计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(2):154−164 [doi: 10.15888/j.cnki.csa.009767] [CSTR: 32024.14.csa.009767]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于广义高效层聚合网络和共享卷积的卡通角色面部检测①

闫博文, 刘永泽, 夏海东, 宋晓强

(石家庄铁道大学 信息科学与技术学院, 石家庄 050043)
通信作者: 刘永泽, E-mail: liuyongze@stdu.edu.cn

摘　要: 卡通角色面部检测是一项比人脸检测更具挑战性的任务, 它涉及许多困难的场景. 针对卡通角色面部间存在巨大差异的特点, 本文提出了一种卡通角色面部检测算法, 命名为 YOLO-DEL. 首先, 基于 GELAN 融合 BDD 设计了 DBBNCSPELAN 模块, 旨在减小模型体积的同时增强检测性能. 接下来, 引入一种称为 ELA 的多尺度注意机制, 用于改善 SPPF 结构, 增强主干模型的特征提取能力. 最后, 设计了新的共享卷积检测头, 使网络更轻便. 同时也用 Shape-IoU 代替原 CIoU 损失函数, 提升模型的收敛效率. 在 iCartoonFace 数据集上进行实验, 通过消融实验验证得到的模型, 并将其与 YOLOv3-tiny、YOLOv5n 和 YOLOv6 等模型进行比较. 改进模型 YOLO-DEL 的 *mAP* 达到 90.3%, 比 YOLOv8 提高了 1.2%, 参数量为 1.69M, 与 YOLOv8 相比参数量降低 47%, GFLOPs 降低 44%. 实验表明, 本文方法能有效提高卡通角色面部的检测精度, 同时缩小网络模型的大小, 验证本文方法的有效性.

关键词: 目标检测; 卡通面部; GELAN; 注意力机制; YOLOv8; 共享卷积

引用格式: 闫博文,刘永泽,夏海东,宋晓强.基于广义高效层聚合网络和共享卷积的卡通角色面部检测.计算机系统应用,2025,34(2):154–164. http://www.c-s-a.org.cn/1003-3254/9767.html

Cartoon Character Face Detection Based on Generalized Efficient Layer Aggregation Network and Shared Convolution

YAN Bo-Wen, LIU Yong-Ze, XIA Hai-Dong, SONG Xiao-Qiang

(School of Information Science and Technology, Shijiazhuang Tiedao University, Shijiazhuang 050043, China)

Abstract: Cartoon character face detection is more challenging than face detection because it involves many difficult scenarios. Given the huge differences between different cartoon characters' faces, this study proposes a cartoon character face detection algorithm, named YOLOv8-DEL. Firstly, the DBBNCSPELAN module is designed based on GELAN fusion BDD to reduce model size and enhance detection performance. Next, a multi-scale attention mechanism called ELA is introduced to improve the SPPF structure and enhance the feature extraction ability of the backbone model. Finally, a new detection head for shared convolution is designed to make the network lighter. At the same time, the original CIoU loss function is replaced by Shape-IoU to improve the convergence efficiency of the model. Experiments are carried out on the iCartoonFace dataset, and ablation experiments are carried out to verify the proposed model. Besides, the proposed model is compared with the YOLOv3-tiny, YOLOv5n, and YOLOv6 models. The *mAP* of the improved model YOLO-DEL reaches 90.3%, 1.2% higher than that of YOLOv8. The parameters amount is 1.69M, 47% lower than that of YOLOv8. The GFLOPs value is 44% lower than that of YOLOv8. Experimental results show that the proposed method effectively improves cartoon character face detection precision while compressing the network model's size. Thus, the proposed method has proved to be effective.

Key words: object detection; cartoon face; GELAN; attention mechanism; YOLOv8; shared convolution

① 基金项目: 河北省自然科学基金面上项目 (F2019210253)
收稿时间: 2024-06-17; 修改时间: 2024-07-10, 2024-08-20; 采用时间: 2024-09-03; csa 在线出版时间: 2024-12-13
CNKI 网络首发时间: 2024-12-13

近年来, 在工业应用的强烈需求的推动下, 卡通媒体受到越来越多的关注. 作为理解这一媒体的第 1 步, 卡通角色面部检测是一个至关重要的任务. 卡通面孔是理解和与虚拟世界互动的重要组成部分, 准确检测这些卡通人物是许多视觉应用的必要前提, 如自动编辑、拍摄、广告推荐和计算机辅助建模. 艺术家根据现实世界的抽象概念创造和想象卡通人物, 因此创造出来的面孔与人类的面孔有很多相似之处. 漫画图像与真人肖像有很强的相似性, 但夸大了某些特定的面部特征. 在虚拟媒体和卡通视频中, 大多数角色都表现出夸张或幽默的面部表情, 这给检测任务带来了新的挑战.

在深度学习成为目标检测领域主流方法之前, 传统卡通角色面部检测方法也有过一定的发展. Matsui 等人[1]在 2017 年提出了一个基于草图的漫画检索系统和新颖的查询方案, 并建立了一个新的漫画图像数据集 Manga109. Kohei 等人[2]提出了一种基于特征提取的卡通人物人脸检测与人脸识别方法, 通过利用肤色、边缘、下颌轮廓、对称性等特征对卡通人物进行人脸检测的方法.

在基于深度学习的检测方法出现之后, 其优势很快体现了出来, 并得到了业内相关研究人员的注意. 基于深度学习的目标检测算法大致可以分为两类: 两阶段算法和单阶段算法. 在两阶段检测算法研究中, 傅俊成[3]运用不同的特征提取网络对卡通人物进行特征提取, 并对原始 Faster R-CNN 算法进行改进, 融合特征金字塔网络, 以及使用 Mask R-CNN 的 ROI Align 代替 ROI Pooling, 对网络进行改进. 结果表明, 融合特征金字塔网络后的改进模型解决了人物尺度差异大以及小目标识别困难的问题, 在识别准确率上有较大的提升. 夏华丽[4]运用特征提取网络 VGG16, 并对原始 Faster R-CNN 算法进行了改进, 替换 VGG16 为 ResNet50 网络, 再融合特征金字塔网络对其进行改进. 对比分析各类别人物识别结果表明, 融合特征金字塔网络后的改进模型在识别准确率上有较大的提升.

