第 58 卷 第 2 期/2021 年 1 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于轻量级注意机制的人脸检测算法

高刘雅，孙冬*，卢一相

安徽大学电气工程与自动化学院，安徽 合肥 230601

摘要 提出一个新的基于轻量级注意力机制的网络框架。在 YOLOv3 主干网络的基础上，使用深度卷积和点卷积代替标准卷积设计特征提取网络，加快模型的训练，提高检测的速度，然后引入注意力机制模块进行模型速度和精度的权衡，最后通过增加多尺度提取更多网络层的特征信息，同时使用 K-means++ 聚类算法进一步优化网络参数。实验结果表明，该方法可以显著提高人脸检测模型的性能，在 Wider Face 数据集上可以达到 94.08%的准确率和 83.97%的召回率，且平均检测时间只需 0.022 s，相比原始 YOLOv3 算法提高了 4.45 倍。

关键词 图像处理；人脸检测；深度学习；轻量级网络；注意力机制；K-means++

中图分类号 TP391.4 **文献标志码** A **doi**：10.3788/LOP202158.0210010

Face Detection Algorithm Based on a Lightweight Attention Mechanism Network

Gao Liuya, Sun Dong*, Lu Yixiang

College of Electric Engineering and Automation, Anhui University, Hefei, Anhui 230601, *China*

Abstract This study proposes a new network framework based on a lightweight attention mechanism and the YOLOv3 backbone network. When designing the feature extraction network, the standard convolutions of the YOLOv3 backbone network are replaced using depthwise and pointwise convolutions, thereby accelerating the model training and increasing the detection speed. Next, the speed and accuracy of the model are weighted using an attention mechanism module. Finally, multiple-scale prediction layers are added to extract more feature information; simultaneously, the network parameters are optimized using the K-means++ clustering algorithm. In an experimental evaluation on face-detection performance, this method considerably improved the face-detection performance, achieving 94.08% precision and 83.97% recall on the Wider Face dataset. The average detection time is 0.022 s, which is 4.45 times higher than that of the original YOLOv3 algorithm.

Key words image processing; face detection; deep learning; lightweight network; attention mechanism; K-means++

OCIS codes 100.4996; 150.0155

1 引 言

目标检测是计算机视觉中最基础的研究内容之一，也是当下计算机视觉中极为重要的一个分支，其目标是发展能够为计算机视觉应用提供所需基本信息的计算模型或技术。目标检测是对给定一幅图像中可变数量的目标进行定位和分类，目标种类与数量的不定性、目标尺度的多样性以及外在环境的干

收稿日期：2020-06-17；**修回日期**：2020-07-02；**录用日期**：2020-07-07

基金项目：国家自然科学基金(61402003)、安徽省高等学校自然科学基金(KJ2018A0012，KJ2019A0023，KJ2019A0022)、赛尔网络下一代互联网技术创新项目(NGII20180612，NGII20180312，NGII20180624)

* **E-mail**：sundong@ahu.edu.cn

扰等都会给目标检测任务带来不同程度的影响。

人脸检测是一类典型的目标检测任务，其方法经历了由传统到深度学习的变迁，传统的目标检测方法受限于图像特征的有效描述，只能手工提取特征，并结合滑动窗口的方式设计网络，进行目标检测，步骤复杂，准确度和实时性差。例如 2001 年由 Viola 和 Jones[1] 提出的 Viola-Jones(VJ)检测器，第一次在不受人体征约束的情况下实现了人脸的实时检测，VJ 算法的特征提取部分采用 Haar 特征[2] 作为特征表示，并使用滑动窗口的策略，虽然引入级联检测来减少人脸目标之外的计算，但尺度的变化和步长的变化仍会使算法本身出现大量的冗余候选框，导致检测速度下降。随后，针对行人检测的方向梯度直方图(HOG)[3] 特征描述算子在 2005 年被首次提出，该算法在保持检测窗口大小不变的前提下缩放输入图像的尺寸，以此来适应不同大小的检测目标，并利用支持向量机(SVM)[4] 训练得到物体的梯度，通过计算梯度的方向来得到检测物体的统计直方图。基于 HOG 特征进行扩展和优化，Felzenszwalb 等[5] 提出了 Deformable Part-Based Model(DPM)算法，采用多组件和图结构的模型策略解决了物体检测中的多视角以及形变问题，但该方法依旧是基于手工特征来设计的，且是针对于某个物体制定固定的激励模板，因此不具有普适性。

