Software Engineering and Applications 软件工程与应用, 2022, 11(2), 185-193
Published Online April 2022 in Hans. http://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2022.112020

Hans汉斯

密集人群小尺度人脸目标检测方法研究

赵　耀，秦　学，王　腾

贵州大学，贵州　贵阳

收稿日期：2022年3月8日；录用日期：2022年3月23日；发布日期：2022年3月30日

摘　要

针对密集人群图像的人脸目标检测，普遍存在检测出的小尺度人脸目标特征少(不足)等问题，本文提出了一种改进YOLO网络的小型预测特征图特征融合的方法。该方法从浅层网络引出特征图，采用改进的DenseNet增强语义特征后，加入到小型预测尺度特征图，用于丰富小型人脸预测尺度的特征语义信息，进而提高小尺度人脸检测效果。在WIDER FACE数据集上对所提方法进行测试，结果表明，所提方法对密集人群小尺度小人脸的检测精度有较好的提升。

关键词

计算机视觉，人脸检测，YOLO，DenseNet

Research on Small-Scale Face Target Detection Method for Dense Population

Yao Zhao, Xue Qin, Teng Wang

Guizhou University, Guiyang Guizhou

Received: Mar. 8th, 2022; accepted: Mar. 23rd, 2022; published: Mar. 30th, 2022

Abstract

For the face target detection of dense crowd images, there are many problems, such as few (insufficient) small-scale face target features. This paper proposes a feature fusion method of small predictive feature map based on improved Yolo network. This method leads out the feature map from the shallow network, uses the improved DenseNet to enhance the semantic features, and adds it to the small-scale prediction scale feature map to enrich the feature semantic information of the small-scale face prediction scale, so as to improve the effect of small-scale face detection. The proposed method is tested on the WIDER FACE dataset. The results show that the proposed method can improve the detection accuracy of small-scale face of dense population.

文章引用: 赵耀, 秦学, 王腾. 密集人群小尺度人脸目标检测方法研究[J]. 软件工程与应用, 2022, 11(2): 185-193.
DOI: 10.12677/sea.2022.112020

Keywords

Computer Vision, Face Detection, YOLO, DenseNet

1. 引言

人脸检测是目标检测的一种特殊类型，人脸检测的任务是识别图像和视频中所有感兴趣的目标(人脸)，判定是否为人脸，并确定目标的位置和大小。由于人脸的角度、背景光照、尺度大小、类人脸等干扰因素，人脸检测是目标检测中极具有挑战性的问题。

不同于一般应用场景下的人脸检测，密集人群下人脸目标具有所占像素少、覆盖面积小等特点。由于受到遮挡、角度、模糊、尺度过小等因素影响，导致密集场景下的人脸检测难度较大，其中以尺度过小的问题尤为明显[1]。

目前，关于人脸检测方法的研究可以分为基于传统方法和基于深度学习方法两个方向。基于传统方法以模板匹配技术和基于 AdaBoost 人脸检测方法为代表。如 Rowley 等[2] [3]提出的一种基于模板匹配技术的人脸检测方法，该方法使用多角度人脸检测系统调整输入人脸角度，然后将调整角度后的人脸输入到训练过的多层感知器中，判断是否为人脸。Viola 等人[4]于 2001 年设计了一种基于 AdaBoost 人脸检测算法，该方法使用简单 Haar-like 特征和级联的 AdaBoost 分类器构造检测器，在实时高效的条件下取得了不错的检测精度[5]。由于上述方法采用的特征是基于人工设计，其特征的稳定性较低，易受光照、角度等外界环境的影响，因此对于复杂环境下的人脸检测性能很难得到保证[6]。

近年来，基于深度学习的算法在计算机视觉领域的各项任务中都取得了惊人的效果，在人脸的检测过程中采用深度学习的方法能获得更稳定的和更丰富的人脸特征信息，常用于密集人群中多目标小尺度的人脸检测任务，而且在大部分复杂环境下的人脸检测比传统方式检测效果更好。

Li 等[7]提出级联卷积的神经网络人脸检测方法(CascadeCNN)一定程度上解决了传统方法在非限制环境中对光照等敏感的问题。Zhang K 等人[8]提出由三个级联组成 MTCNN 模型，能够很好实现人脸检测和人脸对齐任务，级联结构使人脸检测的效果进一步提升。