在单阶段检测算法研究中, 张健[5]提出了一种基于 SSD 的动漫人物识别方法. 作者在增加了正样本损失加权的基础上, 去除了 Conv4-3 检测模块的模型, 在保证检测精度的情况下减小了模型体积. Ogawa 等人[6]提出了一种基于 SSD 的高度重叠对象的检测方法 SSD300-fork, 首先解释了由高度重叠的对象引起的赋值问题. 然后提出了 SSD300-fork 来解决分配问题. 在单阶段检测算法研究中, YOLO 系列凭借其轻量化的模型和检测速度快的性能受到广泛关注. Nguyen 等人[7]提出了一种基于 YOLOv2 网络模型来对漫画角色进行检测, 并制作了一个大型漫画数据集 Sequencity 612. 陈争光[8]提出了一种基于改进 YOLOv3 的卡通头像检测方法. 为了更好地重利用卡通角色的特征, 文章在 YOLOv3 的特征提取网络 Darknet-53 的基础上引入 Dense block. 通过加入一个 Dense block, 以此来提高主干网络对图像中特征的提取能力. 之后, 引入一种高效的特征融合手段基于双向采样的特征融合金字塔 BiFPN, 在融合特征的同时, 还可以精炼锚框中的感受野. Topal 等人[9]提出了一种绘图中的域自适应自监督人脸与身体检测方法. 作者选用了 YOLOX 作为检测模型并探讨了风格迁移和自监督训练的影响. 但以上模型均存在精度低或者模型臃肿的问题.

为解决上述问题, 本文提出了一种基于 YOLOv8 的卷积神经网络模型 YOLO-DEL, 改进后的算法对常规大小物体和复杂场景中的小物体的检测精度有一定提高, 同时, 降低了模型的复杂度. 使用该数据集对 YOLOv3、YOLOv5、YOLOv6、YOLOv8、SSD 和 Faster R-CNN 等网络模型进行了广泛的对比实验. 结果表明, 本文提出的 YOLO-DEL 算法显著优于其他方法, 为卡通角色面部的检测提供了技术支持.

本文的主要贡献如下.

(1) 引入 DBBNCSPELAN 模块, 通过模块的堆叠和各个小模块的不断融合, 形成一个新的网络结构. 这增加了结构的整体深度, 以更低的计算成本实现了更高的分辨率, 并捕获了更多的上下文信息.

(2) 引入共享卷积思想, 对 YOLOv8 的检测头进行改进. 设计轻量级检测头 LWSH, 在损失精度有限的情况下换取机制的模型体积缩减.

(3) 将 YOLOv8 中主干末尾的 SPPF 模块替换为的 SPPF-ELA 模块. 通过 SPPF 模块与 ELA 注意力的融合, 形成一个新的空间金字塔结构, 确保了特征提取过程中上下文信息的更全面保存.

(4) 将 YOLOv8 的 CIoU 损失函数替换为 Shape-IoU[10]损失函数, 提高模型的收敛速度和检测精度.

1 目标检测网络

1.1 YOLOv8 目标检测算法

YOLOv8[11]是一种基于深度学习的目标检测算法, 具有可优化和改进的网络结构, 与以前的版本相比, 提高了准确性和检测效率. YOLOv8 是一种高效且准确

的目标检测算法, 如图 1 所示, 具有检测速度快、精度高的优点, 适用于各种目标检测场景的网络结构. YOLOv8 主要由以下部分组成: (1) 输入层: 输入层接收原始图像数据, 并且在 Mosaic 数据增强, 图像数据是通过缩放、裁剪和排列缝合在一起, 然后经过归一化和预处理, 数据被传递到下一层. (2) 骨干网络: 它由卷积层、C2f 模块和 SPPF 组成, 单元卷积层主要用于提取来自图像的特征信息. YOLOv8 网络采用多个卷积层和残差块结构, 可以有效提高学习能力网络的准确性. C2f 模块跨按更多分支分层, 并调整不同数量的不同比例模型的通道, 这不仅确保轻量级的同时还能获得更丰富的梯度流信息. 空间金字塔池化 SPPF 结构转换特征将任意大小的特征向量映射为固定大小的特征矢量. (3) 特征融合层: 由 C2f 模块、上采样层和 Concat 模块组成. 上采样层是主要用于放大特征图以获得更好的小目标检测效果. (4) 检测层: 检测层是 YOLOv8 网络的输出层, 主要用于检测和定位图像. 使用多个检测层, 每次检测层可以输出多个检测结果. 损失函数用于计算网络预测结果和真实标签以优化网络参数. YOLOv8 使用 VFL Loss 作为分类损失, DFL 损失+CIoU 作为回归损失. VFL 使用非对称加权运算来求解正负样本不平衡的问题. DFL 是在标签处优化的交叉熵区域的形式最接近左右位置, 从而使网络能够更快地关注邻近分布的目标位置的区域.