直到 2012 年以后，卷积神经网络(CNNs)的快速发展给人脸检测任务提供了更加灵活多变的方法，从最初的 VGG[6] 网络到 Inception[7] 网络再到 Resnet[8] 网络，深度学习阶段的算法模型整体上呈现出更深更宽的趋势，但这些方法只是利用卷积神经网进行特征的提取，并没有从本质上改变搜索框提取目标区域的策略，因此这些方法在检测速度上依然没有得到有效的提升。除了通过增加网络的深度来学习更多局部抽象的特征之外，基于区域卷积神经网络(RCNN)的 two-stage 算法(如，SSP-Net[9]，Fast-RCNN[10]，Faster-RCNN[11])进一步构造了区域建议网络，极大地提高了各类目标检测任务中的检测精度，但这类方法训练时间较长，检测速度依旧不能满足快速实时的需求。后期发展而来的直接回归目标框位置的 one-stage 目标检测算法更加注重检测速度的提升，例如 YOLO (You Only Look Once)系列算法[12-14]、包含锚点机制的 SSD (Single shot multibox detector)算法[15] 等，在产生检测框的同时对目标物体的类别概率和位置坐标进行分类和回归，检测速度相对 two-stage 算法有很大提升。后续基于各类神经网络进行了广泛的研究，如使用主流网络进行迁移学习训练[16]，改进网络模型以提取更多更优的局部特征表示[17-19]，使用各类增强算法[20-22] 提高检测效率，设计新的损失函数优化模型利用率[23-24] 等。

以旷视、商汤为代表的在学术界公开竞赛中取得好成绩的厂商也开始发展以实际业务为起点的目标检测算法，除了进一步增加实际数据集合、提升算法性能外，也开始在不降低识别效率的基础上研究网络的轻量化，其目的主要是提升算法的速度，力争将其部署到移动端，扩大算法的实际应用。

本文工作主要分为以下几方面：

1) 结合轻量级网络的设计理念改进主干网络，减少检测过程中的计算量参数，实现人脸检测速度的实质性提升；同时引入注意力机制模块权衡检测速度和检测精度。

2) 借鉴 YOLOv3 算法的多尺度检测输出特性，将原先的 3 个尺度扩增到 4 个尺度，以丰富网络的感受野范围，获取更多的特征描述信息。

3) 使用 K-means＋＋算法初始化 anchor box 的坐标，选择更适合人脸数据集的先验框，提高人脸检测模型的召回率。

2 算法模型设计分析

2.1 轻量级卷积网络

作为一种含有轻量级注意力机制模型的网络，Mobilenetv3[25] 结合了 Mobilenetv1[26] 中的深度可分离卷积(DSC)以及 Mobilenetv2[27] 网络中的具有线性瓶颈的逆残差结构(Inverted residual with linear bottleneck)，并提出用 h-swish 函数代替原先深度卷积网络中的 swish 激活函数，减少模型训练过程的计算量参数，表达式为

$$\text{h-swish}[x]=x\frac{\text{ReLU}(x+3)}{6}, \tag{1}$$

$$\text{swish}[x]=x*\text{sigmoid}(\beta x), \tag{2}$$

$$\text{ReLU}(6)=\min(\max(0,x),6)。 \tag{3}$$

对于输入为 $D_F\times D_F\times M$ 的特征图，在标准卷积核 $D_K\times D_K\times M\times N$ 作用下可得到的特征图尺寸为 $D_F\times D_F\times N$，其中 M 和 N 表示输入输出的通道数，模型计算量为 $D_K\cdot D_K\cdot M\cdot N\cdot D_F\cdot D_F$。

当用深度可分离卷积代替标准卷积后，输入特征图尺寸被分别分解为 $D_K\times D_K\times 1\times M$ 的深度卷积以及 $1\times 1\times M\times N$ 的逐点卷积，对应的输出特征为 $D_F\times D_F\times M$ 和 $D_K\times D_K\times N$，此时模型的计算

量为 $D_K \cdot D_K \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F$。因此，改进后的轻量级网络的参数压缩量为 $\frac{D_K \cdot D_K \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F}{D_K \cdot D_K \cdot M \cdot N \cdot D_F \cdot D_F} = \frac{1}{N} + \frac{1}{D_K^2}$。

2.2　注意机制模块

Mobilenetv3 中注意力机制的引入主要是为了平衡模型的体积、速度和精度。如图 1 所示，SE-Block 主要通过压缩（Squeeze）和激励（Excitation）操作对网络模型之间的特征关系进行校准，实质上是增大有效的特征权重，减小无效或效果作用小的权重，从而实现注意力的集中，加强整体网络的学习能力。对于一个输入为 $W \times H \times C$ 的特征图，首先经过 Squeeze 操作将其压缩成 $1 \times 1 \times C$ 的向量，然后再通过一次 Excitation 操作实现模型的辅助泛化，最后对特征图进行 Scale 操作，所有经过 Squeeze 和 Excitation 操作计算得到的权重分别与输入的原始特征图相应通道的二维矩阵相乘，得到最后输出特征图。