同时，基于深度学习的人脸检测是通用目标检测的一种特殊类型，由于通用目标检测算法性能的提升，相关学者将一些通用目标检测算法应用于人脸检测领域，比如为了减少类内距离，Wang 等[9]在 Faster R-CNN 的基础上提出一种新的损失函数，提升了人脸的检测精度；Jiang 等[10]将 Faster R-CNN 应用到小尺度人脸检测中取得较好的检测效果。上述通用目标检测方法的检测精度较高，但检测速度欠佳，为了提升检测速度，相关学者研究后又提出了以 SSD [11]和 YOLO [12]为代表的 One-Stage 目标检测算法，这种目标算法在检测精度较高的前提下，其检测速度比上述通用目标检测算法效率更高，满足检测任务实时性的一般需求。针对密集人群图像中可能出现的多尺度、小尺度人脸检测等问题，本文选取 YOLOv3 算法进行人脸检测任务，该方法比同类 SSD 算法的检测性能更具优势[12]，更能满足人脸检测任务要求。

针对密集人群中的小尺度人脸检测问题，本文首先将小型目标预测特征尺度前的张量拼接 Concat 改为张量叠加，通过将多个特征向量组合成复向量，获得更多的小尺度人脸检测特征信息，以此提升小尺度人脸检测的召回率。第二，为了丰富小型预测尺度的语义信息，本文将原网络的 4 倍下采样特征图输

出，进行语义特征增强的密集卷积模块操作，再将此特征图与 26 × 26 特征图上采样处进行张量叠加，实现不同层次网络的语义特征融合，从而使网络获取到更多的小尺度人脸检测信息特征，提高人脸检测的效果。

2. 小尺度人脸目标检测的网络结构设计

本文结合小尺度人脸检测特点，在 YOLOv3 网络结构做了针对性的改进和优化，以期更适合密集人群下小尺度人脸目标检测需求。在 YOLOv3 原网络结构设计中，针对小物体检测能力较弱的问题，采用特征金字塔(FPN)和上采样的思想进行多尺度融合，该方法通过自底向上的途径进行特征图的下采样，获得更多的语义特征，然后采用从上到下的路径将更高层特征图进行上采样，获取更多的浅层位置信息，并通过横向连接的方式进行特征融合实现不同层次网络的语义特征和位置特征丰富的效果。最后在多尺度的特征图上进行检测从而进一步提高模型对于小目标物体的检测性能。

通常浅层特征图中的高级语义特征相对较少，但小尺度目标的坐标信息更详细；而深层特征图位置信息模糊，但语义特征相对浅层特征图来说更多。如果能结合浅层特征图的位置信息与深层特征图的语义特征，将能大大提升模型性能。

2.1. 网络结构的改进

为了提高密集人群下小尺度人脸检测的检测效果，本文对 YOLOv3 网络进行了三处优化设计，改进的 YOLOv3 网络结构图如图 1 所示，原网络的 5 个残差块分别由 1×、2×、8×、8×和 4×表示，图中阴影部分模块 subsampled、Densenet3、Convs 和 Add 为改进部分。

由于 YOLOv3 网络的第 2 块残差块 2×的特征图含有更多小目标的位置信息，本文将 2×的输出特征图进行 2 倍下采样 subsampled，然后用改进的密集卷积网络 Densenet3 对 subsampled 操作后的特征图进行特征加强，经 Densenet3 操作后的输出特征图通过 1 × 1 的卷积(Convs)调整通道数，最后，再将 Convs 操作后的输出特征图与第 4 块残差块 8×经上采样后的输出特征图融合，从而使小型目标预测特征图在获得更多人脸语义特征的情况下，保留更详细的小尺度人脸的位置信息。

原网络采用第 3 块残差块 8×输出的特征图对小目标进行检测，即图 1 中的尺度 Scale3，本文对尺度 Scale3 上的特征图的拼接方式进行调整，将特征堆叠(Concate)改为特征叠加(Add)，见公式(1)和公式(2)。

$$S_{Concat}=\sum_{i=1}^{C}\mathrm{Conv}\left(X_i,K_i\right)+\sum_{i=1}^{C}\mathrm{Conv}\left(Y_i,K_{C+i}\right) \tag{1}$$

$$S_{Add}=\sum_{i=1}^{C}\mathrm{Conv}\left(\left(X_i+Y_i\right),K_i\right)=\sum_{i=1}^{C}\mathrm{Conv}\left(X_i,K_i\right)+\sum_{i=1}^{C}\mathrm{Conv}\left(Y_i,K_{C+i}\right) \tag{2}$$

