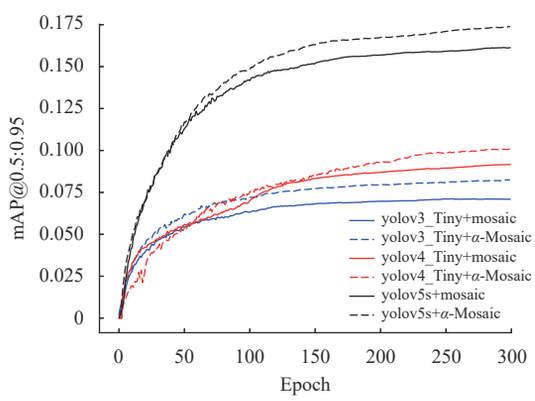


图5 α-Mosaic在不同模型上的表现

Fig.5 α-Mosaic performance on different models

1.2 轻量化和多尺度改进

由于无人机这类移动端设备的算力普遍不高，将复杂的目标检测模型部署到这类设备上时，常常难以获得理想的实时性推理运算效果。因此有必要针对无人机设备的硬件特性和应用场景，对目标检测算法进行优化，以提高无人机目标检测的效率和实时性。基于此，本文提出利用ShuffleNet[16]轻量化骨干网络中的部分模块替换原YOLOv5中的Backbone主干网络和特征度融合网络中的部分结构，以降低网络计算成本、提高模型的推理速。

为了减小网络轻量化导致的检测精度下降、适应性变差的问题，本文在原有的检测尺度上增加了一个微小物体检测头，以提高网络对小物体的检测精度和鲁棒性，使算法能够更好地适应不同尺度的目标检测需求。改进后的网络称之为LM-YOLO (Lightweight Multiscale)，结构如图6所示。

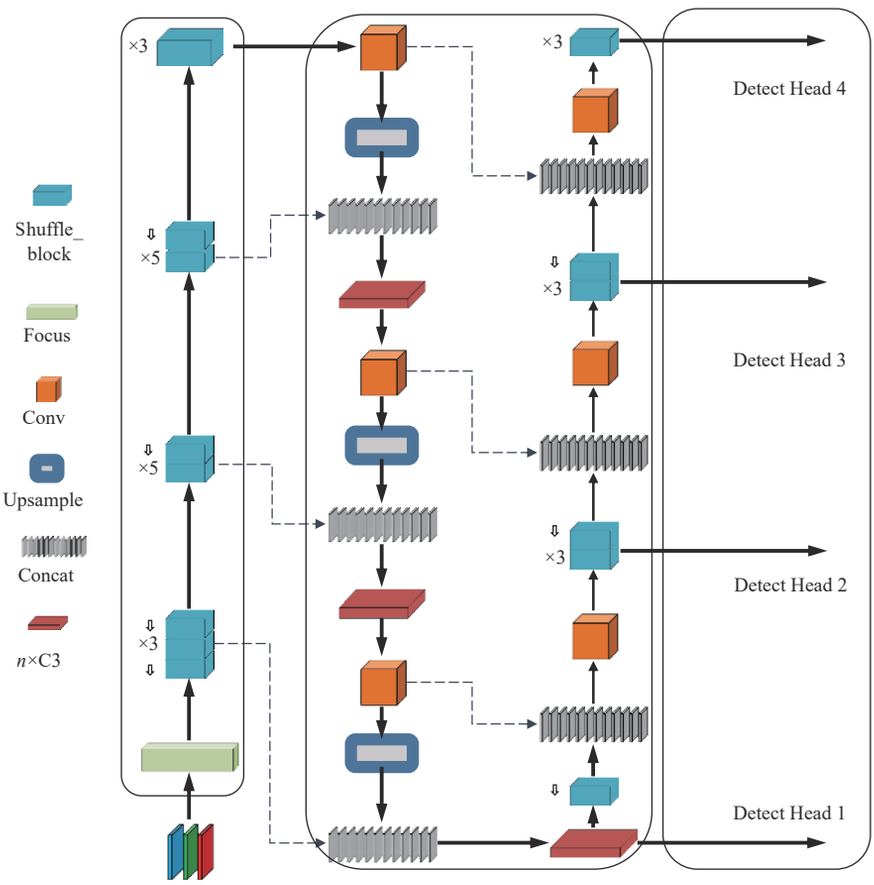


图 6　LM-YOLO网络结构

Fig.6　LM-YOLO network structure

1.3　并行卷积注意力模块(R-CBAM)