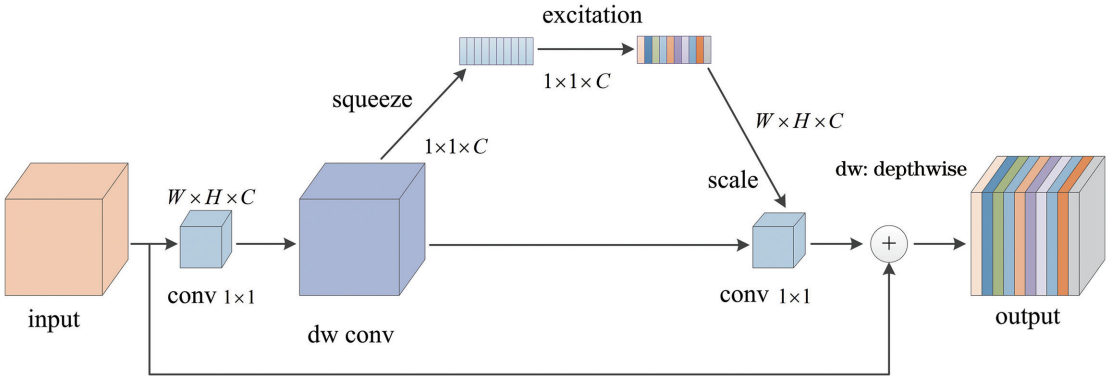

图 1　SE-Block 结构示意图

Fig. 1　SE-Block structure diagram

2.3　YOLOv3 多尺度检测

YOLOv3 采用类似于特征金字塔（FPN）[28] 的多尺度检测策略来提高对小物体的检测精度。YOLOv3 的多尺度特征可以同时兼顾大中小目标的检测，预测阶段分别从三个不同尺度的 Feature Map 上提取相应的特征作为 YOLOv3 检测的输入，三个尺度的维度分别为 $13 \times 13 \times N$、$26 \times 26 \times N$、$52 \times 52 \times N$。YOLOv3 的网络结构如图 2 所示。

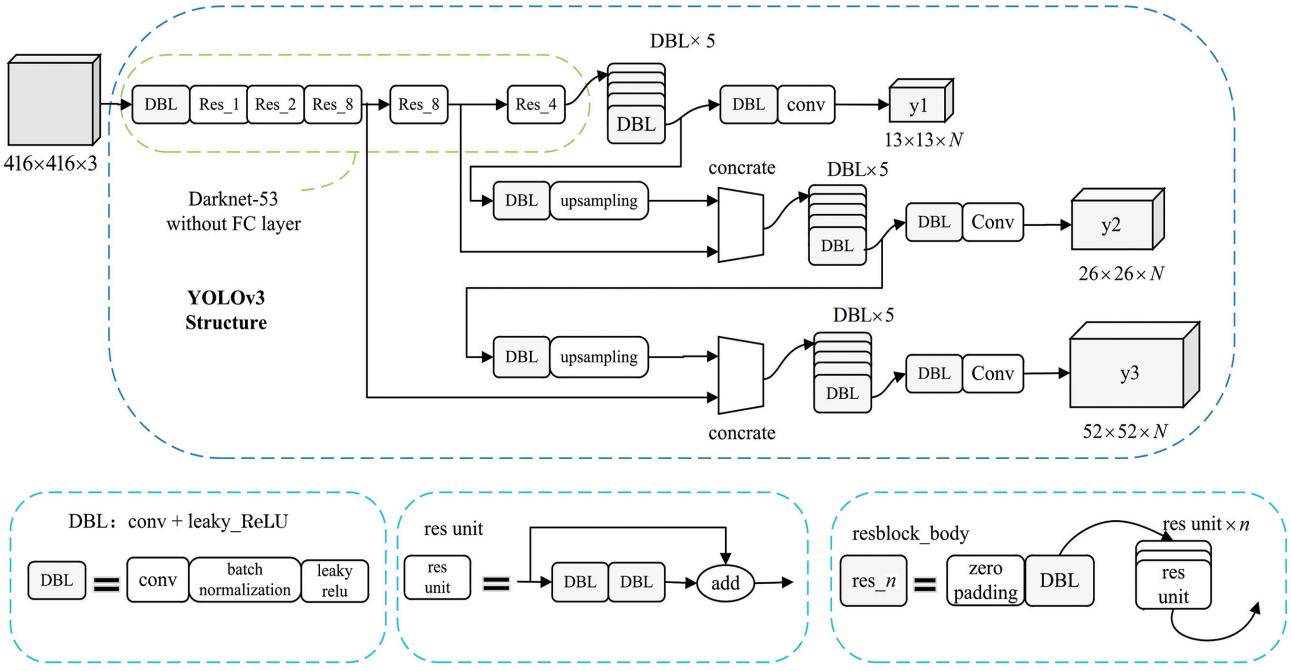

图 2　YOLOv3 可视化结构

Fig. 2　Visual structure of YOLOv3 network

3　改进的轻量级注意力机制模型

3.1　改进的网络结构

改进后的模型使用 Mobilenetv3 主干网络替换掉 YOLOv3 中的 Darknet-53 网络，如图 3 所示，即用 Mobilenetv3 的深度可分离卷积和点卷积替换原来 YOLOv3 中的标准卷积，极大地削减了主干网络中卷积部分的运算量，使得网络的整体计算量大大减少。为了在预测阶段获得更高的精度，网络进行特征提取时必须兼顾浅层网络的位置特征和深层网络的语义特征，本文设计 4 个不同尺度进行输出预测，分别负责图片中大、中、小、极小目标的检测。