Concat 和 Add 的单个通道的输出计算如公式(1)和(2)，式中 $X_i,Y_i\left(i=1,\cdots,C\right)$ 表示输入的特征，Conv() 表示卷积操作，C 表示通道数，K_i 表示卷积核。由文献[13]知，通过采用 Add 操作，使网络模型获得更丰富的人脸语义特征信息。

原网络利用第 4 个残差块 8×的输出特征图用于中等目标的检测，即图 1 中的尺度 Scale2，为了调整特征图的通道数，本文将 Scale2 上的 1 × 1 的卷积层去除，使网络在一定程度上减少了网络模型的计算量。删除的 1 × 1 卷积层位置如图 1 虚线处。

2.2. 密集卷积模块 Densenet3 的优化设计

密集卷积网络(DenseNet)是 Huang 等[14]人提出的一种网络特征复用，避免梯度消失的方法，该方法借鉴了 ResNet [15]与 GoogLeNet [16]思想，其核心思想是通过特征重用和旁路设置，将模块内任一卷积

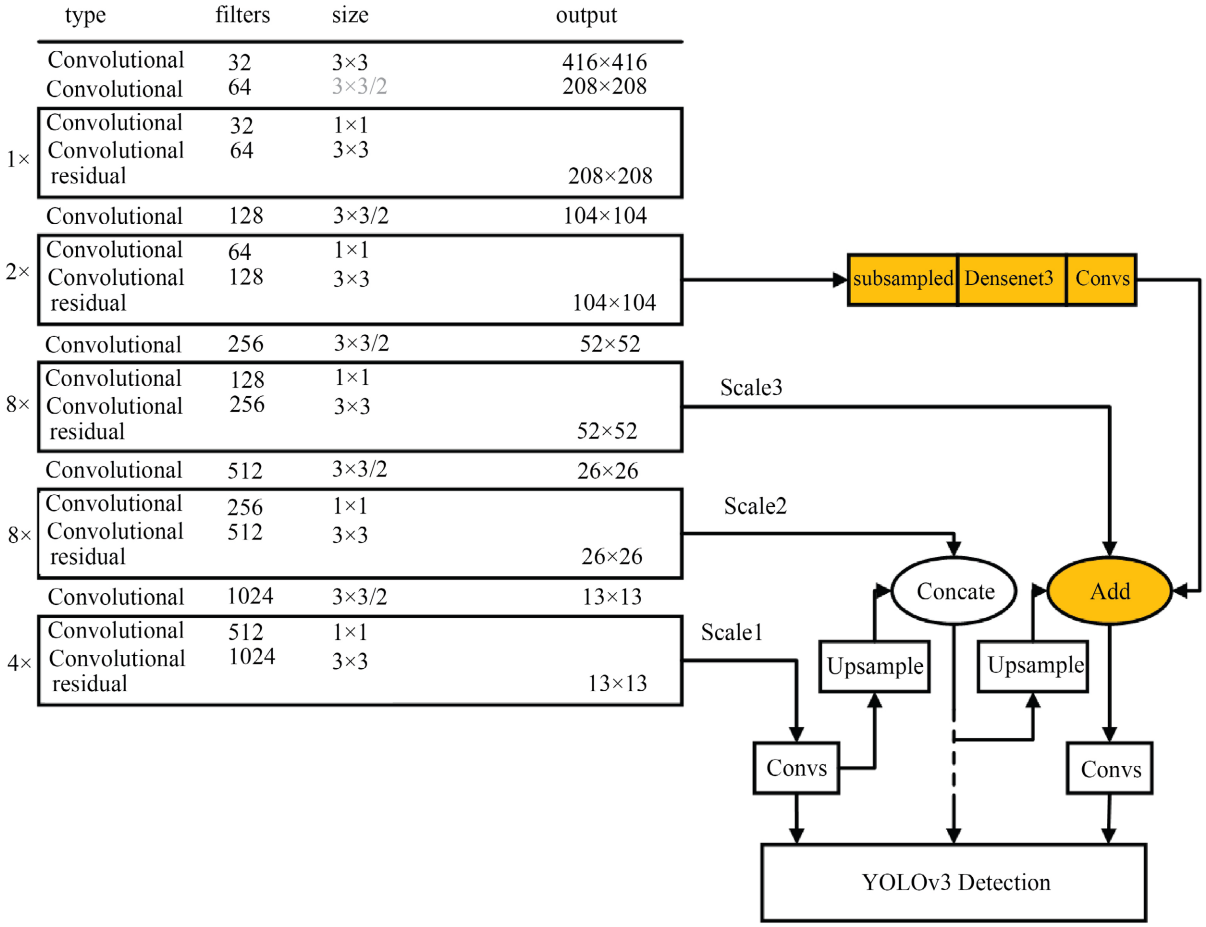
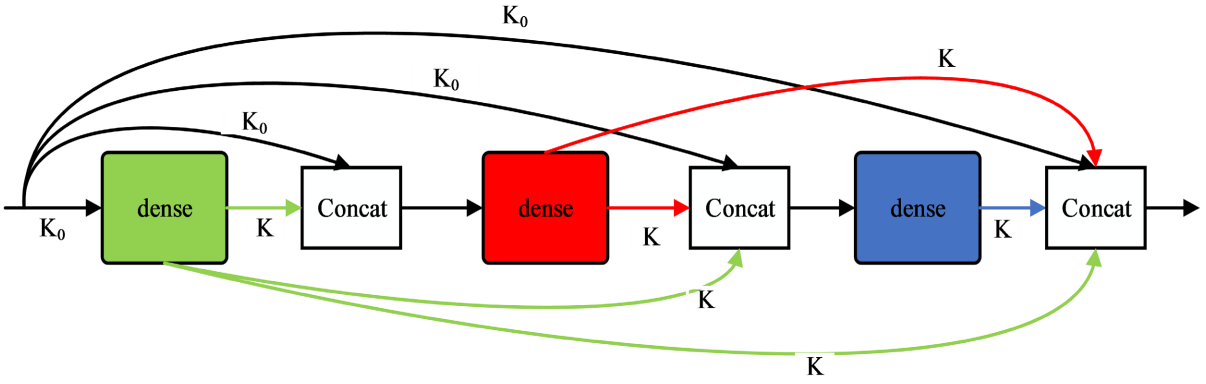