使用无人机进行航拍目标检测时，由于航拍图像数据中目标的尺寸较小且背景复杂，导致对小目标物体难以准确检测，因此提升神经网络的检测性能是至关重要的。本文从优化卷积特征提取能力的角度出发，通过改进CBAM[17](Convolutional Block Attention Module)设计了一种并行结构的空间通道注意力机制R-CBAM(Repeat Convolutional Block Attention Module)，以提升卷积对小目标的特征提取能力，其结构如图7所示。

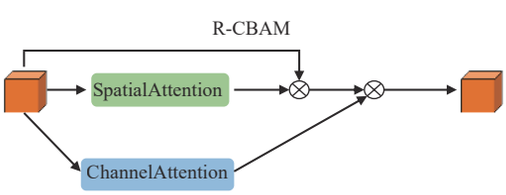


图 7　R-CBAM结构

Fig.7　R-CBAM structure

在CBAM模型中，通道注意力和空间注意力以串行方式连接，即通道域的输出作为空间域的输入特征。这种结构对于空间注意力模块来说存在一定的问题，因为在通道域注意力中每个卷积通道都会被叠加一个权重，从而改变了原始输入特征的空间细化信息，而卷积网络对于空间特征的变化非常敏感，叠加的权重可能会引起网络的过拟合，最终导致网络的性能下降。基于此，本文对CBAM的结构做了改进，将通道注意力中的全连接层替换为1×1卷积，并将通道域和空间域注意力模块调整到并行状态。改进之后，原始输入特征 $\boldsymbol{X} \in \boldsymbol{T}^{B\times C\times W\times H}$ (B, C, W, H 分别为卷积特征的批次、通道数、宽、高) 将分别作为两个维度输入空间通道注意力模块，首先，空间域提取到的特征信息与原始输入特征进行相乘得到空间域混合信息，接着将空间域混合信息与通道域的输出特征进行相乘得到完整的空间通道混合注意力特征信息。这样，两个维度的特征信息不会相互干扰，不同的作用域可以充分利用原始输入特征的细化信息，可以有效避免因权重叠加带来的过拟合效应。

定义以 $\boldsymbol{F}_{\mathrm{sa}}(\boldsymbol{X})$ 表示空间域输出特征，$\boldsymbol{F}_{\mathrm{ca}}(\boldsymbol{X})$ 表示通道域输出特征，$\boldsymbol{F}^{1}_{\mathrm{MLP}}(\boldsymbol{X})$、$\boldsymbol{F}^{2}_{\mathrm{MLP}}(\boldsymbol{X})$ 分别表示2个全连接神经网络层的输出特征，$\boldsymbol{F}_{\mathrm{R-CBAM}}(\boldsymbol{X})$ 表示整个注意力模块的输出特征，则以 $\boldsymbol{X} \in \boldsymbol{T}^{B\times C\times W\times H}$ 为输入特征

时，R-CBAM的推理过程如式(4)~(8)所示。

$$\boldsymbol{F}_{\mathrm{sa}}(\boldsymbol{X})=\sigma(\mathrm{Conv}(\mathrm{Cat}(\mathrm{Mean}(\boldsymbol{X}),\mathrm{Max}(\boldsymbol{X})))) \tag{4}$$

$$\boldsymbol{F}^{1}_{\mathrm{MLP}}(\boldsymbol{X})=\mathrm{Conv}(\mathrm{ReLu}(\mathrm{Conv}(\mathrm{AvgPool}(\boldsymbol{X})))) \tag{5}$$

$$\boldsymbol{F}^{2}_{\mathrm{MLP}}(\boldsymbol{X})=\mathrm{Conv}(\mathrm{ReLu}(\mathrm{Conv}(\mathrm{MaxPool}(\boldsymbol{X})))) \tag{6}$$

$$\boldsymbol{F}_{\mathrm{ca}}(\boldsymbol{X})=\sigma(\boldsymbol{F}^{1}_{\mathrm{MLP}}(\boldsymbol{X})+\boldsymbol{F}^{2}_{\mathrm{MLP}}(\boldsymbol{X})) \tag{7}$$

$$\boldsymbol{F}_{\mathrm{R-CBAM}}(\boldsymbol{X})=(\boldsymbol{F}_{\mathrm{ca}}(\boldsymbol{X})\times\boldsymbol{X})\times\boldsymbol{F}_{\mathrm{sa}}(\boldsymbol{X}) \tag{8}$$