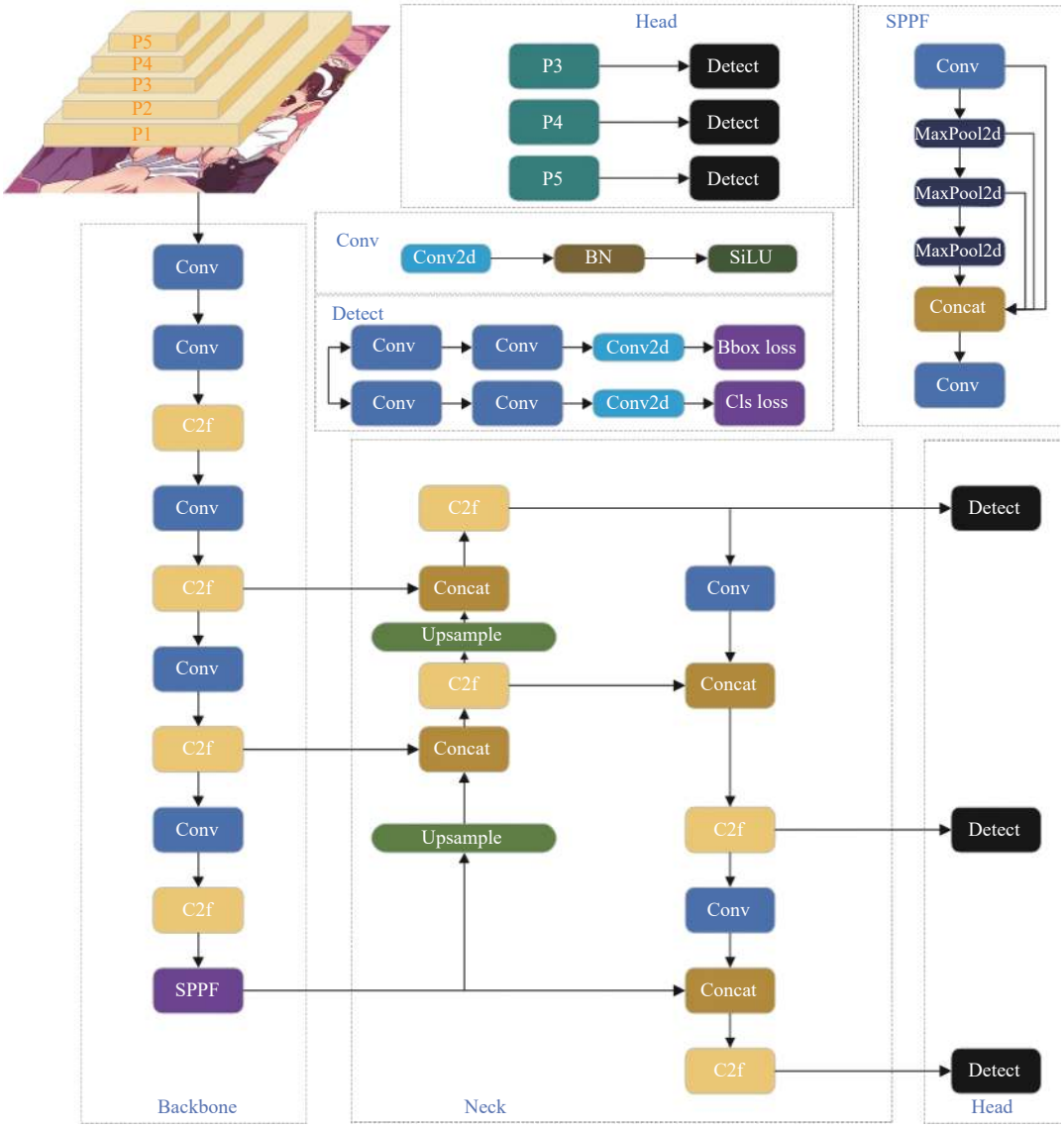

图 1 YOLOv8 结构图

1.2 广义高效层聚合网络

新网络架构 GELAN (generalized efficient layer aggregation network)[12]通过结合 CSPNet 和 ELAN 这两种采用梯度路径规划设计的神经网络架构, 设计了兼顾轻量级、推理速度和准确性的广义高效层聚合网络 (GELAN) . 它的整体架构如图 2 所示. 将最初仅使用卷积层堆叠的 ELAN 的能力推广到可以使用任何模块的新架构. 在 YOLOv9 中, 使用 RepNCSPELAN4 作为特征提取模块, 从名字上知, 该模块是一个 Rep+CSP+ELAN 的组合网络.

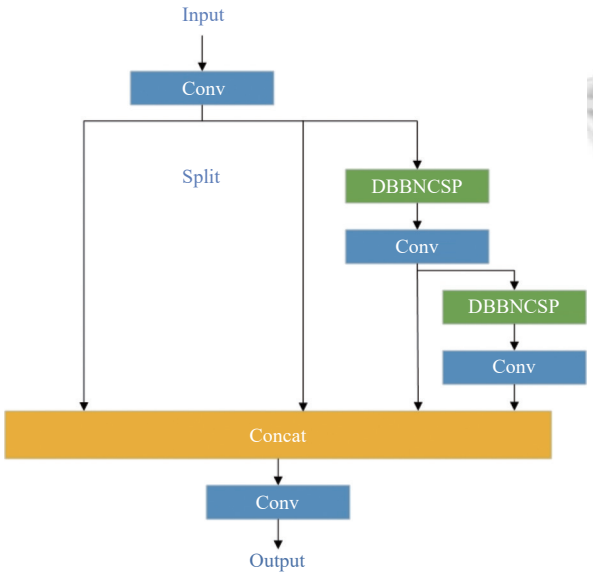

图 2 RepNCSPELAN4 结构图

而 RepConvN, 是在重参数化卷积的基础上去掉了恒等连接. 将重参数化卷积应用到残差模块或者用到基于拼接的模块中去. 但是在代码中使用了最简单的重参数化卷积, 并没有使用提出的这个结论. 该思想取自于 RepVGG, 基本思想是在训练的时候引入特殊的残差结构辅助训练, 这个残差结构是经过独特设计的, 在实际预测的时候, 可以将复杂的残差结构等效于一个普通的 3×3 卷积, 这个时候网络的复杂度下降, 但是网络的预测性能却没有下降. 因为残差网络本身存在恒等连接, 而原本的重参数化卷积 RepConv 也有恒等连接, 两者之间起了冲突, 所以要去掉原本重参数化卷积 RepConv 中的恒等连接, 成为 RepConvN.

RepConvN 结构如图 3 所示, 在推理时, 会将多个卷积重参数化, 转换成一个卷积操作. 在此基础上, 可以得到 RepNBottleneck 与 RepNCSP, 如图 4 和图 5 所示, 进而得到 RepNCSPELAN4.Rep 优化计算, CSP 丰富梯度、减少冗余、降低计算量, ELAN 做高效的特征聚合.

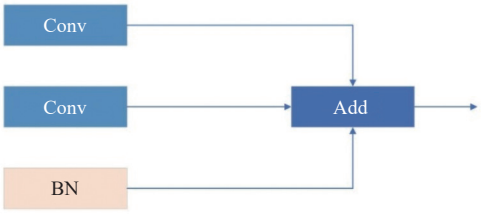

图 3 RepConvN 结构图

图 4 RepNBottleneck 结构图

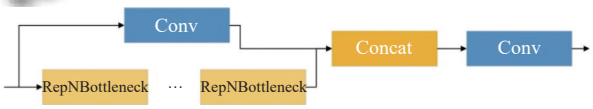

图 5 RepNCSP 结构图

2 改进算法网络结构

骨干、颈部和检测头是用于目标检测的典型网络组件. 虽然 YOLOv8 算法已经相当高效, 但它对卡通人脸的检测性能并不理想. 这是由于从检测网络中提取的特征及其质量相对较低, 并且没有专门针对该领域的设计. 在这项工作中, 我们对原来的 YOLOv8 网络进行了升级, 图 6 展示了提出的网络架构.

YOLO-DEL 的整体架构由几个关键组件组成, 利用 DBB 主干进行鲁棒特征提取, 在空间金字塔池化结构融入 ELA 注意力机制, 并集成了 LWSH 检测头和 Shape-IoU 损失函数, 增强了整体目标检测能力.

2.1 DBBNCSPELAN 模块

多元分支模块 (diverse branch block)[13]的核心思想是结合不同规模和复杂度的多分支结构来丰富卷积块的特征空间. 多分支构造包括卷积序列、多尺度体积和平均池化, 增强了对单个卷积的表示能力, 有效强化了 CNN 的核心骨架. DBB 工作的示意图如图 7 所示.

多分支网络结构具有不同规模的感受野. 网络的宽度和数量与普通网络结构相比, 增加了参数, 有利于提高性能. 但是, 随着网络的多分支化, 会大大增加网络的内存消耗和参数的数量. 利用结构重参数化方法, 可以将复杂的多分支网络无损失地转化为单分支网络. 这样, 转换后的网络不仅具有转换前的优良性能, 而且结构简单.