Figure 1. Improved YOLOv3 network structure
图 1. 改进的 YOLOv3 网络结构

层的输入包含前面所有卷积层的输出，使特征得到充分复用的同时防止出现特征消失，从而提高网络的性能。本文采用了 DenseNet-B 结构[15] (如图 2 所示)，其 K_0 表示网络的输入通道数，K 表示通道增长率。每一个 Dense 单元由 1 × 1 的卷积核和 3 × 3 的卷积核以及卷积核所对应的激活函数和批归一化层堆叠而成。相比较于另外两种 DenseNet 结构，该网络通过选用较小的通道增长率 K 既能避免在融合各个通道特征时使网络变得很宽，又能在提取特征时降低维度减少网络的计算量。



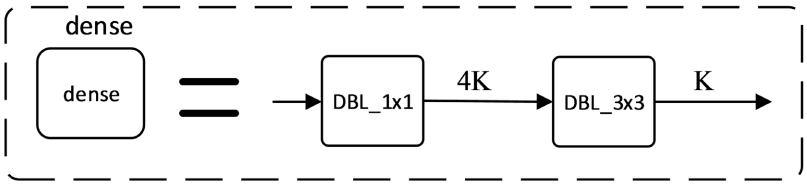


Figure 2. Dense convolution network (The figure on the top shows the dense convolution network structure, and the figure on the bottom shows the unit structure of dense)
图 2. 密集卷积网络(上图为密集卷积网络结构，下图为 dense 的单元结构组成)

针对小尺度人脸检测所占像素少，覆盖面积小等特点，本文对密集卷积网络进行改进，对 DenseNet-B 基本结构中特征融合方式进行了调整，将张量拼接(Concate)方式改为张量叠加(Add)，为了满足通道融合条件，改进密集卷积网络须保证输入通道数和通道增长率一致。改进的密集卷积网络采用并行策略将不同感受野的特征图中的特征向量组合成复向量，以此获得高层特征图和低层特征图的语义信息，从而丰富特征图的小尺度人脸检测特征信息，最终有效提高了模型检测的效果。其改进密集卷积网络图如图 3 所示。

