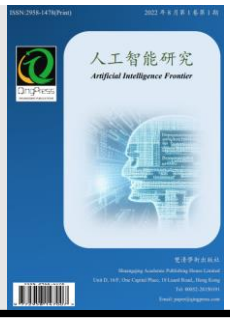
人工智能研究
Artificial Intelligence Frontier
ISSN: 2958-1478 (print)
雙清學術出版社
Contents lists available at www.qingpress.com
Journal homepage: qingpress.com/zh-cn/journals/45



基于注意力机制和 YOLOX 的火焰烟雾检测算法

朱博超[1]，徐照[2*]
(1. 东南大学软件学院（苏州)，江苏 苏州 215000；
2. 东南大学土木工程学院，江苏 南京 210000)

摘要：火灾是严重威胁人类生命安全与造成巨大财产损失的主要灾害之一，对火焰和烟雾进行检测能有效预防火灾发生。针对现有基于机器视觉方法对火焰和烟雾进行检测的方法具有的准确率低、效率低等问题，本文提出一种基于使用多种注意力机制改进的 YOLOX 目标检测算法。Swin Transformer 是一种基于多头自注意力与滑动窗口的层级式深度神经网络。我们的模型基于注意力机制和 YOLOX，通过使用 Swin Transformer 网络作为目标检测的主干网络，结合空间注意力与通道注意力机制，添加模糊损失，让模型具有感知全局特征能力。实验结果表明，在相同数据集上，基于注意力机制改进的 YOLOX 相对未修改的 YOLOX 目标检测评价指标 mAP 提高了 5. 75%，火焰与烟雾检测的准确度获得了极大提升。

关键词：火焰, 烟雾, 注意力机制, 目标检测

Fire and Smoke Detection Algorithm Based on an Improved YOLOX with Attention

Bo-chao ZHU[1], Zhao XU[2†]
(1. Department of Software Engineering, Southeast University, Suzhou Jiangsu 215000, China;
2. Department of Civil Engineering, Southeast University, Nanjing Jiangsu 210000, China)

基金项目：教育部人文社科基金(20YJAZH114)；江苏省自然科学基金(BK20201280);国家自然科学基金(72071043)
作者简介：朱博超，男，主要研究方向：计算机视觉
*通讯作者信息：徐照， 研究方向:BIM、工程管理等， Email: xuzhao@seu.edu.cn

Article history: Received April 21, 2023 Accepted May 5,, 2023 Available online May 11, 2023
To cite this paper: 朱博超，徐照 (2023). 基于注意力机制和YOLOX 的火焰烟雾检测算法.人工智能研究, 第1 卷,第2 期, 16-25.
Doi: https://doi.org/10.55375/aif.2023.2.2

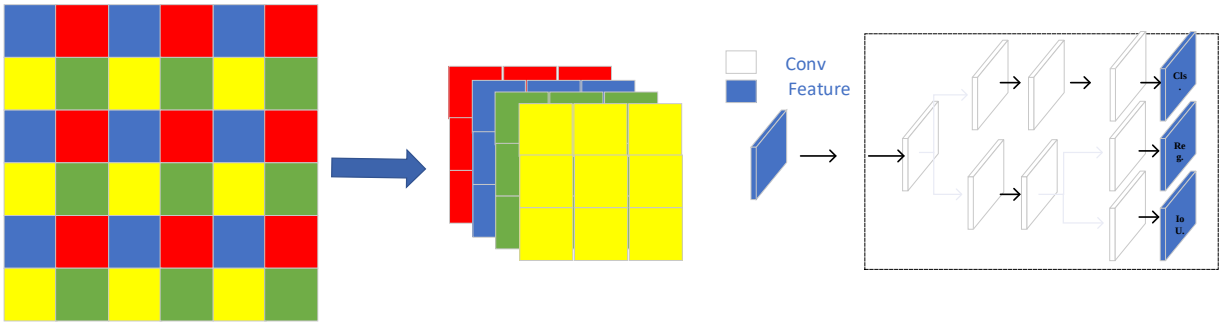
Abstract: Fire disaster is one of the major disasters that seriously threaten human life's safety and cause huge property losses. Detecting the flame and smoke can effectively prevent fire disasters. In response to the low accuracy and efficiency of existing machine vision-based methods for detecting flames and smoke, this article proposes a YOLOX object detection algorithm using multiple attention mechanisms for improvement. Swin Transformer is a hierarchy deep neural network with multi-head self-attention and shift window. Our model is based on the attention mechanism and YOLOX by using Swin Transformer as the backbone of object detection, combining spatial attention and channel attention, adding blur loss so that the model can perceive global features. The experiment results revealed that, on our dataset, the improved YOLOX is 5.75% better than the original one in mAP, and the accuracy of fire and smoke detection is greatly improved.