name	input size	operator	s	onput size
the first feature extraction layers	416×416×3	conv2d	2	208×208×16
	208×208×16	bneck, 3×3	1	208×208×16
	208×208×16	bneck, 3×3	2	104×104×24
	104×104×24	bneck, 3×3	1	104×104×24
the second feature extraction layers	104×104×24	bneck, 5×5	2	52×52×40
	52×52×40	bneck, 5×5	1	52×52×40
	52×52×40	bneck, 5×5	1	52×52×40
the third feature extraction layers	52×52×40	bneck, 3×3	2	26×26×80
	26×26×80	bneck, 3×3	1	26×26×80
	26×26×80	bneck, 3×3	1	26×26×80
	26×26×80	bneck, 3×3	1	26×26×80
	26×26×80	bneck, 3×3	1	26×26×112
	26×26×112	bneck, 3×3	1	26×26×112
the fourth feature extraction layers	26×26×112	bneck, 5×5		13×13×160
	13×13×160	bneck, 5×5		13×13×160
	13×13×160	bneck, 5×5		13×13×160
	13×13×160	conv 2d, 1×1		13×13×960

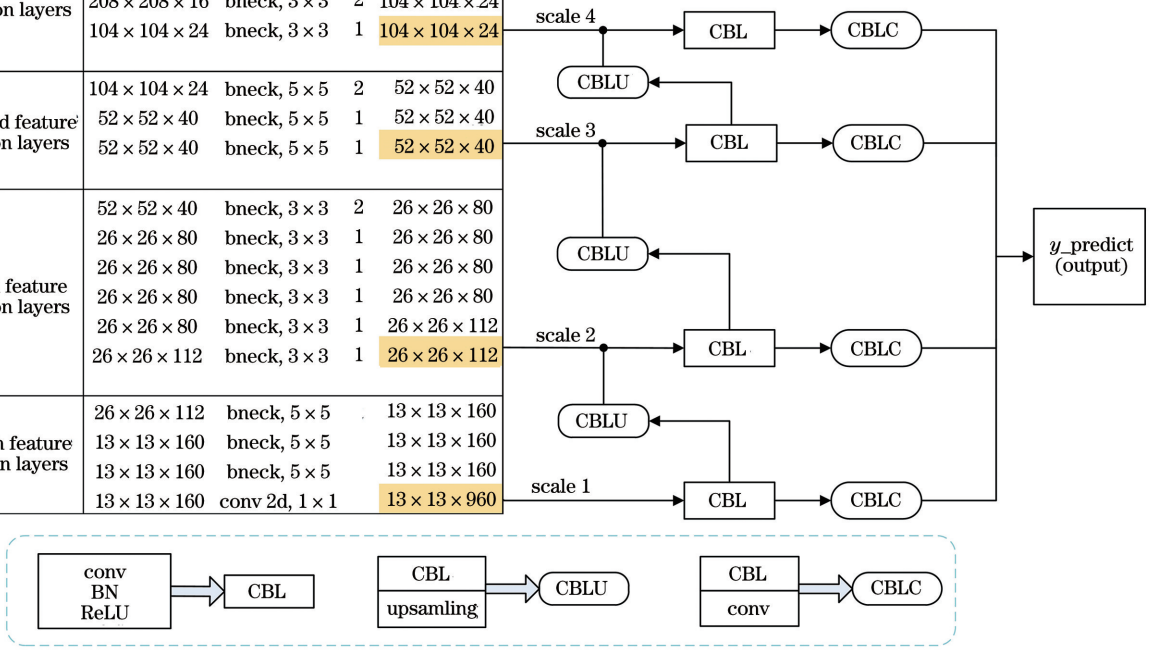

图 3　改进后的轻量级多尺度网络

Fig. 3　Improved lightweight multi-scale network

因为不同的特征映射具有不同的接受域和上下文信息，所以在不同的特征图上检测不同尺度的人脸是合理的。类比 YOLOv3 的多尺度特性，尺度 4 会对尺度 3 输出的卷积结果进行一次上采样，再与 104×104 的 Feature Map 相连接，最后通过多个卷积层输出 Bounding box 的预测信息。