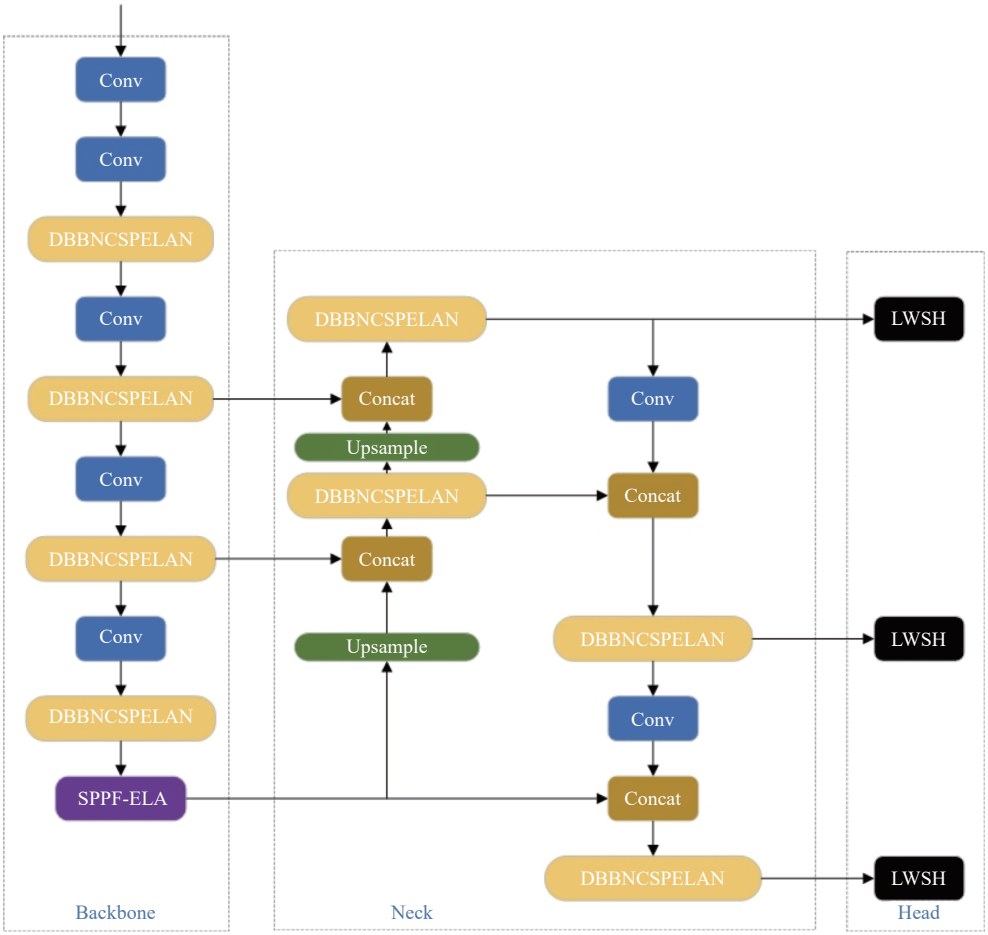

图 6 YOLO-DEL 结构图

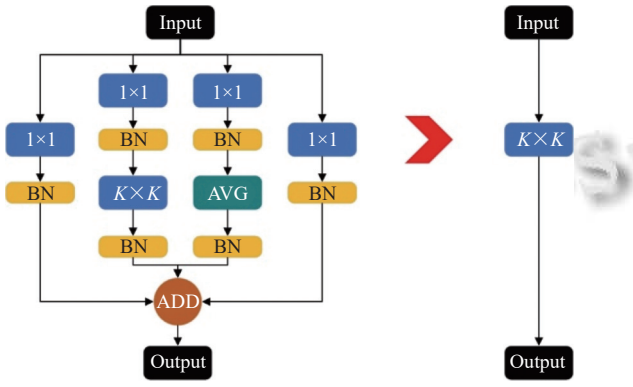

图 7 DBB 结构图

卷积层和平均池化层之后是 BN 层. 为了将复杂的多分支结构转化为单分支结构, 涉及多个融合步骤, 包括 BN 层与层间的体积融合、不同大小卷积的融合等.

将 DBB 融入 GELAN 之中, 使用 DBB 替换 RepConvN, 进而得到 DBBNBottleneck 与 DBBNCSP, 如图 8 与图 9 所示, 最终得到 DBBNCSPELAN.

DBBNCSPELAN 架构的设计目标是保持模型的轻量化, 同时增强其对小目标的检测性能. 通过将 GELAN 与 DBB 相结合, 生成更全面的特征表示, 并提高模型的通用性, 从而使检测性能更加优越. DBB 模块增强了模型的可泛化性, 从而增强了模型的鲁棒性, 同时减少了参数量. 这种调整可以更轻松地提高其实时性能. DBBNCSPELAN 结构如图 10 所示.

图 8 DBBNBottleneck 结构图

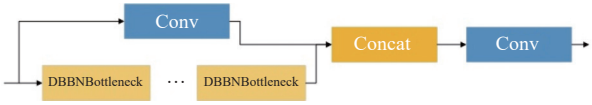

图 9 DBBNCSP 结构图

2.2 轻量级共享检测头

批归一化 (batch normalization, Batch Norm 或 BN) 已经被认为是深度学习中非常有效的组成部分, 在很大程度上有助于推动计算机视觉领域的发展以及其他

领域. BN 通过在小批内计算的均值和方差对特征进行归一化, 许多实践已经证明了这一点. 可以简化优化并使非常深的网络收敛. 批统计数据的随机不确定性也可以作为有利于泛化的正则化器. BN 已经成为许多最先进的计算机视觉算法的基础. 尽管 BN 取得了巨大的成功, 但它也表现出一些缺点, 这些缺点也是由它在批尺寸上的独特规范化行为引起的. 特别是, BN 需要在足够大的批量下工作. 小批会导致对批统计数据的估计不准确, 而减少 BN 的批大小会极大地增加模型误差. 本节引入群归一化 GN (group normalization)[14]作为 BN 的一种简单的替代方案. 提出 GN 作为一个层, 将通道划分为组, 并对每组内的特征进行规范化. GN 不利用批处理维度, 其计算与批处理大小无关. 将检测头中的卷积模块 BN 层替换为 GN 层, 可以提升检测头定位和分类的性能.

通过使用共享卷积, 将 3 个检测头卷积共享, 可以大幅减少参数数量, 这使得模型更轻便. 在使用共享卷积的同时, 为了应对每个检测头所检测的目标尺度不一致的问题, 使用 Scale 层对特征进行缩放. 综合以上, 可以得到轻量级共享检测头 LWSH (light weight shared convolutional detection head), 如图 11 所示, 可以让检测头做到参数量更少、计算量更少的情况下, 尽可能减少精度的损失.