Keywords: Fire, Smoke, Attention, Object Detection

火灾是一种较为常见的灾害，一旦发生火灾，不仅会造成严重的人员伤亡，而且会产生巨大的财产损失。传统的火灾检测方法主要依赖于各种传感器对检测区域的温度、光谱和烟粒浓度进行探测，以此确定是否发生火灾。然而，火灾会引起烟雾浓度、温度、气流等其他信息的变化，导致该方法难以在室外宽阔场景或环境恶劣的场景中有效的工作。近年来，基于深度学习的图像分类、目标检测等计算机视觉技术被广泛运用于火灾检测上。卷积神经网络 (CNN) 是深度学习中最常用的一种网络，常常被用来对火灾图像进行特征提取。由 CNN 提取的特征常被用作分类和回归，获得检测结果以完成下游检测任务。吴凡[1]使用 DenseNet-121[2] 作为 YOLOv3[4] 的主干网络以达到增强提取火焰和烟雾特征能力的目的。张为[3]等提出通过嵌入空洞卷积模块改进 YOLOv3 的火灾检测方法，通过在 DarkNet-53 中使用空洞卷积来增大各层网络的感受野。王国睿[5]通过在 YOLOv4[6] 中引入多头注意力机制与层级式特征金字塔提高了检测精度。陆雅诺[7]基于 anchor-free 的目标检测算法通过引入有效通道注意力机制 (ECA) 提高了 CenterNet[8] 在无人机图像上的森林火灾检测精度。Muhammad[9]提出使用一种轻量级的网络 SqueezeNet 来进行火灾探测，该网络采用小卷积核替代全连接层，保持其他复杂模型类似的精度，具备更小的模型和更低的计算成本。喻丽春[10]通过改进自顶向下的特征金字塔和分割损失函数来优化 Mask RCNN[11]，提高了火焰检测与实例分割的准确度。

目前基于卷积神经网络的深度学习算法在处理视觉任务时具有准确性高、成本低、速度快等优势[20]，但相比之下，它们在处理视觉要素和物体之间的关系方面表现不如 Transformer。Transformer[12] 最早应用于自然语言处理领域，它的提出是为了解决循环神经网络模型难以并行训练，同时需要大量存储空间记忆完整的序列信息。Transformer 在自然语言处理领域的成功应用，使得相关学者开始探讨和尝试其在计算机视觉领域的应用。视觉 Transformer 模型的优势在于其构建了全局信息交互机制，有助于建立更为充分的特征表示，因此在图像分类、目标检测、图像分割、视频理解、图像生成以及点云分析等领域取得媲美甚至领先卷积神经网络的效果[13]。

本文提出将 Transformer 模块作为 YOLOX[14] 目标检测算法的主干网络，加入多种注意力机制，增强主干网络对全局火焰与烟雾特征的学习能力，提高模型对火焰初期形成的较小的火焰目标与稀疏的烟雾的检测能力。通过在自制数据集上的实验表明，本方法相较于原版 YOLOX 目标检测评价指标 (mAP) 提高了 5.75%。