式中：σ为Sigmoid函数，Conv为核为7×7的卷积模块，Mean为均值函数，ReLu为Rectified Linear Unit激活函数，Max为取最大值函数，AvgPool和MaxPool分别为均值池化和最大值池化函数。

为证明R-CBAM的有效性，在VisDrone2019 - DET数据集上以相同的训练参数，设置了5组实验。由于本研究扩充了YOLOv5的检测尺度，为保证实验统一性，以下均是在LM-YOLO的基础上，通过在Head网络结构部分添加不同的注意力模块来获取的实验结果，使用mAP@0.5和mAP@0.5:0.95作为各个模型的性能评价标准，具体实验数据如表2所示。

表 2 不同注意力模块性能对比

Table 2 Performance comparison of different attention modules

模型(α-mosaic)	mAP@0.5	mAP@0.5:0.95
LM-YOLO	33.5%	18.0%
+ SENet	34.3%↑	18.4%↑
+ ECA	32.5%↓	17.6%↓
+ CBAM	33.6%↑	17.8%↓
+ R-CBAM(our)	**34.4%↑**	**18.4%↑**

由表2中的数据可以看出，相比较于LM-YOLO基线算法，在分别加入SENet、ECA、CBAM、R-CBAM这4种注意力结构时，本文所提注意力模块性能指标表现出了最优的结果，其中mAP@0.5、mAP@0.5:0.95指标相较于基线分别提高了0.9个百分点和0.4个百分点，证明本文所提R-CBAM模型是一种有效的改进方法。

1.4 ShuffleViT全局注意力模块

Transformer[18]源于2017年谷歌在NLP方向提出的一种适用于文本分析的算法，该模型不同于传统大量使用卷积模块的神经网络，而是采用一种独特的基于自注意力的编解码机制，该编码机制可以高效地获取全局信息。2020年，DOSOVITSKIY A等[19]借鉴Tansformer设计思想，提出了一种适用于视觉领域的神经网络模型ViT(Vision Transformer)，该模型将Transformer中的注意力机制与卷积模块结合，在多个视觉任务上取得了最优的效果。但是ViT也存在着一些问题，比如不具备空间归纳偏差的能力、难以训练以及参数庞大的问题。针对这些问题苹果公司[20]在2021年提出了一种适用于移动设备的改进模型MobileViT(Light-Weight, General-Purpose, and Mobile-Friendly Vision Transformer)，该模型使用了一种新的切片方式，将卷积获取局部特征的优点和ViT可以高效获取全局信息的特点相结合，不但优化了网络结构，而且使网络具备了不丢失切片位置信息的能力，但MobileViT仍然存在着较高的训练成本和推理速度不理想的问题。

从进一步提升无人机航拍小目标检测网络的全局信息提取能力的角度出发，本文利用ShuffleNet中的“混洗”机制和MobileViT中的切片机制对ViT中的Transformer结构进行了改进，设计了一种新的全局信息提取模块ShuffleViT。该模块通过将输入特征进行切片拆分，再进行块内信息打乱，在保留块序列信息不变的情况下，实现块内的局部特征信息交互的同时降低了计算量，从而提升网络的全局信息提取能力并且降低了网络的训练成本，其结构如图8所示。