3.2　K-menas＋＋ 聚类算法

目标检测任务中使用聚类算法的目的是使先验框(anchor box)与标注框(ground truth)的交并比(IOU，η)尽可能大，因此目标函数采用 η 作为衡量的标准，距离公式的定义为

$$d(R_{\text{box}},R_{\text{centroid}})=\min\sum_{R_{\text{box}}=0}^{n}\sum_{R_{\text{centroid}}=0}^{k}(1-\eta_{R_{\text{centroid}}}^{R_{\text{box}}}),\tag{4}$$

其中，R_{box} 为样本标签的目标框，R_{centriod} 为聚类中心，n 和 k 分别指数据集中的样本数量和类别数量，合适的 η 可以很好地权衡模型的复杂度和检测召回率。

对于 anchor 的设计，选择 K-means＋＋ 算法对人脸目标进行聚类，图 4 对比了原始聚类算法和 K-means＋＋ 算法在人脸数据集上的聚类结果。依据改进后 4 尺度预测特征，最终确定 12 个 anchor 值，并均匀分布在 4 个尺度上，每一种尺度可以预测三个 Bounding box。除此之外，对于每种尺度都会

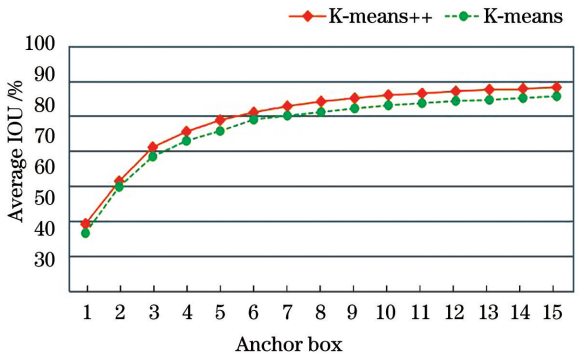

图 4　平均交并比与锚点框的关系

Fig. 4　Relationship between average IOU and anchor box

引入一些卷积层来进一步提取特征，之后再输出预测 box 的信息。本次实验中，由 K-means++ 聚类确定的 12 个人脸目标的 anchor 值分别为(22,26)、(26,33)、(31,37)、(31,48)、(36,42)、(41,53)、(52,65)、(67,88)、(90,120)、(124,168)、(196,260)、(361,470)。

3.3 损失函数

以 Darknet 训练的 YOLOv3 源码为依据，分析总结 YOLOv3 损失函数，主要包括坐标误差、置信度误差以及分类误差。

$$\alpha = E_{\text{coord}} + E_{\text{IOU}} + E_{\text{class}}, \tag{5}$$

其中，

$$E_{\text{coord}} = \lambda_{\text{coord}} \sum_{i=0}^{S\times S} \sum_{j=0}^{M} I_{ij}^{\text{obj}} \left[(t_{x_i} - t'_{x_i})^2 + (t_{y_i} - t'_{y_i})^2 \right] + \lambda_{\text{coord}} \sum_{i=0}^{S\times S} \sum_{j=0}^{M} I_{ij}^{\text{obj}} \left[(\sqrt{t_{w_i}} - \sqrt{t'_{w_i}})^2 + (\sqrt{t_{h_i}} - \sqrt{t'_{h_i}})^2 \right], \tag{6}$$

$$E_{\text{IOU}} = \sum_{i=0}^{S\times S} \sum_{j=0}^{M} I_{ij}^{\text{obj}} \left[c'_i \log c_i + (1 - c'_i) \log(1 - c_i) \right] + \lambda_{\text{noobj}} \sum_{i=0}^{S\times S} \sum_{j=0}^{M} I_{ij}^{\text{noobj}} \left[c'_i \log c_i + (1 - c'_i) \log(1 - c_i) \right], \tag{7}$$

$$E_{\text{class}} = \sum_{i=0}^{S\times S} I_i^{\text{obj}} \sum_{c\in C} \left\{ p'_i(c) \log p_i(c_i) + [1 - p'_i(c)] \log[1 - p_i(c_i)] \right\}, \tag{8}$$

式中：t_{x_i}、t_{y_i}、t_{w_i}、t_{h_i} 是预测的目标物体中心点的横纵坐标值、宽度以及高度；t'_{x_i}、t'_{y_i}、t'_{w_i}、t'_{h_i} 是相对应的真实值；λ_{coord}、λ_{noobj} 分别为预测坐标时的惩罚系数以及没有检测目标时置信度的惩罚系数；输入图像被划分为 $S\times S$ 的网格，每个网格中被预测的物体的边框数为 M；$I_{ij}^{\text{obj}}=1$，$I_{ij}^{\text{noobj}}=0$ 是指在第 i 个网格中第 j 个候选目标的边框中能够成功检测出物体，$I_{\text{obj}}^{ij}=0$，$I_{ij}^{\text{noobj}}=1$ 是指在第 i 个网格中第 j 个候选目标的边框中没有物体被检测出来；c 是目标所属的类别，c'_i 指第 i 个网格中目标所属类别的真实置信度，c_i 指第 i 个网格中目标所属类别的预测置信度，p 代表概率计算，C 指类别总个数。