(a)通道整合　　(b)解耦预测头

图 1　YOLOX 使用的算法

1　相关技术

1.1 YOLOX

YOLOX 是旷视科技在 2021 年发表的目标检测文章，综合了之前 YOLO 系列网络的优点如 YOLOv5 提出的 Focus 通道增广技术和 Mosaic 数据增强技术，YOLOv4 使用的 CSPDarknet 骨干网络， YOLOv1[15] 提出的 anchor−free 思想。该算法创新性地提出了解耦预测头 (decoupled head) 和 SimOTA 动态正样本匹配方法。解耦预测头如图 1 所示，一直以来，YOLO 工作都是仅使用一个分支就同时完成置信度、分类以及回归三部分的预测，Decoupled Head 使用两个并行分支去分别做分类和回归的预测，提高了检测的精度和模型收敛速度。Focus 模块将每个 2x2 的相邻像素划分为一个 patch，然后将每个 patch 中相同位置（同一颜色)像素给拼在一起就得到 4 个特征图，增大了输入数据的通道数且有效降低了计算复杂度。

1.2 Transformer

Transformer 是由谷歌在 2017 年提出的一种基于注意机制的深度神经网络。注意力机制（Attention Mechanism) 源于对人类视觉的研究。由于信息处理的瓶颈，人类会选择性地关注一部分信息，忽略其他信息。研究人员将 Transformer 应用于计算机视觉任务上且取得了不错的效果，甚至超越了卷积神经网络。Vision Transformer[16] 首次证明 Transformer 架构可以直接应用于图像领域，提出了将图像打散为一系列大小相同的子块作为 Transformer 的输入序列进行处理。DETR[17] 首先利用 Transformer 实现了一种端到端的目标检测方法，通过增加了一个 Transformer 编码器和解码器，避免了非极大值抑制过程，提高检测速度。Swin Transformer[18] 学习了 CNN 中常用的一些归纳偏置如平移不变性等先验知识，通过层级化的方式构建层次化 Transformer，提出了一种滑动窗口注意力机制，计算相邻窗口之间的图像存在的空间关系。

YOLOX 提出的多种优化方法有效的提高了神经网络在目标检测上的效果。注意力机制是改进深度卷积神经网络的一种有效方法。基于注意力机制的视觉 Transformer 能够在检测早期获得优于传统卷积神经网络的感受野。结合两者的优点有助于在火灾形成的早期检测到火焰与烟雾目标。

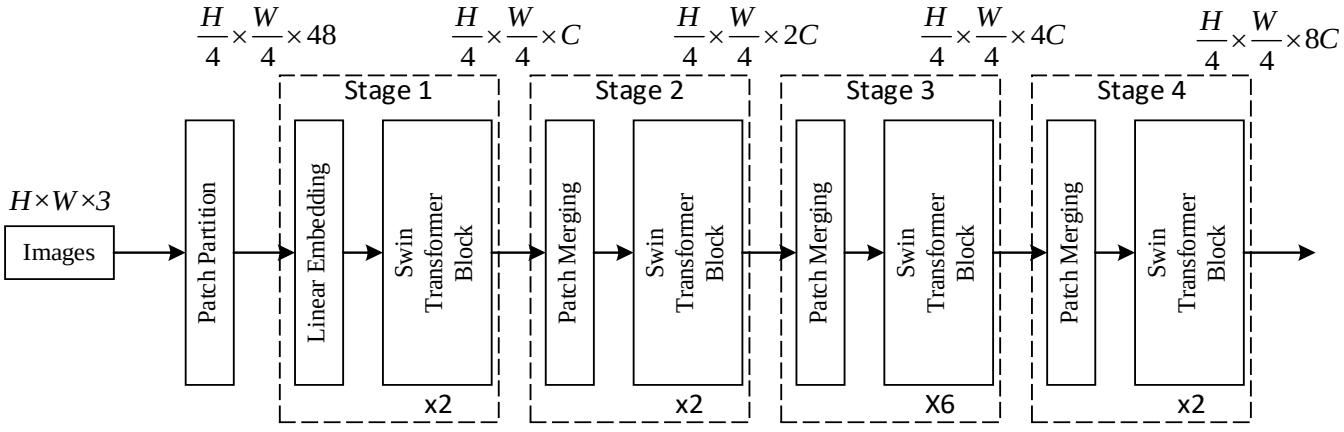


图 2　主干网络

其中，H 和 W 代表输出特征图尺寸， C 代表特征图通道数。

2　方法

2.1　数据集

实验使用的数据集为自制的火焰烟雾图像数据集。图像来源于网络，共 10827 张图像。其中火焰图像包含多种场景下的火灾图像如森林火灾、交通事故火灾、工厂火灾等等。此外还包括一些常见的火焰图像如烛火、煤气火焰等，以增强图像类别的多样性。