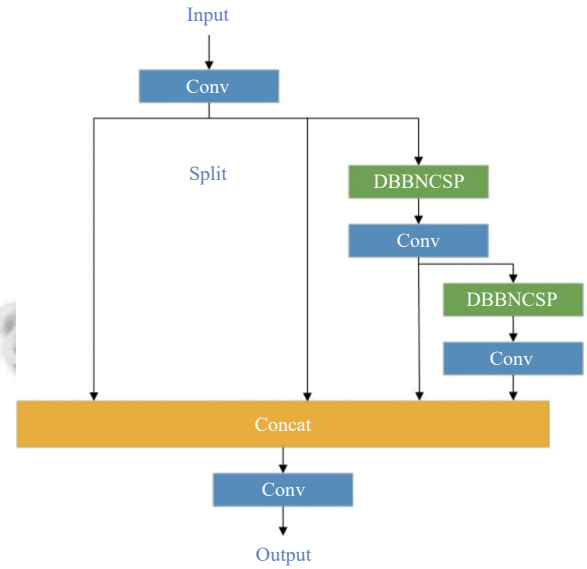

图 10 DBBNCSPELAN 结构图

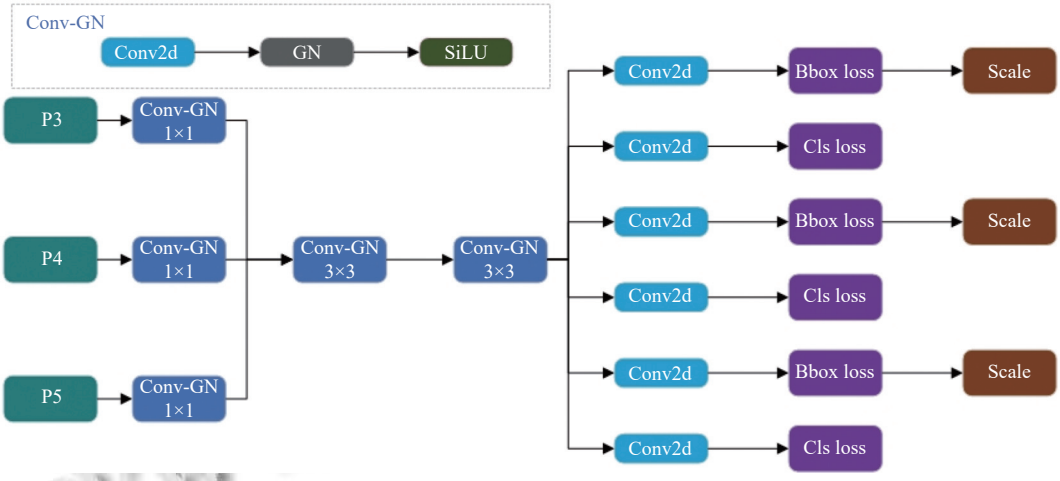

图 11 LWSH 结构图

2.3 SPPF-ELA 模块

由于卡通角色面部检测的情况复杂多变, 为了提高模型对关键特征的提取能力, 在 YOLOv8 骨干网络中的 SPPF 模块引入 ELA (efficient local attention)[15]注意力机制, 使网络忽略无关背景信息干扰, 注意到更多有效特征信息. SPPF-ELA 模块结构如图 12 所示.

图 12 SPPF-ELA 结构图

注意机制能够有效地增强深度神经网络的性能, 在计算机视觉领域得到了广泛的认可. 然而, 现有的方法往往难以有效地利用空间信息. 为了解决这些限制, 本文引入了一种高效局部注意 (ELA) 方法, 该方法以简单的结构实现了实质性的性能改进.

高效局部注意模块作为一个计算单元, 旨在提高对重要目标位置的准确识别. 为了清楚地解释 ELA, 本节将首先介绍 CA (如图 13 所示). CA 主要包括两个步骤: 坐标信息嵌入和坐标注意生成. 作者提出了一种利

用条形池代替空间全局池来捕获远程空间依赖关系的智能方法. 卷积块的输出表示为$R^{H\times W\times C}$, 分别表示高度、宽度和通道维度 (即卷积核的数量). 为了应用条形池化, 在两个空间范围内对每个通道进行平均池化: 沿水平方向$(H,1)$和沿垂直方向$(1,W)$. 这将产生高度为h的第c通道的输出表示, 以及宽度为w的第c通道的输出表示. 这些可以使用式 (1) 和式 (2) 进行数学表示.

$$z_c^h(h)=\frac{1}{H}\sum_{0\leqslant i<H}x_c(h,i) \tag{1}$$

$$z_c^w(w)=\frac{1}{W}\sum_{0\leqslant j<W}x_c(j,w) \tag{2}$$

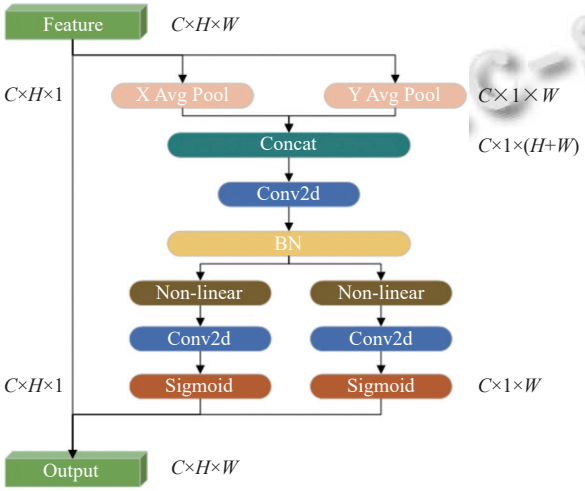

图 13 CA 注意力结构图

CA 通过利用条带池来捕获空间维度上的远程依赖关系, 在精度上有了显著的提高, 特别是对于更深层次的网络. 但 BN 阻碍了 CA 的泛化能力, 而 GN 解决了这些缺陷. 由式 (1) 和式 (2) 导出的定位信息嵌入是信道内的顺序信号. 因此, 通常使用一维卷积比二维卷积更适合处理这些序列信号, 1D 卷积不仅擅长处理序列信号, 而且与 2D 卷积相比也更轻量. CA 虽然使用了 2 次二维卷积, 但它使用的是 1×1 卷积核, 这限制了特征的提取能力. 因此, ELA 采用核大小为 5 或 7 的一维卷积, 有效地增强了定位信息嵌入的交互能力. 这种修改使整个 ELA 能够准确定位重要区域.

从式 (1) 和式 (2) 中得到, ELA 采用了一种新颖的编码方法来生成精确的位置注意图. 由式 (1) 和式 (2) 得到的z_h和z_w不仅捕获了全局感官场, 还捕获了精确的位置信息. 为了有效地利用这些特性, 本文设计了简单的处理方法. 应用一维卷积来增强水平和垂直方向的位置信息. 随后利用 GN (group normalization) 对增强的位置信息进行处理, 得到位置注意在水平方向和垂直方向上的表示, 如式 (3) 和式 (4) 所示.

$$y^h=\sigma(G_n(F_h(z_h))) \tag{3}$$

$$y^w=\sigma(G_n(F_w(z_w))) \tag{4}$$

在上述描述中, 将非线性激活函数表示为σ, 并将 1D 卷积表示为F_h和F_w. 选择将F_h和F_w的卷积核设置为 5 或 7. 一般来说, 卷积核数为 7 的表现会更好, 尽管参数的数量会稍微大一些. 得到的位置注意力在水平和垂直方向上的表示分别用y^h和y^w表示. 最后通过式 (5) 可以得到 ELA 模块的输出, 记为Y. ELA 整体结构如图 14 所示.

$$Y=x_c\times y^h\times y^w \tag{5}$$

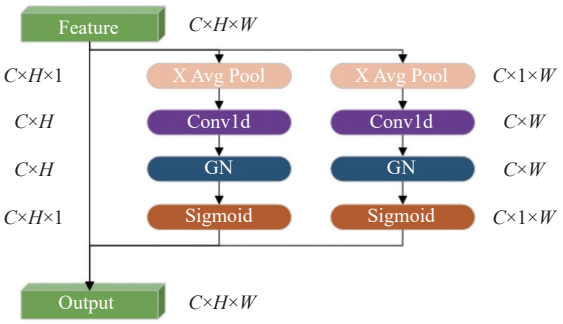

图 14 ELA 注意力结构图

2.4 Shape-IoU 损失函数

过去的边界盒回归方法主要是通过在 *IoU* 上添加新的几何约束来实现更精确的回归, 但忽略了边界盒本身的形状和尺度也会对边界盒回归产生影响. 为了进一步提高回归的精度, 本文引入了新一代的边界回归损失 Shape-IoU 来替换 YOLOv8 中的 CIoU 损失函数.