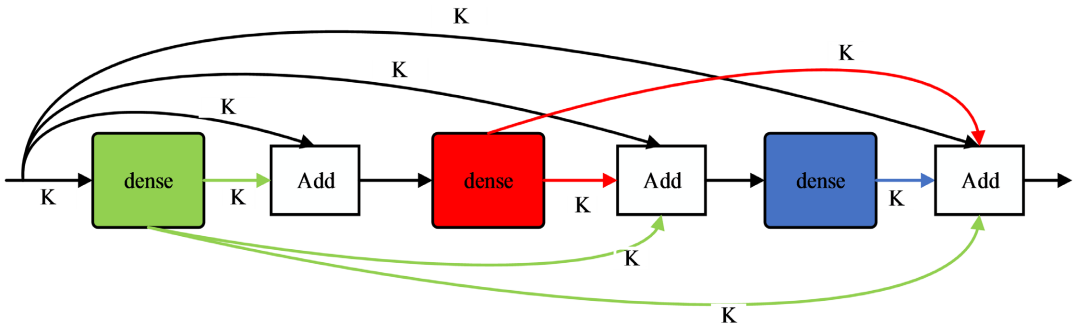


Figure 3. Improved dense convolution network
图 3. 改进的密集卷积网络

3. 实验设计

3.1. 数据集选择

本文将在 WIDER FACE [17]数据集上进行训练以及测试验证。该数据集是当前已知人脸数据集中人脸数量最多、检测环境最为复杂数据集之一。该数据集含图像 3 万多张，其中训练集和验证集图像占比 50%，测试集图像占比 50%，其所有图像的人脸标注数量接近 40 万个，数据集平均每张图像 13 张人脸。由于数据集的人脸数量多且较为密集更加符合本文数据集的选取要求。

实验中考虑到 WIDER FACE 数据集中测试集未公开其人脸标注的标签，选取 WIDER FACE 数据集中的占比总数据 40%的训练集(Training)和占比总数据 10%的验证集(Validation)为模型数据集。选取的人脸数据集经预处理转换成 VOC 格式的局部数据集，共包含 16,102 张图片，包含标注的人脸有 196,144 张，然后将训练集和测试集以 9:1 的比例进行划分，得到训练集 14,491 张图像，包含 174,489 张人脸；测试集 1611 张图像，包含 21,655 张人脸。由于数据集中的人脸种类齐全，且存在尺寸、遮挡、光照、表情等多种影响因素，比较符合密集人群场景下人脸的复杂环境，同时可以更好地验证模型的泛化性。

3.2. 数据集选择

对于模型的性能评价需要相对客观的一些指标去度量。本文采用精度(Precision)、召回率(Recall)、平

均精度均值(mAP)、AP 以及 F_1 值五个重要指标对人脸检测模型评价，同时还考虑到模型是否具有良好的实时性，其评价指标的定义以及公式如下：

$$P = \frac{S_{TP}}{S_{TP} + S_{FP}} \times 100\% \tag{3}$$

$$R = \frac{S_{TP}}{S_{TP} + S_{FN}} \times 100\% \tag{4}$$

公式(3)中 P 表示人脸的精度，即本次预测正确的人脸数占预测人脸数的百分比，S_{TP} 代表标注为人脸且被正确预测的个数，S_{FP} 表示未被标记为人脸而被预测为人脸的个数，公式(3)中的 $S_{TP} + S_{FP}$ 为所有的预测结果为人脸的个数。公式(4)中 R 表示人脸的召回率，即本次预测正确的人脸数占测试集数据中标注的总人脸数的百分比，也就是标注人脸数被模型成功预测的个数，S_{FN} 表示测试集中标注人脸但未被正确预测的个数，公式(4)中的 $S_{TP} + S_{FN}$ 为测试集中的标注了的人脸总个数。

$$F_1 = \frac{2}{\frac{1}{P} + \frac{1}{R}} = 2 \times \frac{PR}{P + R} \tag{5}$$

$$AP = \int_0^1 P(R)\,dR \tag{6}$$

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N} \tag{7}$$

上式(5)中，F_1 值表示人脸的准确率与召回率之间的调和均值，上限为 1，下限为 0，模型的调和均值越接近 1，差异性越小，从而模型性能越好。在目标检测中，准确率与召回率关系往往是相互制约的，一方升高，则另一方就低，为了更好地权衡平均精度均值与召回率，本文引入人脸检测的 AP 指标，即 P-R 曲线其下方面积为 AP 的值，如公式(6)所示。公式(7)中 mAP (mean average precision)表示的是人脸的预测平均精度的均值，由于在本实验中只有人脸一个类别，这里的 mAP 值就等于 AP 值，同时 mAP 是当前目标检测中最重要的模型衡量指标之一。