由于图像数量较多，对每张图像进行人工标注成本过高，因此选择使用预训练模型的方法。先使用 Labelimg 工具手动标注 1000 张图像，并以此训练出一个轻量的 YOLOv5s 检测模型。再使用该训练好的模型对所有图像进行二次标注，标注完成后再次对标注数据进行人工修改与补充。

2.2 主干网络

直接把注意力机制从自然语言处理领域直接应用到视觉领域存在一些挑战，这个挑战主要来自于两个方面：第一个挑战是尺度上的问题。一张图片里面有大小尺寸不同的物体，不同尺度的物体在卷积神经网络中可以通过感受野来解决，这种尺度上的不统一在自然语言处理中不存在。另一个挑战是图像的分辨率太大。如果以像素点作为输入的基本单位，对序列的长度进行注意力计算会导致数据量爆炸。为了解决这两个挑战，本文使用 Swin Transformer 作为主干网络。通过借鉴卷积神经网络的设计理念以及先验知识，构建了一种层级式的注意力机制，它的特征通过移动窗口的方式学习。

首先，为了减少序列的长度、降低计算复杂度，通过把一张图像拆散成大小相同的图像子块(Patch)，并用窗口选择固定数量的子块，在小窗口之内计算自注意力，整张图的计算复杂度与图片的大小呈线性关系，在一个窗口范围内计算自注意力的效果接近卷积神经网络。对于如何生成多尺寸的特征，Swin Transformer 使用通道像素整合（Patch merging)方法。通道像素整合把不相邻的小像素组合成一个大像素，增大感受野并且获得多尺寸的特征，输出多尺度特征。

主干网络的整体网络结构如图 2 所示。每个阶段可以输出尺寸与通道数不同的特征图，这种层次化结构与 ResNet[19]等典型的卷积神经网络相似。本文与原版 YOLOX-s 保持一致，超参数通道数 C 设置为 128。

2.3 多头自注意力

本文使用了多头自注意力层。图 3 显示了两个 Transformer 块。输入到第一个 Transformer 块的特征首先通过层次标准化(Layer Normalization)进行归一化，再经过窗口多头自注意力层(W-MSA)进行特征的学习，随后使用全连接层，期间使用残差网络保证网络模型不会出现梯度爆炸或梯度消失，这样完成第一层的输出特征的计算。第二次进入 Transformer 块与第一次类似，不同的是窗口多头自注意力层的计算特征部分需要进行一次滑动窗口(SW-MSA)的操作。数据通过注意力块的前向传播如下式所示。

$$\hat{z}^{l} = W-MSA(LN(\hat{z}^{l-1})) + \hat{z}^{l-1} \quad (1)$$

$$\hat{z}^{l} = MLP(LN(\hat{z}^{l})) + \hat{z}^{l} \quad (2)$$

$$\hat{z}^{l+1} = SW-MSA(LN(\hat{z}^{l})) + \hat{z}^{l} \quad (3)$$

$$\hat{z}^{l+1} = MLP(LN(\hat{z}^{l+1})) + \hat{z}^{l+1} \quad (4)$$

其中，W-MSA 代表窗口多头自注意力 ， Z 代表输入输出数据 ， LN 代表层归一化， MLP 代表全连接层。

自注意力可以让神经网络在提取特征图的迭代过程中，自发的注意需要被关注的像素，忽视背景像素。使用多个自注意力模块可以让模型学习到不同位置像素的权重，提高特征提取的能力。