如图 15 所示, 假设 GT 盒不是正方形, 有长边和短边, 当偏差和形状偏差相同且不全为 0 时, 回归样本中边界盒形状和尺度的差异会导致其 *IoU* 值的差异. 对于相同尺度的边界盒回归样本, 当回归样本的偏差和形状偏差相同且不均为 0 时, 边界盒的形状会对回归样本的 *IoU* 值产生影响. 沿边界盒短边方向偏移和形状偏移所对应的 *IoU* 值变化更为显著. 对于具有相同形状边界框的回归样本, 当回归样本偏差和形状偏差相同且不均为 0 时, 相对于较大规模的回归样本, 较小规模边界框回归样本的 *IoU* 值受 GT 盒形状的影响更为显著. Shape-IoU 正是考虑了其中盒形状的影响, 以达到更合理的损失函数计算.

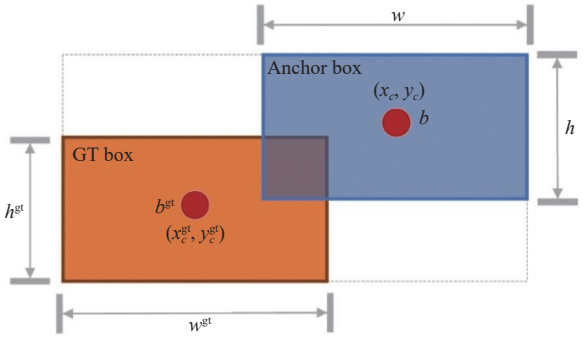

图 15 预测框与真实框示意图

Shape-IoU 的计算公式如下所示:

$$IoU = \frac{|B \cap B^{gt}|}{|B \cup B^{gt}|} \tag{6}$$

$$ww = \frac{2 \times (w^{gt})^{scale}}{(w^{gt})^{scale} + (h^{gt})^{scale}} \tag{7}$$

$$hh = \frac{2 \times (h^{gt})^{scale}}{(w^{gt})^{scale} + (h^{gt})^{scale}} \tag{8}$$

$$distance^{shape} = hh \times \frac{\left(x_c - x_c^{gt}\right)^2}{c^2} + ww \times \frac{\left(y_c - y_c^{gt}\right)^2}{c^2} \tag{9}$$

$$\Omega^{shape} = \sum_{t=w,h} (1 - e^{-\omega t})^{\theta}, \ \theta = 4 \tag{10}$$

$$\begin{cases} \omega_w = hh \times \dfrac{|w - w^{gt}|}{\max(w, w^{gt})} \\ \omega_h = ww \times \dfrac{|h - h^{gt}|}{\max(h, h^{gt})} \end{cases} \tag{11}$$

其中, *scale* 为尺度因子, 与数据集中目标的尺度有关, *ww* 和 *hh* 分别为水平方向和垂直方向的权重系数, 其值与 GT 盒的形状有关. 其对应的边界盒回归损失为:

$$L_{Shape\text{-}IoU} = 1 - IoU + distance^{shape} + 0.5 \times \Omega^{shape} \tag{12}$$

3 实验与分析

3.1 实验数据

实验使用的第 1 个数据集为爱奇艺在 2020 年公开的数据集 iCartoonFace[16], 该数据集由 389 678 张 5 013 个卡通人物的图像组成, 这些图像带有身份, 边界框, 姿势和其他辅助属性, 是目前图像识别领域规模最大、质量高、注释丰富的数据集, 涵盖了图像识别领域的多个方面, 包括近重复、遮挡和外观变化. 数据集中的卡通人物分布广泛, 包括了多个国家卡通作品中的卡通人物. 该数据集中用于面部检测的图片共 60 000 张, 其中训练集 50 000 张, 测试集 10 000 张.

实验使用的第 2 个数据集为 Manga109-Face. 数据集从 Manga109[1]官方网站中申请并获取, 共 21 142 幅图像, 按照 9:1 的比例将数据集划分为训练集和测试集. Manga109 由专业漫画艺术家绘制的漫画图像组成. 本文实验只对角色面部进行检测, 因此需要剔除 Manga-109 中除 Face 以外的所有标签. 处理之后得到数据集 Manga109-Face.

3.2 实验环境和评价指标

本文选择准确率 (*P*), 召回率 (*R*) 和平均精度 (*mAP*) 作为评价指标来评估模型的性能. 公式如下:

$$P = \frac{TP}{TP + FP} \tag{13}$$

$$R = \frac{TP}{TP + FN} \tag{14}$$

$$AP = \int_0^1 P(R)\,dR \tag{15}$$

$$mAP = \frac{1}{n}\sum_{i=0}^{n} AP_i \tag{16}$$

其中, *TP* 表示正确识别为正类的正样本数量, *FP* 表示错误识别为正类的负样本数量, *FN* 表示错误识别为负类的正样本数量, *n* 表示数据集中样本类别的数量. 此外, 还考虑了浮点运算 (GFLOPs) 和参数量, 以便更准确地评估轻量级模型的性能. GFLOPs 代表“每秒千兆浮点运算”, 是衡量深度学习模型计算复杂性的常用单位. 它表示每秒执行的 10 亿次浮点操作 (FLOPs) 的数量. GFLOPs 用于衡量模型的计算需求, 通常用于评估模型对计算资源、速度和效率的需求. 更高的 GFLOPS 值意味着模型需要更多的例如 CPU 和 GPU 的计算资源, 用于训练或推理. 这也可能表明模型需要更多的时间来完成其计算.

本文的实验在 Windows 操作系统下进行, 采用 AMD 5600H CPU、NVIDIA GeForce RTX4070 GPU、8 GB 显存、Python 3.11.4 + PyTorch 2.0.0 + Cuda 11.8 环境, 640×640 输入图像大小, 采用梯度下降法进行训练. 初始学习率设为 0.01, 权值衰减因子设为 5E–4, 批大小设为 16, 动量因子设为 0.937, 共训练 200 轮.