由上述公式可知，损失函数中既使用了平方和损失函数计算坐标误差，也采用了二元交叉熵损失函数对置信度和分类误差进行计算，该方法可以有效地避免训练过程中的梯度消失问题。

4 实验与结果分析

4.1 数据集和评价指标

Wider Face[29]人脸数据集与当前许多公开的数据集相比，具有更加明显的人脸面部特征，收集了具有不同尺度、不同典型姿态、不同遮挡重叠等高度可变性的人脸目标，可以更好地验证模型的泛化能力。实验中使用的数据集是经预处理转换成 VOC 格式的局部数据集，共包含 61 种不同环境下的 12880 张训练图片，以及 3226 张测试图片。

检测效果的好坏往往需要通过一定的指标去度量。本研究用于评价的重要指标除了包含检测的精确率和召回率之外，还考虑到模型是否具有较好的实时性。相关指标的公式定义为

$$P_{\text{precision}} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FP}}} \times 100\%, \tag{9}$$

$$R_{\text{recall}} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FN}}} \times 100\%。 \tag{10}$$

用 P 表示 $P_{\text{precision}}$，R 表示 R_{recall}，则进一步权衡 P-R 的关系表示为

$$F_1 = \frac{2}{\frac{1}{P} + \frac{1}{R}} = 2\frac{PR}{P+R}, \tag{11}$$

其中：N_{TP} 表示所有被检测出的人脸中被正确分类为人脸的个数；N_{FP} 表示检测框误将背景标注成人脸的个数；N_{FN} 表示测试集中没有被检测到的人脸数量。

4.2 实验结果及分析

4.2.1 训练结果

本次实验中，对模型分两个阶段进行训练，第一阶段的 epoch 范围设置为 0 到 180，第二阶段设置为 180 到 400，随着训练过程中 epoch 数量的不断增加，网络模型的损失函数值逐渐趋于稳定，如图 5 所示，

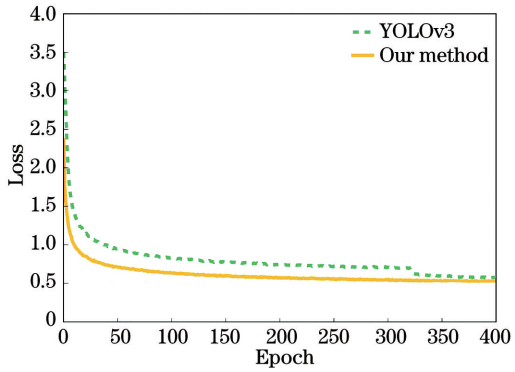

图 5　不同模型损失函数变化曲线

Fig. 5　Loss curves of different models

图中展示了原始 YOLOv3 网络和改进后网络的损失曲线对比图，上侧曲线代表了原始算法的变化趋势，下侧曲线是本文方法的损失函数变化曲线。可见改进后的网络初始损失函数值较小，且算法的收敛速度更快，相比原算法，模型的总体损失值更小。

4.2.2　不同算法检测指标的定量分析

本研究用来检测算法效果的目标图片中共含有 7681 张人脸，为了证明本文所提方法的优越性，使用 Faster R-CNN、YOLOv3、mv3-YOLOv3（只改变 YOLOv3 主干网络，未增加预测尺度）分别在 Wider Face 数据集上进行训练测试。不同算法模型的检测结果对比如表 1 所示。由结果可知，YOLOv3 算法相较于 two-stage 中经典的 Faster-RCNN 算法，各项评价指标效果更佳，检测速度提高近 30 倍，但检测的召回率依旧很低。引入轻量级深度可分离卷积结构作为主干网络的 mv3-YOLOv3 模型在进行人脸检测时，保证了与原算法相近的精确度，但检测速度得到大幅度提升，这说明轻量级网络在实时性检测方面确实具有更好的稳健性。进一步改进网络的预测尺度后，算法的检测率进一步提高了 1.77 个百分点，同时召回率提高了 10.2 个百分点，这说明改进后的算法可以在精确度和召回率之间实现更好的平衡，网络模型的整体性能也更加稳定，更适合于进行复杂环境背景下的人脸检测。

表 1　不同算法检测指标对比

Table 1　Comparison of detection indicators of different algorithms

Detection algorithm	Precision /%	Recall /%	AP /%	F1 /%	Average time /s
Faster-RCNN	88.41	62.59	61.39	73.29	3.077
YOLOv3	92.31	73.81	73.13	82.11	0.098
mv3-YOLOv3	91.30	70.64	69.82	79.65	0.026
Our algorithm	**94.08**	**83.97**	**83.33**	**88.73**	**0.022**