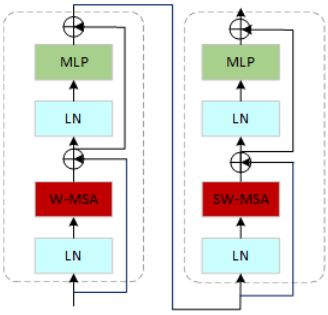


图 3　自注意力 Transformer 块

2.4　空间与通道自注意力机制

通道注意力模块通过网络计算出输入图像各个通道的重要性（权重)，关注信息较多的通道，忽视信息少的通道，从而达到提高特征表示能力。使用通道注意力机制可以充分利用输入的图像的通道信息，对通道信息特征进行校正，校正后的特征可保留有价值的特征，剔除没价值的特征。通道注意力机制如图所示。

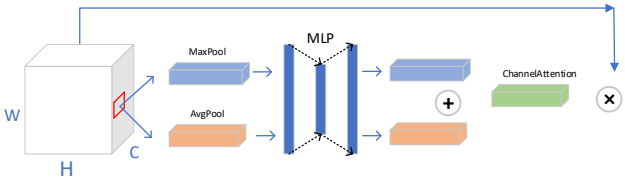


图 4 通道注意力机制

空间注意力机制与通道注意力机制类似，对于输入的图片，空间注意力机制更关注图像像素上的特征，对各种形变数据在空间中进行转换并自动捕获重要区域特征。该技术能够保证图像在经过裁剪、平移或者旋转等操作后，不影响最终产生的结果。通道注意力机制如图 5 所示。

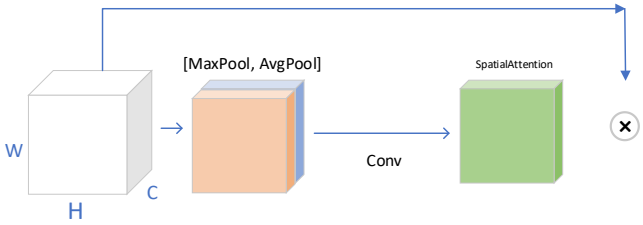


图 5 空间注意力机制

混合注意力机制则是将上述通道注意力和空间注意力可以通过串联、或者并联的方式进行组合。

在 Swin Transformer 主干网络中的 4 个 stage 中会输出 4 个不同尺寸与通道的特征图。通过上采样、下采样的方式拼接这些特征图，在不同尺寸的特征图上进行预测。本文对即将上采样或者下采样的特征图提取注意力信息，具体的实现如图 6 所示。在 stage1 中，特征图的尺寸比较大，但是通道数较小，因此在最上层使用空间注意力机制。对于最下层 stage4，特征图的尺寸较小，但是通道数多，因此使用通道注意力机制。而中间的层则使用混合注意力机制。