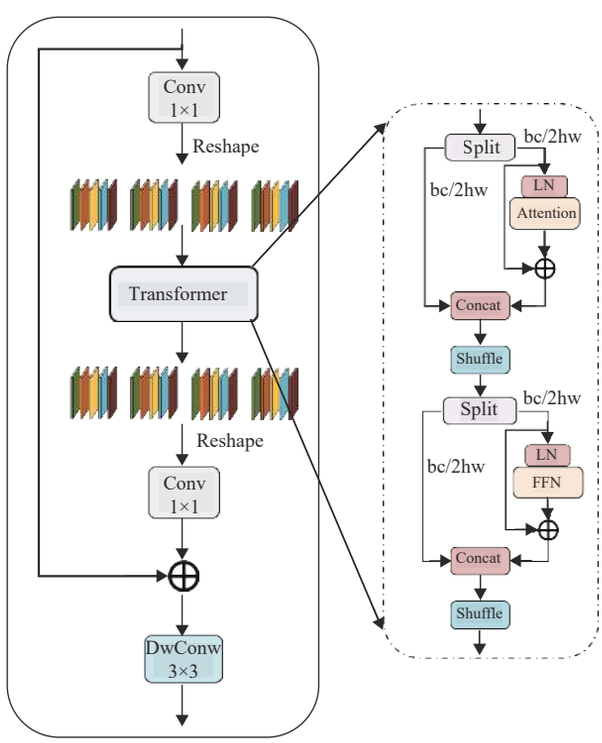


图 8 ShuffleViT结构

Fig.8 ShuffleViT structure

与ViT模型不同的是，ShuffleViT在数据处理上直接将输入特征$\boldsymbol{X}\in\boldsymbol{T}^{B\times C\times W\times H}$ (B，C，W，H分别为卷积特征的批次、通道数、宽、高) 通过一个1×1分组

卷积投影到固定D维度$\boldsymbol{X}\in\boldsymbol{T}^{B\times D\times W\times H}$，然后将特征分割为$P^h\times P^w$ ($P^h,P^w\leqslant H,W$；h，w分别为分块特征的高和宽) 大小的块，处理之后的特征为$\boldsymbol{X}_p\in\boldsymbol{T}^{B\times P^h\times P^w\times\frac{HW}{P^hP^w}\times D}$。接着，将特征$\boldsymbol{X}_p$按第2个维度分块处理为两部分：$\boldsymbol{X}_{\text{ps1}},\boldsymbol{X}_{\text{ps2}}\in\boldsymbol{T}^{B\times\frac{P^hP^w}{2}\times\frac{HW}{P^hP^w}\times D}$，其中$\boldsymbol{X}_{\text{ps2}}$作为自注意力模块的输入，$\boldsymbol{X}_{\text{ps1}}$作为残差连接与$\boldsymbol{X}_{\text{ps2}}$按第2个维度进行拼接，这里进行切片处理的操作是在每个切片内部进行的，并不会打乱每个块之间的位置信息。然后，利用“混洗”机制对拼接特征在第2个维度进行“打乱”处理，实现块内信息交互，值得注意的是，这里进行的打乱是在切片内部进行的像素打乱，不会改变原有的块之间的位置信息。通过这种改进方式，ShuffleViT的计算量大幅下降，不但能进行全局信息提取，而且不会丢失切片的位置信息。

定义$\boldsymbol{X}_{\text{ps1}}$、$\boldsymbol{X}_{\text{ps2}}$表示分片之后的特征，$\boldsymbol{F}_{\text{attn}}$为多头注意力机制的输出，$\boldsymbol{F}_{\text{ffn}}$为全连接解码层的输出，则对于输入特征$\boldsymbol{X}\in\boldsymbol{T}^{B\times C\times W\times H}$，ShuffleViT的输出计算推导过程如式(9)~(14)所示。

$$\boldsymbol{X}_{\text{ps1}},\boldsymbol{X}_{\text{ps2}}=\text{Split}(\boldsymbol{X}) \tag{9}$$

$$\boldsymbol{F}_{\text{attn}}\left(\boldsymbol{X}_{\text{ps1}},\boldsymbol{X}_{\text{ps2}}\right)=\text{Cat}(\text{Attn}(\text{LN}(\boldsymbol{X}_{\text{ps2}}))\boldsymbol{X}_{\text{ps2}},\boldsymbol{X}_{\text{ps1}}) \tag{10}$$

$$\boldsymbol{X}_{\text{shuffle}}=\text{Shuffle}(\boldsymbol{F}_{\text{attn}}) \tag{11}$$