4.2.3　检测结果分析

改进后的网络和原始 YOLOv3 算法的检测效果对比如图 6 所示，第一行展示了 YOLOv3 网络的检测结果，第二行是改进后的检测效果。由图可见，改进后的模型对重叠人脸、光照肤色差异下的人脸以及密集人群中的人脸的检测效果更佳。如左侧图中，YOLOv3 算法把两张重叠的人脸当成一个目标检测输出，但改进后的网络可以正确地区分并检测出两张人脸，这说明改进后的网络具有更好的分类性能；中间一组对比结果中，原始算法漏检了图像两

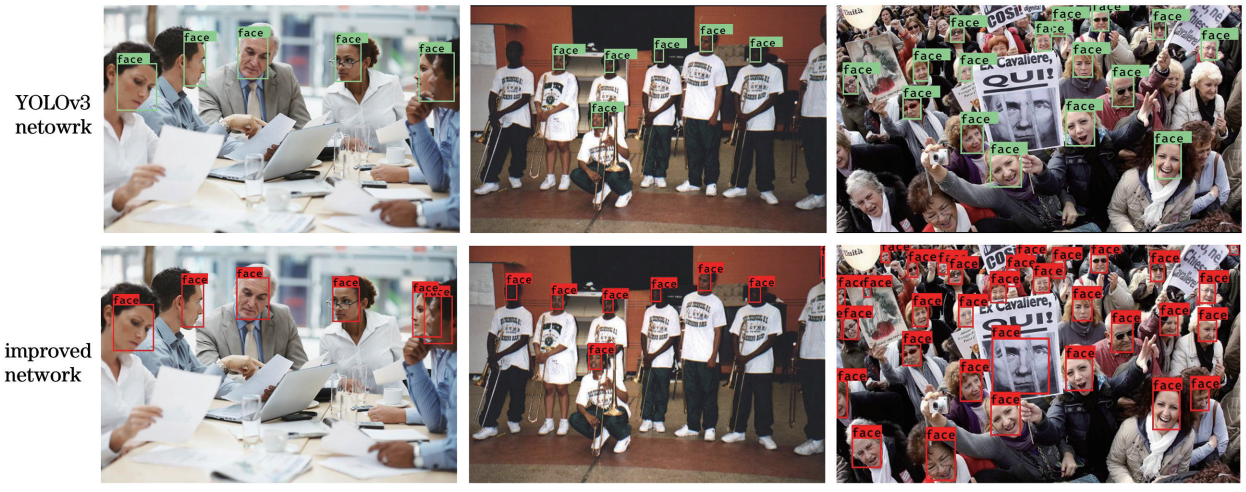

图 6　YOLOv3 网络和改进后网络检测效果对比

Fig. 6　Detection results between YOLOv3 and improved network

侧肤色较暗的人脸目标，而改进后的网络则可以学习更多的细节特征，具有更高的检测效率；对于人脸分布密集、姿态多样化的场景（如右侧对比图），改进后的算法体现出更大的优越性，对目标人脸的召回率远远高于原始算法，且可以准确检测出不同尺度、不同类别（真实人脸和海报人脸共存）的人脸目标。

图 7 进一步展示了改进模型在不同事件环境下的人脸检测效果图，列出了更多定量性的输出结果，所有定量性指标既包括了 Bounding box 的位置和类别信息，也包括检测出的人脸的 score 得分。检测任务可具体分为：多尺度共存、聚焦模式不同、光照差异、面部遮挡、多种面部表情姿态、不同肤色差异、不同稀疏稠密等，Bounding box 的得分直观反映了人脸检测的查准率，并可以通过输出的检测框数量评估改进后算法在人脸检测任务上的召回率。从整个实验结果上看，改进后的网络模型不仅实现了快速人脸检测的目的，还可以在查准率和召回率之间获取更好的平衡，对自然场景下的人脸检测任务更加具有普适性。

图 7　改进后网络对不同图像的定量性检测结果

Fig. 7　Quantitative detection results with improved network for different images

5　结　　论

针对检测实时性差和检测召回率低的问题，提出含有注意力机制的轻量级网络模型，通过在 Wider Face 数据集上进行实验，证明改进后的算法能够显著地提高自然场景下人脸检测的速度，并且在获得较高召回率的同时还能确保准确率维持在较高的范围内。该模型对现实生活中的人脸识别、人脸检测场景的发展具有积极的作用，在未来的研究中，我们将考虑把轻量级模型嵌入到实际应用设备中，并根据特定场景进一步改进 anchor 匹配策略，以提高模型的泛化能力。

参　考　文　献

[1] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features[C]//Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), December 8-14, 2001, Kauai, HI, USA. New York: IEEE Press, 2001: I.

[2] Papageorgiou C P, Oren M, Poggio T. A general framework for object detection[C]//Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271), January 7-7, 1998, Bombay, India. New York: IEEE Press, 1998: 555-562.