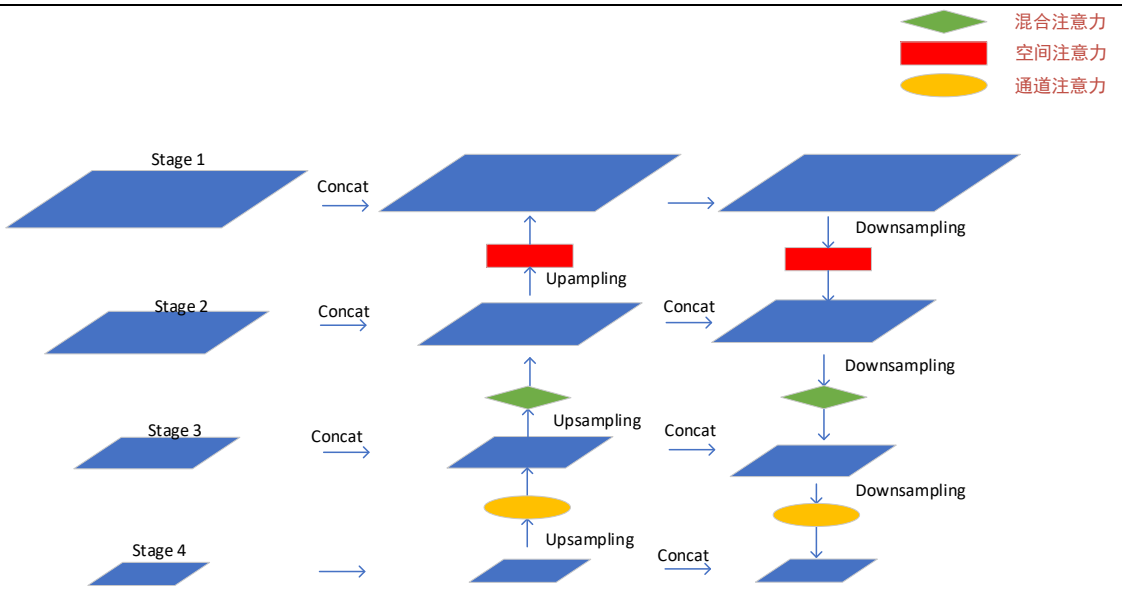


图 6 融合注意力机制的特征金字塔

2.5 模糊损失函数

在分析检测的图像预测框时，模型总是倾向于把模糊的物体识别为烟雾。因此，本文提出一种模糊损失，将其加入原有的损失函数进行优化。

在训练的每一轮迭代过后输出预测框，本文使用拉普拉斯算子对预测框内部计算模糊损失。如果选择的区域是模糊的，相应的模糊损失应当更小。对于图像，拉普拉斯算子的表达式如式(5)所示：

$$\nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = 4f(x,y) - f(x+1,y) - f(x-1,y) - f(x,y+1) - f(x,y-1) \tag{5}$$

其中 f(x+1，y)表示坐标(x+1，y)像素的值，对应的算子矩阵 g 如下所示：

$$g = \begin{pmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{pmatrix}$$

模糊损失的计算方法如下式所示：

$$val = (box_{pred} * -g).sum() \tag{6}$$

$$loss_{blur} = k^{val} / len(bs) \tag{7}$$

其中 k 是控制模糊损失的可学习参数，初始化为 0.9。len(bs)为每个 batch 图片数量。

3 实验

3.1 实验环境

本实验环境为：操作系统 Ubuntu，CPU Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz，128G 内存，显卡 Nvidia Tesla T4。实验在 Python3.9，CUDA11.6，Pytorch1.11 上进行，模型均使用 openmmlab 框架提供的预训练模型进行初始化。

3.2 数据增强

数据增强技术可以有效地增加样本的多样性，提高模型在不同环境下的鲁棒性。本次实验中采用了多种数据增强方法，其中 Mosaic 方法将四幅预处理图像拼接成一幅图像，丰富了被检测物体的背景，Random Affine 方法随机进行仿射变换，对图像进行旋转、缩放、平移和剪切操作， Augment HSV 方法随机调整色度，饱和度以及明度，Random horizontal flip 方法随机水平翻转，数据增强效果如图 7 所示。

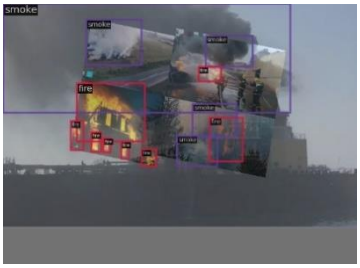


图 7　数据增强效果

3. 3 实验结果

数据集共 10827 张图片，按照 8:1:1 划分为训练集、验证集与测试集。根据实验结果显示，改进后的 YOLOX 模型的训练损失曲线如图 8 所示。从图中可以看出，随着训练轮数的增加，损失曲线逐渐趋于平稳。当 Epoch 数达到大约 250 时，模型逐渐收敛，训练过程没有出现过拟合的情况。

改进 YOLOX 与原版 YOLOX 的 mAP 曲线如图 9 所示。以 Swin Transformer 为主干网络的 YOLOX mAP 为 78. 15%。