$$\boldsymbol{X}'_{\text{ps1}},\boldsymbol{X}'_{\text{ps2}}=\text{Split}(\boldsymbol{X}_{\text{shuffle}}) \tag{12}$$

$$\boldsymbol{F}_{\text{ffn}}(\boldsymbol{X}'_{\text{ps1}},\boldsymbol{X}'_{\text{ps2}})=$$

$$\text{Cat}(\text{FFN}(\text{LN}(\boldsymbol{X}'_{\text{ps2}}))+\boldsymbol{X}'_{\text{ps2}},\boldsymbol{X}'_{\text{ps1}}) \tag{13}$$

$$\boldsymbol{X}_{\text{stransformer}}=\text{Shuffle}(\boldsymbol{F}_{\text{ffn}}) \tag{14}$$

式中：LN为LayerNorm归一化方法，Attn为多头注意力模块，Cat为张量拼接函数，Split为分片操作，Shuffle为“混洗”机制中的打乱操作。

2　实验结果及分析

2.1　训练参数及实验环境设置

改进过程中每个实验的初始学习率均设置为0.01，BatchSize为64，动量大小设置为0.937，权重延迟大小为0.000 5，预热训练迭代3次，预热期间动量大小为0.8，输入特征大小均为640×640。采用α-Mosai、Mosaic数据增强方法和自适应Anchor，总迭代次数为300个Epoch。

实验过程所使用的环境及软硬件配置参数如表3所示。

表 3　实验环境

Table 3　Experimental environment

软硬件配置	参数
操作系统	Ubuntu 20.04.1
CPU	Intel® Xeon® Gold 5218 CPU@2.3 Hz
GPU	TITIAN RTX(24 GB)
内存	252 GB
编程语言	Python 3.8
深度学习框架	PyTorch 1.7，CUDA 11.0

2.2　数据集

为了在相同的条件下衡量不同算法的性能，本节所有实验均采用VisDrone2019-DET数据集。该数据集是ICCV2019 VisDrone挑战赛发布的数据集，包含pedestrain、people、bicycle、car、van、truck、tr -icycle、awning-tricycle、bus、motor共10个类别，总共8 629张图像，其中训练集6 471张，验证集548张，测试集1 610张，所有图像均采自中国14个不同的城市。

2.3　评价指标

为评估不同模型算法的综合性能，以下实验采用模型体积(单位：MB)、参数量、浮点运算量(Giga Floating-point Operations Per Second，GFLO -PS)、不同阈值范围的平均值均值精度(mean Average Precision, mAP@0.5和mAP@0.5:0.95)以及每秒处理帧数(Frames Per Second，FPS)衡量每个模型的性能效果。

2.4　实验结果对比与分析

2.4.1　消融实验及分析

为验证本文提出的α-Mosaic数据增强方法、ShuffleViT全局信息提取模块和R-CBAM并行空间通道注意力机制的有效性，在YOLOv5(4.0)的结构基础之上，通过将不同模块依次与原网络进行融合，设计了6消融实验，每组实验均设置相同的训练参数和相同的实验环境。实验结果见表4所示。

其中“√”表示添加该模块，S-M表示本文针对于α-Mosaic数据增强算法所提出的改进模块，S-B表示经过轻量化处理之后的网络，D-H表示在S-B网络基础之上添加微小物体检测头(Detect Head)，R-C表示本文基于CBAM模块所改进的空间通道注意力模块R-CBAM，S-ViT表示本文所提出的轻量化全局注意力模块ShuffleViT。改进1、改进2、改进3、改进4、改进5分别为每次对算法做出改进后的代称。

从表4可以看出，相较于基线算法YOLOv5s，改进1在替换使用本文提出的α-Mosaic数据增强方法

表 4 消融实验

Table 4 Ablation experiment

模型	S-M	S-B	D-H	R-C	S-ViT	体积/MB	参数量/M	计算量/GFLOPS	mAP@0.5/%	mAP@0.5:0.95/%	FPS
YOLOv5s						14.4	7.08	16.6	31.3	16.1	137
改进1	√					14.4	7.08	16.6	32.5	17.0	137
改进2		√	√			5.8	2.63	10.2	33.5	18.0	96
改进3	√	√	√			5.8	2.63	10.2	34.6	18.3	96
改进4	√	√	√	√		5.8	2.63	10.2	35.5	19.0	93
改进5	√	√	√	√	√	8.1	3.68	12.2	36.4	19.6	82