3.3 不同模型检测性能对比

本文将 YOLOv3、YOLOv5、YOLOv6、SSD 和 Faster R-CNN 算法与 YOLO-DEL 算法进行了比较. 结

果见表 1, *mAP* 较本文算法分别提高了 3.33%、3.31%、2.48%、39.33% 和 38.23%, 本文算法参数量分别为其 14%、67%、39%、2.7% 和 1.3%. 与文献[8,9]对比, *mAP* 分别提高 37.17% 和 2.61%. YOLO-DEL 与其他主流模型的直观对比表明, YOLO-DEL 在 *mAP* 以及 GFLOPs 和参数量等计算参数相关的性能方面都优于其他模型.

表 1 在 iCartoonFace 数据集上对比实验

对比模型	*mAP* (%)	参数量 (M)	GFLOPs
SSD[17]	51.03	61.72	24.1
Faster R-CNN[18]	52.13	138.76	—
YOLOv3-tiny[19]	87.03	12.13	19.0
YOLOv5n[20]	87.05	2.51	7.2
YOLOv6n[21]	87.88	4.24	11.9
YOLOv8n[11]	89.10	3.01	8.2
文献[8]	53.19	—	—
文献[9]	87.75	—	—
本文	90.36	1.69	4.6

为体现模型的鲁棒性, 本文同时使用数据集 Manga-109-Face 进行对比实验. 如表 2, 与 YOLOv3-tiny、YOLOv5n、YOLOv6n、YOLOv8n、SSD 和 Faster R-CNN 等方法进行了比较, 在精度与模型大小方面, 本文算法在数据集 Manga109-Face 上同样体现出了优势. 相比于文献[6,9] , 本文算法 *mAP* 分别提高 15.12% 和 3.46%, 证明了本文算法优于同领域其他模型.

表 2 在 Manga109-Face 数据集上对比实验

对比模型	*mAP* (%)	参数量 (M)	GFLOPs
SSD[17]	67.10	138.76	—
Faster R-CNN[18]	15.70	61.72	24.1
YOLOv3-tiny[19]	87.13	12.13	19.0
YOLOv5n[20]	90.11	2.51	7.2
YOLOv6n[21]	90.76	4.24	11.9
YOLOv8n[11]	90.91	3.01	8.2
文献[6]	76.20	—	—
文献[9]	87.86	—	—
本文	91.32	1.69	4.6

3.4 不同注意力机制检测性能对比

为了精确定位关键区域, 抑制无关信息, 提高模型在复杂背景或多目标场景下感知物体的灵敏度和准确性, 在 YOLO 网络结构主干的 SPPF 模块中嵌入了 4 种不同类型的注意机制进行实验, 帮助网络关注与有效目标相关的关键信息, 提取目标区域, 进一步提高了网络的检测能力. 对 5 种不同情况的注意机制的影响进行综合分析和比较, 旨在确定在不改变原意的情况下最有效的注意机制. 分别添加不同的注意力机制, 包括 DA、CBAM、SimAM 和 ELA 注意力. 表 3 通过比较 SPPF 模块中添加的注意机制的效果, 我们可以得出结论, 加入 ELA 注意机制的 *mAP* 最高, 达到 90.5%. 与原来的 YOLOv8 相比, *mAP* 提高了 1.4%, ELA 的加入显著提高了网络检测性能, 超过了 DA、SimAM 和 CBAM 注意力机制带来的性能提升. ELA 在过滤有效特征信息方面表现出卓越的效率, 因此我们采用 ELA 来进一步提高网络的性能.

表 3 注意力机制对比实验结果

注意力机制	*P* (%)	*R* (%)	*mAP* (%)	参数量 (M)	GFLOPs
YOLOv8	90.1	81.1	89.1	3.01	8.2
+CBAM	90.4	81.8	89.6	3.22	8.2
+DA	90.5	82.0	89.8	3.27	8.4
+SimAM	90.2	81.4	89.3	3.01	8.2
+ELA	90.4	83.4	90.5	3.07	8.2

3.5 不同损失函数检测性能对比

为了验证 Shape-IoU 的优越性, 我们使用 Shape-IoU 和一些主流损失函数对 YOLOv8 进行了对比实验. 实验结果见表 4. 使用 Shape-IoU 作为边界盒损失函数时, 模型的 *mAP* 值最高, 其中比 YOLOv8 基线模型使用 CIoU 时 *mAP* 高 0.8%, 同时仍比使用 SIoU 和 WIoU 时分别高 0.6% 和 0.4%, 说明使用 Shape-IoU 作为边界盒回归的检测效果最好.

3.6 消融实验

在模型主干中, 我们用 DBBNCSPELAN 模块替换了 C2f. 此外, 在 SPPF 模块中引入了 ELA 注意力. 在 neck 中, 同样用 DBBNCSPELAN 替换 C2f, 并用 LWSH 替换原本的 Head 部分. 这些改进的目的是增加对关键特征的关注, 同时减少计算复杂度和参数计数, 减轻检测负载, 增强不同尺度大小目标的特征提取能力. 最后将原本 CIoU 损失函数替换为 Shape-IoU 损失函数. 为了评估这些改进对模型性能的影响, 我们进行了消融实验, 逐步添加这些增强模块并训练模型, 目的是演示不同模块的增强效果.

表 4 损失函数对比实验结果 (%)

损失函数	*P*	*R*	*mAP*
CIoU[22]	90.1	81.1	89.1
SIoU[23]	90.7	81.4	89.3
Wise-IoU[24]	89.7	81.8	89.5
Shape-IoU[10]	89.7	82.5	89.9

如表 5, 模型 1 为 YOLOv8 基线模型. 模型 2 是添加了 DBBNCSPELAN 模块的情况. 同样模型 3、模型 4

和模型 5 是分别加入 LWSH、SPPF-ELA 和 Shape-IoU 模块的结果. 实验证明分别单独使用 SPPF-ELA 和 Shape-IoU 模块对模型 *mAP* 有改善作用, 分别单独使用 DBBNCSPELAN 和 LWSH 对减少模型参数量和计算量有很好的效果. 模型 6 和模型 7 是在模型 2 基础上逐渐增加改进模块 LWSH 和 SPPF-ELA 的组合模型, 它们都取得了很好的效果. 与其他模型相比, 模型 8 表明, 当 4 个模块结合使用时, *mAP* 的数值最高. 这些综合改善导致参数量降低 47%, GFLOPs 降低 44%, *mAP* 提升 1.2%.

表 5　消融实验

模型	DBBNC-SPELAN	LWSH	SPPF-ELA	Shape-IoU	*mAP* (%)	参数量 (M)	GFLOPs
1	×	×	×	×	89.1	3.01	8.2
2	√	×	×	×	89.4	2.26	6.2
3	×	√	×	×	88.2	2.36	6.6
4	×	×	√	×	90.5	3.07	8.2
5	×	×	×	√	89.9	3.01	8.2
6	√	√	×	×	88.7	1.62	4.6
7	√	√	√	×	89.6	1.69	4.6
8	√	√	√	√	90.3	1.69	4.6

3.7　实验结果可视化

为进一步验证模型性能, 在数据集 iCartoonFace 和 Manga109-Face 上开展可视化实验. 图 16 为可视化实验结果, 展示了不同背景下 YOLOv8 和 YOLO-DEL 的检测可视化结果. 图 16(a)、(c)、(e) 为 YOLOv8 的检测结果, 图 16(b)、(d)、(f) 为 YOLO-DEL 的检测结果.