4. 实验结果及分析

本文在 Windows10 操作系统下完成软件平台的搭建和实验，GPU 选用 NVIDIA GeForce RTX3060，软件工具：Anaconda3；python 版本 3.6；Pytorch-gpu ≥ 1.6.0。

4.1. 训练结果

本文在实验准备阶段，为了加快模型训练进程，减少时间消耗，采用迁移学习的方式，将包含 person 类特征的 coco 数据集上的 YOLOv3 的权重作为本次实验的预训练权重，同时为了进一步加快人脸特征的学习速度，本实验室采用冻结 YOLOv3 网络除分类层以外的主干层的方式，并设置超参数学习率和迭代次数，分别为 $2e^{-3}$ 和 50 个 epoch，如图 4 所示，由于前 50 个 epoch 只训练分类层是否为人脸，原模型和改进后的 YOLOv3 损失下降明显，且后者的下降速度要比前者更快。通过冻结主干的方式训练网络之后得到的权重用于正式训练，将正式训练设置为 51 个 epoch 到 150 个 epoch 和 151 个 epoch 到 500 个 epoch 两个阶段，起始训练的学习率为 $2e^{-4}$。从图中可以得出损失在第一个阶段依旧保持下降态势，但比准备阶段下降缓慢，后 350 个 epoch 改进的 YOLOv3 网络和原网络损失下降非常缓慢，到第 300 个 epoch 后，改进的 YOLOv3 网络损失值稳定在 0.41 左右，原网络模型的损失值维持在 0.55 左右。由此可见改进后的网络模型的收敛速度更快，损失值更小。

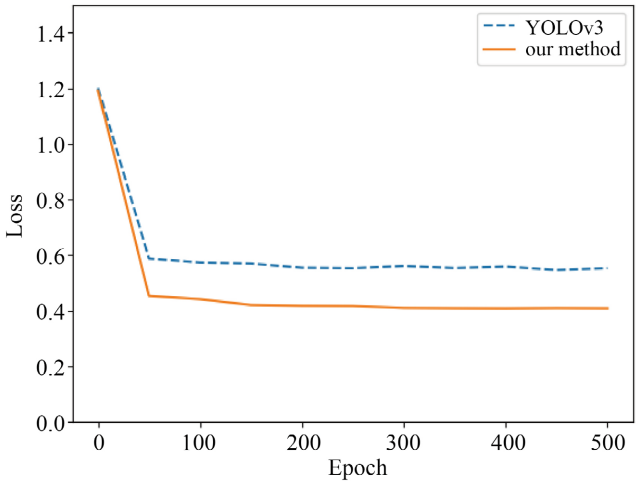


Figure 4. Training loss curve of YOLOv3 model before and after improvement
图 4. 改进前后 YOLOv3 模型的训练损失曲线

4.2. 算法检测指标分析

由于测试集中只存在一部分图像的人脸数多，本文对测试集进行筛选，选取测试集中的密集小人脸的图像 100 张，包含 3234 张人脸，将挑选的测试集用于模型性能的测试。同时，本文将实验测试交并比(IoU)的阈值设定为 0.5。

Table 1. Comparison of detection indexes of small face in dense population
表 1. 密集人群小人脸检测指标对比

Detection algorithm	Precision/%	Recall/%	mAP/%	F_1/%	Average time/s
YOLOv3	75.72	72.26	72.54	74	0.021
Our-YOLOv3	78.12	73.49	74.20	76	0.022

密集小人脸检测结果对比如表 1 所示，与原网络相比，改进后的 YOLOv3 网络检测速度下降了 0.001 s，但小人脸的检测准确率由 75.73%提高到 78.12%，召回率由 72.26%提高到 73.49%，F_1 值由 74%提高到 76%，mAP 值由 72.54%提高到 74.20%，实验结果表明改进后 YOLOv3 网络对密集人群的小人脸检测性能有较好的提升。