后，在网络体积、参数量、计算量不变的情况下，mAP@0.5提高了1.2个百分点，mAP@0.5:0.95提高了0.9个百分点，这说明该方法有效改善了Mosaic对于小目标增强恶化情况，提升了输入网络中的有效特征。改进2在只经过轻量化处理和引入微小物体检测头时，网络的计算量下降38.55%、参数量下降62.8%、模型体积下降59.7%的同时，mAP@0.5提升了2.2个百分点，mAP@0.5:0.95提升了1.9个百分点，证明增加微小物体检测头则可以在网络“减重”的同时保证精度的提升。改进3在改进2的基础之上将原有的Mosaic数据增强方法替换为α-Mosaic方法，相比较于改进2模型，改进3在网络体积、参数量、计算量不变的情况下，mAP@0.5提高了1.1个百分点，mAP@0.5:0.95提高了0.3个百分点，进一步证明了α-Mosaic数据增方法的有效性。改进4是在改进3的基础之上引入并行卷积注意力模块(R-CBAM)，在几乎不改变网络的参数量和计算量的情况下mAP@0.5提高了0.9个百分点，mAP@0.5:0.95提高了0.7个百分点，证明R-CBAM可以通过提升卷积通道之间的竞争合作关系和对局部重要特征的关注能力进而提升网络的检测精度。改进5是在改进4的基础之上引入本文提出的ShuffleViT全局注意力模块，该模块的作用主要是加强网络对于“难检测目标”(尺寸过小、图像不清晰)的检测能力，相较于改进4，改进5虽然在模型复杂度上有所提升，但mAP@0.5提高了0.9个百分点，mAP@0.5: 0.95提高了0.6个百分点，这证明ShuffleViT可以更好地提高模型的性能和鲁棒性。

2.4.2 对比实验及分析

为进一步验证所提算法相比同类型算法的优越性，将本文算法与当前一些主流的目标检测算法在VisDrone2019-DET数据集上进行实验对比，实验结果如表5所示。

表 5 对比实验

Table 5 Comparison experiment

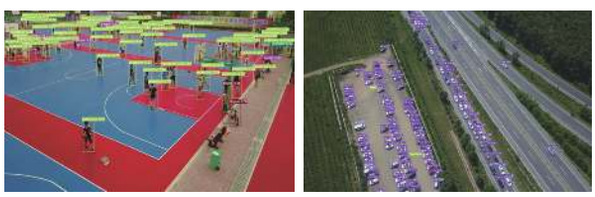
模型	输入尺寸	体积/MB	参数量/M	计算量/GFLOPS	mAP@0.5/%	mAP@0.5:0.95/%	FPS
YOLOv5s(4.0)	640×640	14.4	7.08	16.6	31.3	16.1	117
YOLOv4-Tiny	640×640	46.33	6.02	16.5	17	9.2	168
YOLOv3-Tiny	640×640	17.5	8.69	13.0	16.4	7.09	156
SSD(VGG-16)	512×512	197	60.30	23.0	26.6	9.33	37
Faster-RCN(ResNet50)	640×640	316	38.2	39.3	29.7	16.3	9
RetinaNet (ResNet50)	640×640	140	39	211	13.9	6.86	31
CenterNet	640×640	14.1	11.26	20.6	26.2	17.9	27
Our	**640×640**	**8.1**	**3.68**	**12.2**	**36.4**	**19.6**	**82**