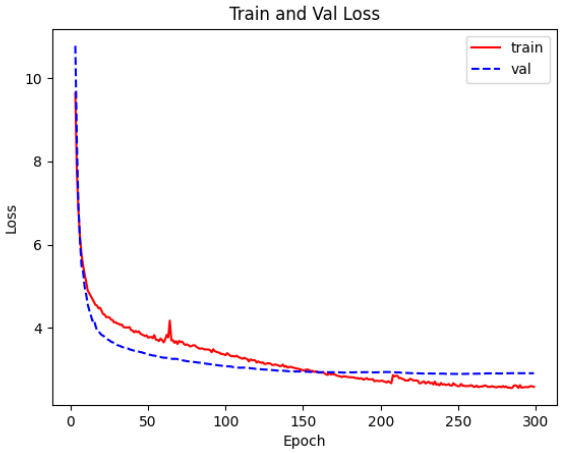


图 8　训练与验证损失

为了验证改进模型的检测性能，将其与主流模型 YOLOv3、Faster−RCNN 对比，结果如表 1 所示。表中 FPS 数据均是使用单张 Tesla T4 的测试结果，输入图片尺寸均调整为(640，640)。由表 1 中的数据可知，在同样的数据集下，改进的 YOLOX 相较于原始 YOLOX 算法 mAP 提高了 5. 75%，且远高于主流的一阶段与两阶段目标检测算法。在保证高精度的同时，模型的推理速度并未大幅下降。

表 1　主流目标检测模型性能对比

method	mAP	FPS
Faster−RCNN	71. 20%	30. 2
YOLOv3	71. 00%	31. 8
YOLOX	72. 40%	**34. 8**
Ours	**78. 15%**	33. 0

此外，本文还对比了使用模糊损失前后的性能，如表 2 所示。可以发现 mAP 有接近 1%的提升，并且检测速度没有太大影响。

表 2　使用模糊损失性能对比

method	mAP	FPS
Without blur loss	77. 22%	33. 5
With blur loss	**78. 15%**	33. 0

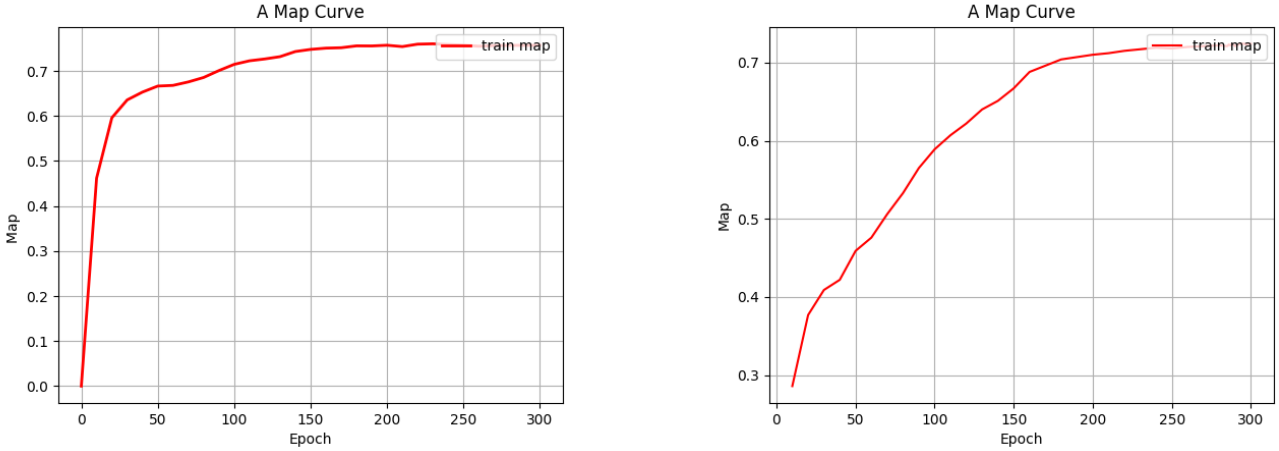


图 9 两种方法 Map 对比

4　结论

YOLOX 在目标检测领域效果胜过以往的 YOLO 系列算法，本文针对火灾场景提出了一种改进的 YOLOX 模型。该方法在 YOLOX 模型基础上，引入了 Swin Transformer 主干网络，加入了通道整合技术与多种注意力机制，加入模糊损失方法，提高了火焰烟雾的检测性能。检测结果如图 10 所示。可以看到，对于远距离的火焰和部分烟雾，改进后的模型能够很好的检测，但是对于充满屏幕的大范围的烟雾，存在一定的漏检。实验表明，基于注意力机制改进的 YOLOX 能够提高火焰与烟雾检测任务的准确性。