两种网络计算密集人群的小尺度人脸检测的平均精度(AP)曲线，如图 5 所示，左图为原网络的平均精度(AP)曲线、右图为改进的 YOLOv3 网络的平均精度(AP)曲线，平均精度融合精确率和召回率两者的优势来衡量检测算法的性能，由图 6 对比可知，与原网络相比，改进的 YOLOv3 网络在密集人群小尺度人脸的检测性能上，平均精度由 72.54%提高到了 74.20%，从而反映了对于结构上的微调和引入改进密集卷积网络的 YOLOv3 在小人脸的召回率和精度上都有一定的提升。

4.3. 检测结果分析

改进后的网络和原始 YOLOv3 算法的测试效果对比如图 6 所示，第一列展示了原 YOLOv3 网络的人脸检测结果，第二列展示的是改进后的 YOLOv3 网络的人脸检测结果，由图 6 可知右边图片检测的人脸数要比左边图片检测到的人脸数更多，表明在测试集中改进的 YOLOv3 算法可以学习到更多的人脸特征，

具有更高的检测精度；同时通过图片所具有的人脸数量，验证了本文实验在密集人群下的小尺度人脸检测的人脸检测，改进的 YOLOv3 网络比原网络有更好的分类性，其检测效果更好。

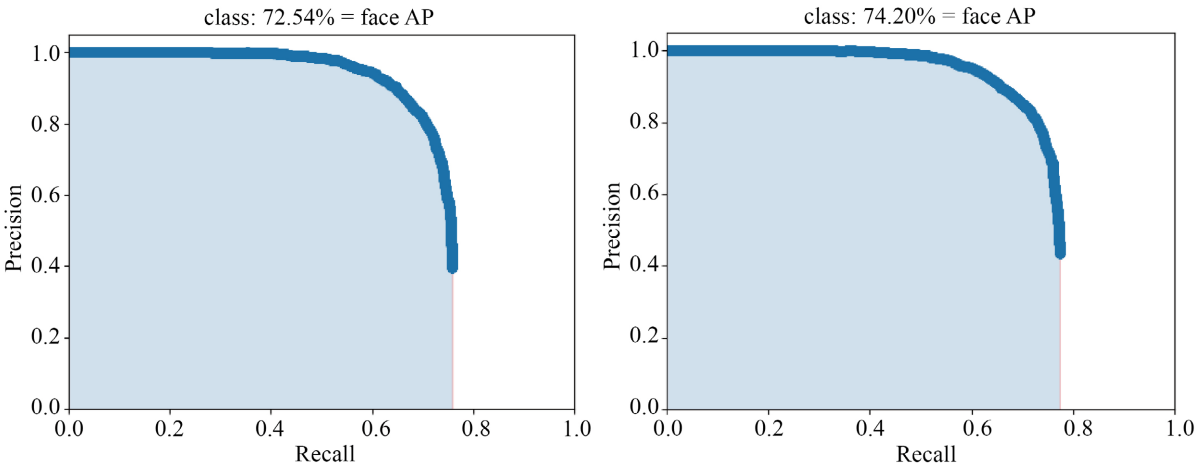


Figure 5. AP curves of the two networks (The left figure is the AP curve of the original YOLOv3 network, and the right figure is the AP curve of the improved YOLOv3 network)
图 5. 两种网络的 AP 曲线(左图为原 YOLOv3 网络的 AP 曲线，右图为改进后 YOLOv3 网络的 AP 曲线)

Figure 6. Visual comparison between YOLOv3 algorithm and improved YOLOv3 face detection results
图 6. YOLOv3 算法与改进 YOLOv3 人脸检测的结果可视化对比

5. 结论

针对密集人群图像的人脸目标检测，普遍存在小尺度人脸目标特征少(不足)等问题，本文对 YOLOv3 网络做了如下改变，首先通过对小型物体预测特征尺度前的特征融合方式进行调整，使网络获得更多的