从表5中的数据可以看出，相比较于其他典型算法，本文算法在保证较快的推理速度的前提下，其精度和模型复杂度表现出明显的优越性，这说明本文算法在实际应用中能够更快速地进行推理运算，并且具有更高的预测准确性。与轻量化网络YOLOv3-Tiny、YOLOv4-Tiny相比，本文所提算法虽然在检测速度上稍慢于两者，但在其他几项指标上均大幅度领先。在mAP@0.5和mAP@0.5:0.95两项指标上，本文算法分别比YOLOv3-Tiny、YOLOv4-Tiny高19.4、20和10.4、12.51个百分点。相比较于基线算法，本文算法参数量下降了48.0%，计算量下降了26.5%，模型体积下降43.8%，而mAP@0.5提高了4.9个百分点，mAP@ 0.5:0.95提高了3.3个百分点。综合来看，本文提出的算法在α-Mosaic数据增强方法、ShuffleViT和R-CBAM注意力机制的加持下，相比较于其他典型目标检测算法，不仅能够有效地提高无人机航拍小

目标检测的分类和定位能力，还能够保证较快的推理速度，这表明本文算法比其他算法更具性能优势。

2.5　算法有效性分析

为了更加直观地对本文算法的检测效果进行评估，在VisDrone2019-DET测试集中选取几组复杂真实场景图像进行测试，部分测试图像检测结果如图9所示。从4个检测样本图像可以看出，即使图像背景复杂、目标分布不均匀且高度密集、光线不充足，本文方法在各种不同的复杂场景中依然展现出了较高的检测性能。



(a) 样例 1　　(b) 样例 2

(c) 样例 3　　(d) 样例 4

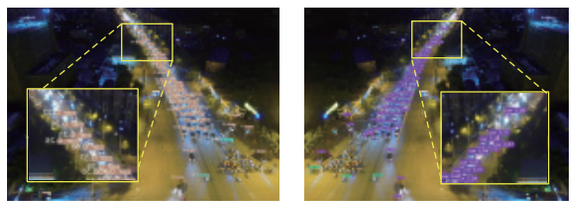
图 9　本文算法在测试集上的检测效果

Fig.9　The detection performance of the algorithm in this article on the test set

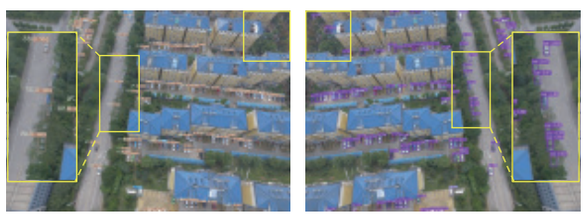
为验证本文算改进算法相较于基线算法(YOLOv5s)在无人机航拍视角下的检测优化提升效果，选取VisDrone2019-DET测试集中两组不同背景条件的小目标图像进行检测并进行可视化对比分析。如图10所示，其中左侧图像为基线YOLOv5s算法的检测推理结果，右侧为本文改进算法的推理检测结果。由检测结果可知，相较于基线算法，本文改进算法能够更加精准地识别出遮挡性目标和密集型小目标，并且能够避免漏检。这表明，本文的改进方法是有效的，并且在处理相关无人机航拍目标图像时表现出比基线算法更好的检测性能。

3　结论

本文提出一种基于改进YOLOv5的多尺度轻量化无人机航拍目标检测网络。为了提高络的输入有效特征，提出了一种改进Mosaic的数据增强方法，通过放大经Mosaic处理的输入图像数据，使得小目标、极小目标的有效输入特征信息增多。为了进一步提



(a) 样例 1 (YOLOv5s)　　(b) 样例 2 (GA-YOLO)



(c) 样例 3 (YOLOv5s)　　(d) 样例 4 (GA-YOLO)

图 10　YOLOv5s与本文算法对比

Fig.10　Comparison between YOLOv5s and this algorithm

高网络的检测能力，提出了基于“混洗”机制和ViT结合的改进注意力模块和基于CBAM的改进注意力模块，前者在保留分块序列位置信息不丢失的情况下，实现了块内信息交互，同时降低了计算量，提高了网络对于全局信息的提取能力；后者提升了网络不同通道之间的竞争合作关系和对于局部重要特征的关注程度。消融实验和对比实验证明了本文方法在检测精度、计算复杂度和体积大小上的优越性，但因本文采用的是四尺度检测，这在一定程度上减缓了网络的推理速度，因此，在未来的工作中，如何既保持网络的精度又保持更快的推理速度是一个重要的研究方向。

参考文献：

[1] 曹家乐, 李亚利, 孙汉卿, 等. 基于深度学习的视觉目标检测技术综述[J]. 中国图象图形学报, 2022, 27(6): 1697-1722.
CAO J L, LI Y L, SUN H Q, *et al*. A survey on deep learning based visual object detection [J]. China Journal of Image and Graphics, 2022, 27(6): 1697-1722.