[3] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]// Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2005, San Diego, CA, USA. New York: IEEE Press, 2005: 886-893.

[4] Cortes C, Vapnik V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273-297.

[5] Felzenszwalb P F, Girshick R B, McAllester D. Cascade object detection with deformable part models[C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 13-18, 2010, San Francisco, CA, USA. New York: IEEE Press, 2010: 2241-2248.

[6] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-09-04)[2020-05-30]. https://arxiv.org/abs/1409.1556.

[7] Szegedy C, Liu W, Jia Y Q, et al. Going deeper with convolutions[C]// Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 1-9.

[8] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[9] He K M, Zhang X Y, Ren S Q, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[10] Girshick R. Fast R-CNN[C]// Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, Santiago, Chile. New York: IEEE Press, 2015: 1440-1448.

[11] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//IEEE Transactions on Pattern Analysis and Machine Intelligence, New York: IEEE Press, 2017: 1137-1149.

[12] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 779-788.

[13] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6517-6525.

[14] Redmon J, Farhadi A. YOLOv3: an incremental improvement[EB/OL]. (2018-04-08)[2020-05-30]. https://arxiv.org/abs/1804.02767.

[15] Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector[M]//Computer Vision-ECCV 2016. Cham: Springer International Publishing, 2016: 21-37.

[16] Duan Z J, Li S B, Hu J J, et al. Review of deep learning based object detection methods and their mainstream frameworks[J]. Laser & Optoelectronics Progress, 2020, 57(12): 120005.
段仲静, 李少波, 胡建军, 等. 深度学习目标检测方法及其主流框架综述[J]. 激光与光电子学进展, 2020, 57(12): 120005.

[17] Jiang H Z, Learned-Miller E. Face detection with the faster R-CNN[C]// 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), May 30-June 3, 2017, Washington, DC, USA. New York: IEEE Press, 2017: 650-657.

[18] Wang Y, Zheng J C. Real-time face detection based on YOLO[C]// Proceedings of the 2018 1st IEEE International Conference on Knowledge Innovation and Invention (ICKII), July 23-27, 2018, Jeju, South Korea. New York: IEEE Press, 2018: 221-224.

[19] Zhang J L, Wu X W, Hoi S C H, et al. Feature agglomeration networks for single stage face detection [J]. Neurocomputing, 2020, 380: 180-189.

[20] Ling Y, Chen Y. Salient object detection with multiscale context enhanced fully convolutional network[J]. Journal of Computer-Aided Design & Computer Graphics, 2019, 31(11): 2007-2016.
凌艳, 陈莹. 多尺度上下文信息增强的显著目标检测全卷积网络[J]. 计算机辅助设计与图形学学报, 2019, 31(11): 2007-2016.

[21] Xie X L, Li C X, Yang X G, et al. Salient object detection algorithm based on dual-attention recurrent convolution[J]. Acta Optica Sinica, 2019, 39(9): 0915005.
谢学立, 李传祥, 杨小冈, 等. 双注意力循环卷积显著性目标检测算法[J]. 光学学报, 2019, 39(9): 0915005.

[22] Zhao B, Wang C P, Fu Q, et al. Multi-scale infrared pedestrian detection based on deep attention mechanism[J]. Acta Optica Sinica, 2020, 40(5): 0504001.
赵斌, 王春平, 付强, 等. 基于深度注意力机制的多尺度红外行人检测[J]. 光学学报, 2020, 40(5): 0504001.

[23] Wang R S, Tian J Z, Jin C L. Joint face detection and alignment using focal loss-based multi-task convolutional neural networks[C]// Sun Z, He R, Feng J, et al. Biometric Recognition. Cham: Springer, 2019: 266-273.

[24] Ren Z J, Lin S Z, Li D W, et al. Mask R-CNN object detection method based on improved feature pyramid[J]. Laser & Optoelectronics Progress, 2019, 56(4): 041502.
任之俊, 蔺素珍, 李大威, 等. 基于改进特征金字塔的 Mask R-CNN 目标检测方法[J]. 激光与光电子学进展, 2019, 56(4): 041502.

[25] Howard A, Sandler M, Chu G, et al. Searching for mobilenetv3[C]// Proceedings of the IEEE International Conference on Computer Vision,

October 27-November 2, 2019, Seoul, Korea. New York: IEEE Press, 2019: 1314-1324.

[26] Howard A G, Zhu M, Chen B, et al. Mobilenets: efficient convolutional neural networks for mobile vision applications[EB/OL]. (2017-04-17)[2020-05-30]. https://arxiv.org/abs/1704.04861.

[27] Sandler M, Howard A, Zhu M L, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 4510-4520.

[28] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 936-944.

[29] Yang S, Luo P, Loy C C, et al. WIDER FACE: a face detection benchmark[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 5525-5533.