小尺度人脸检测特征信息，从而达到提升小尺度人脸检测的检测召回率的效果；其次为了进一步丰富小型预测尺度的语义信息，本文采用原网络的第 2 块残差块输出特征图，进行多次特征图操作(主要是 2 倍下采样卷积和经改进的密集卷积模块操作获得不同感受野的融合)，最后将多次特征图操作后的输出特征图与原网络的第 4 块残差块经上采样后的输出特征图融合，实现不同网络层次的语义信息的特征融合，从而使网络获取到更多的小尺度人脸检测信息特征，提高人脸检测的效果。同时改进的 YOLOv3 算法如果未来应用工程实践，模型计算量、精确率、网络结构将是其主要研究方向。

参考文献

[1] 徐光柱, 屈金山, 雷帮军, 刘鸣, 石勇涛. YOLO 和分块-融合策略结合的稠密人脸检测方法[P]. 中国专利, CN112541483A. 2021-03-23.

[2] Rowley, H.A., Baluja, S. and Kanade, T. (1998) Neural Network-Based Face Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**, 23-38. https://doi.org/10.21236/ADA341629

[3] Rowley, H.A., Baluja, S. and Kanade, T. (1998) Rotation Invariant Neural Network-Based Face Detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Santa Barbara, 25 June 1998, 38-44. https://doi.org/10.21236/ADA341629

[4] Viola, P. and Jonens, M. (2001) Rapid Object Detection Using a Boosted Cascade of Simple Feature. *IEEE Computer Society Conference on Computer Vision & Pattern Recognition*, Kauai, 8-14 December 2001, 511.

[5] Viola, P. and Jones, M.J. (2004) Robust Real-Time Face Detection. *International Journal of Computer Vision*, **57**, 137-154. https://doi.org/10.1023/B:VISI.0000013087.49260.fb

[6] Mathias, M., Benenson, R., Pedersoli, M., *et al*. (2014) Face Detection without Bells and Whistles. In: *European Conference on Computer Vision*, Springer, Cham, 720-735. https://doi.org/10.1007/978-3-319-10593-2_47

[7] Li, H., Lin, X., *et al*. (2015) A Convolutional Neural Network Cascade for Face Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, 7-12 June 2015, 5325-5334. https://doi.org/10.1109/CVPR.2015.7299170

[8] Zhang, K., Zhang, Z., Li, Z., *et al*. (2016) Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, **23**, 1499-1503. https://doi.org/10.1109/LSP.2016.2603342

[9] Wang, H., Li, Z., Ji, X., *et al*. (2017) Face R-CNN.

[10] Jiang, H. and Learned-Miller, E. (2017) Face Detection with the Faster R-CNN. 12*th IEEE International Conference on Automatic Face & Gesture Recognition*, Washington DC, 30 May-3 June 2017, 650-657. https://doi.org/10.1109/FG.2017.82

[11] Liu, W., Anguelov, D., Erhan, D., *et al*. (2016) SSD: Single Shot Multibox Detector. In: Leibe, B., Matas, J., Sebe, N., *et al*., Eds., *Lecture Notes in Computer Science*, Springer, Cham, Vol. 9905, 21-37. https://doi.org/10.1007/978-3-319-46448-0_2

[12] Redmon, J. and Farhadi, A. (2018) YOLOv3: An Incremental Improvement. https://arxiv.org/abs/1804.02767

[13] 赵柳, 陆军, 刘杨. MAEA-DeepLab: 具有多特征注意力有效聚合模块的语义分割网络[J]. 中国科学技术大学学报, 2020, 50(8): 1170-1180.

[14] Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q. (2018) Densely Connected Convolutional Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 18-23 June 2018, 4700-4708.

[15] He, K.M., Zhang, X.Y., Ren, S.Q., *et al*. (2016) Deep Residual Learning for Image Recognition. In: *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE Press, Washington DC, 770-778. https://doi.org/10.1109/CVPR.2016.90

[16] Szegedy, C., Liu, W., Jia, Y., *et al*. (2015) Going Deeper with Convolutions. In: *Proceedings of the* 2015 *Conference on Computer Vision and Pattern Recognition*, IEEE Press, Washington DC, 1-9. https://doi.org/10.1109/CVPR.2015.7298594

[17] Yang, S., Ping, L., Chen, C.L., *et al*. (2016) Wider Face: A Face Detection Benchmark. *IEEE Conference on Computer Vision & Pattern Recognition*, Las Vegas, 27-30 June 2016, 5525-5533. https://doi.org/10.1109/CVPR.2016.596