[2] 戴文君, 常天庆, 张雷, 等. 图像目标检测技术在坦克火控系统中的应用[J]. 火力与指挥控制, 2020, 45(7): 147-152.
DAI W J, CHANG T Q, ZHANG L, *et al*. Application of image target detection technology in tank fire control system [J]. Fire and Command Control, 2020, 45(7): 147-152.

[3] LIO W, ANGUELOV D, ERHAN D, *et al*. SSD: single shot multibox detector[C] //Computer Vision–ECCV 2016: 14th European Conference. Amsterdam, Netherlands: Springer International Publishing, 2016: 21-37.

[4] ZHAI S, SHANG D, WANG S, *et al*. DF-SSD: an improved SSD object detection algorithm based on DenseNet

and feature fusion [J]. IEEE Access, 2020, 8: 24344-24357.

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

[6] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR) . Hawaii: IEEE, 2017: 7263-7271.

[7] REDMON J, FARHADI A. Yolov3: An incremental improvement[EB/OL]. arXiv: 1804.02767 (2018-04-08) [2023-02-07]. https://arxiv.53yu.com/abs/1804.02767.

[8] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. Yolov4: optimal speed and accuracy of object detection[EB/OL]. arXiv: 2004.10934 (2020-04-22) [2023-02-07]. https://arxiv.org/abs/2004.10934.

[9] LIN T Y, GOYAL P, GIRSHICK R, *et al*. Focal loss for dense object detection[C] //Proceedings of the IEEE International Conference on Computer Vision. Hong Kong: IEEE, 2017: 2980-2988.

[10] GIRSHICK R. Fast R-CNN[C] //Proceedings of the IEEE International Conference on Computer Vision. Santiago Chile: IEEE, 2015: 1440-1448.

[11] REN S, HE K, 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, 2016, 39(6): 1137-1149.

[12] PURKAIT P, ZHAO C, ZACH C. SPP-Net: deep absolute pose regression with synthetic views[EB/OL]. arXiv: 1712.03452 (2017-12-09) [2023-02-09]. https://arxiv.53yu.com/abs/1712.03452.

[13] LI P, CHE C. SeMo-YOLO: a multiscale object detection network in satellite remote sensing images[C] //2021 International Joint Conference on Neural Networks (IJCNN) . Shenzhen: IEEE, 2021: 1-8.

[14] TAN L, LV X, LIAN X, *et al*. YOLOv4_Drone: UAV image target detection based on an improved YOLOv4 algorithm [J]. Computers & Electrical Engineering, 2021, 93: 107261.

[15] WANG M, LI Q, GU Y, *et al*. SCAF-net: Scene context attention-based fusion network for vehicle detection in aerial imagery [J]. IEEE Geoscience and Remote Sensing Letters, 2021, 19: 1-5.

[16] ZHANG X, ZHOU X, LIN M, *et al*. Shufflenet: an extremely efficient convolutional neural network for mobile devices[C] //Proceedings of the IEEE Eonference on Computer Vision and Pattern Recognition. Wellington New Zealand: IEEE, 2018: 6848-6856.

[17] WOO S, PARK J, LEE J Y, *et al*. CBAM: convolutional block attention module[C] //Proceedings of the European Conference on Computer Vision (ECCV) . Munich: EACV, 2018: 3-19.

[18] GUO M H, LU C Z, LIU Z N, *et al*. Visual attention network [J]. Computational Visual Media, 2023, 9(4): 733-752.

[19] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, *et al*. An image is worth 16x16 words: transformers for image recognition at scale[EB/OL]. arXiv: 2010.11929 (2021-06-03) [2023-02-11]. https://arxiv.53yu.com/abs/2010.11929.

[20] MEHTA S, RASTEGARI M. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer [EB/OL]. arXiv: 2110.02178 (2022-03-04) [2023-02-11]. https://arxiv.53yu.com/abs/2110.02178.

（责任编辑：杨耀辉　英文审核：熊荣斌）