图 10　检测结果

参考文献：

[1] 吴凡(2020). 基于深度学习的火灾检测算法研究与实现(硕士学位论文, 杭州电子科技大学).
Fan Wu. (2020). Research and implementation of fire detection Algorithm based on Deep Learning(Unpublished Master thsis, Hanzhou Electronic Science and Technology University, Hangzhou, China).

[2] Huang, G., Liu, Z., Pleiss, G., Van Der Maaten, L., & Weinberger, K. Q. (2019). Convolutional networks with dense connectivity. IEEE transactions on pattern analysis and machine intelligence, 44(12), 8704-8716. .

[3] 张为 & 魏晶晶(2020). 嵌入 DenseNet 结构和空洞卷积模块的改进 YOLO v3 火灾检测算法. 天津大学学报(自然科学与工程技术版)(09), 976-983.
Wei Zhang, Jinjin Wei. (2020). Improved yolov3 fire detection algorithm embedded with densenet structure and atrous convolution module. Journal of Tianjin University (Natural Science and Engineering Technology), 53(9), 976 - 983.

[4] Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. arXiv preprint arXiv:1804. 02767.

[5] 王国睿(2021). 基于 Transformer 改进 YOLO v4 的火灾检测方法. 智能计算机与应用(07), 86-90.
Guorui Wang. (2021). Improved yolov4 fire detection method based on transformer. Intelligent Computers and Applications. 11(07):86 - 90.

[6] Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M. (2021). Scaled-yolov4: Scaling cross stage partial network. In Proceedings of the IEEE/cvf conference on computer vision and pattern recognition (pp. 13029-13038).

[7] 陆雅诺, & 陈炳才(2021). 融合注意力机制的无锚点森林火灾检测算法. 计算机与现代化.
Yanuo Lu, Bingcai Chen. (2021). Anchor-free forest fire detection algorithm based on attention mechanism. Computer and Modernization. 11:61 - 66+76.

[8] Zhou, X., Wang, D., & Krähenbühl, P. (2019). Objects as points. arXiv preprint arXiv:1904. 07850.

[9] Muhammad, K., Ahmad, J., Lv, Z., Bellavista, P., Yang, P., & Baik, S. W. (2018). Efficient deep CNN-based fire detection and localization in video surveillance applications. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 49(7), 1419-1434.

[10] 喻丽春 & 刘金清(2020). 基于改进 Mask R-CNN 的火焰图像识别算法. 计算机工程与应用(21), 194-198.
Lichun Yu, Jinqing Liu. (2020). Flame image recognition algorithm based on improved maskr-cnn. Computer Engineering and Applications. 56(21):194 - 198.

[11] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).

[12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[13] Yonglin Tian, Yutong Wang, Jiangong Wang, et al. Key Issues in Vision Transformers: State of the Art and Prospects [J]. Acta Automatica Sinica, 2022, 48(04):957-979. DOI:10. 16383/j. aas. c220027.

[14] Ge, Z., Liu, S., Wang, F., Li, Z., & Sun, J. (2021). Yolox: Exceeding yolo series in 2021 arXiv preprint arXiv:2107. 08430.

[15] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).

[16]Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010. 11929.

[17]Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16 (pp. 213-229). Springer International Publishing.

[18]Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 10012-10022).

[19]He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[20]叶铭亮, 周慧英 & 李建军(2022). 基于改进 Swin Transformer 的森林火灾检测算法. 中南林业科技大学学报 (08), 101-110. doi:10. 14067/j. cnki. 1673-923x. 2022. 08. 010.

Mingliang Ye, Huiying Zhou, Jianjun Li. (2022). Forest fire detection algorithm based on improved Swin Transformer. Journal of Central South University of Forestry and Technology, (08), 101-110.