在图 16(a) 中, 预测总体上是准确的, 但错误地将右上方的角色手臂识别为卡通角色面部且存在小目标漏检情况, 如黄圈所示. 相比之下, 图 16(b) 正确检测了图片中远处的小目标面部, 且没有错认. 在图 16(c) 中, 它将下方的角色口袋错误地识别为面部, 而图 16(d) 准确地检测了所有面部. 图 16(e)(f) 为 Manga109-Face 数据集实验可视化结果, 图 16(e) 中右下角没有检测出来的小目标面部在图 16(f) 中得到了正确的检测. 从图 16 显示结果来看, 提出的 YOLO-DEL 算法集成了 4 个关键改进, 显著增强了模型的检测性能, 整体面部检测结果准确性有所提升, 同时减少了参数量和计算量, 突出了本文所提算法的性能改进.

综上所述, YOLOv8 本身性能优秀, 这就是选择它作为基线模型的原因. 但与未经改良的 YOLOv8 相比, YOLO-DEL 能够更好地检测卡通人脸, 更适合于检测. 通过对 YOLOv8 进行有针对性的改进, 网络模型在检测方面的性能得到了成功的提高.

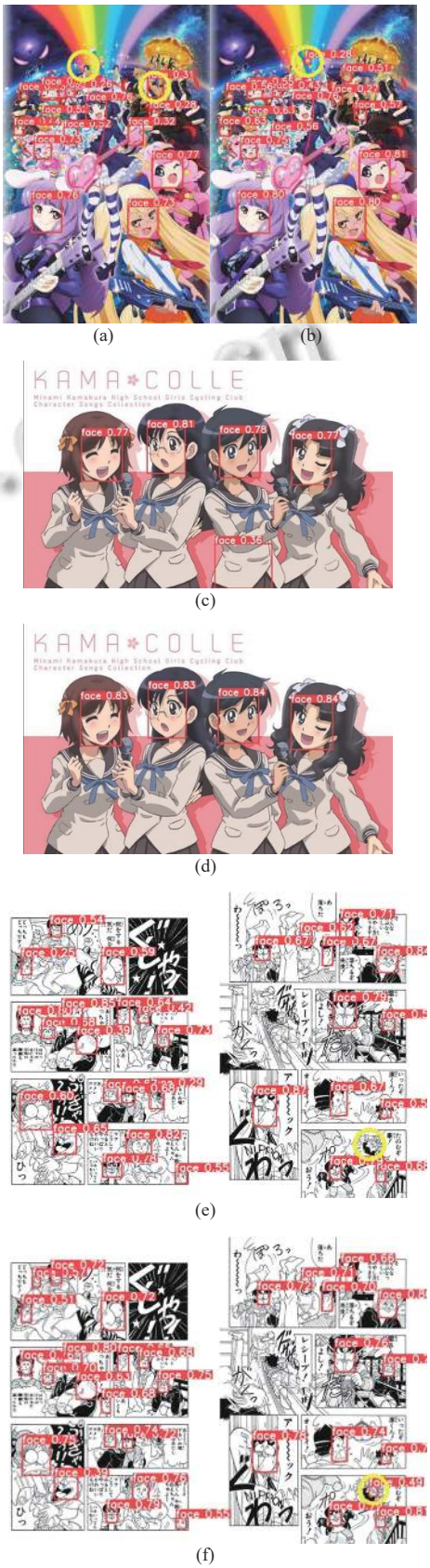

图 16　检测结果可视化图

4 结论与展望

本文针对卡通角色面部目标检测中图像背景复杂、目标小的特点, 提出一种基于 YOLOv8 的卡通角色面部检测模型 YOLO-DEL. 首先, 使用改进重参数化结构 DBBNCSPELAN, 保证了检测速度和精度的平衡. 其次, 在模型主干中引入一个名为 ELA 的注意力机制, 并重新设计 SPPF-ELA 模块, 提高了小目标的检测性能. 接着, 设计了共享卷积检测头, 减小了模型体积. 最后, 引入损失函数 Shape-IoU 代替原本的 CIoU 损失函数, 加快了模型的收敛速度. 进一步, 在 iCartoonFace 和 Manga109-Face 两个数据集上开展了实验, 在实验结果中 *mAP* 分别达到了 90.36% 和 91.32%, 相比 YOLOv8n 提高了 1.26% 和 0.41%, 参数量降低了 47%, GFLOPs 降低了 44%. 此外, 与 YOLOv3、YOLOv5 和 YOLOv6 等网络模型相比, YOLO-DEL 在各个方面都优于这些现有模型, 且在保持较高检测精度的同时显著降低了计算量. 未来我们将进一步优化模型, 以增强模型在具有挑战性的场景中的适应性和有效性.

参考文献

1 Matsui Y, Ito K, Aramaki Y, *et al*. Sketch-based manga retrieval using Manga109 dataset. Multimedia Tools and Applications, 2017, 76(20): 21811–21838. [doi: 10.1007/s11042-016-4020-z]

2 Kohei K, Kohei H, Kohei T. Face detection and face recognition of cartoon characters using feature extraction. Proceedings of the 2012 IIEEJ Image Electronics and Visual Computing Workshop. Kuching, 2012. 18.

3 傅俊成. 基于 Mask R-CNN 的卡通人物识别 [硕士学位论文]. 北京: 北京邮电大学, 2022.

4 夏华丽. 基于 Faster R-CNN 戏曲卡通人物识别的研究 [硕士学位论文]. 长沙: 湖南师范大学, 2020.

5 张健. 基于神经网络的动漫人物识别研究 [硕士学位论文]. 成都: 电子科技大学, 2019.

6 Ogawa T, Otsubo A, Narita R, *et al*. Object detection for comics using Manga109 annotations. arXiv:1803.08670, 2018.

7 Nguyen NV, Rigaud C, Burie JC. Comic characters detection using deep learning. Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Kyoto: IEEE, 2017. 41–46.

8 陈争光. 基于 YOLOv3 网络的卡通头像检测研究 [硕士学位论文]. 武汉: 中南财经政法大学, 2022.

9 Topal BB, Yuret D, Sezgin TM. Domain-adaptive self-supervised face & body detection in drawings. Proceedings of the 32nd International Joint Conference on Artificial Intelligence. Macao: IJCAI, 2023. 1432–1439.

10 Zhang H, Zhang SJ. Shape-IoU: More accurate metric considering bounding box shape and scale. arXiv:2312.17663, 2024.

11 沈学利, 王灵超. 基于 YOLOv8n 的无人机航拍目标检测. 计算机系统应用, 2024, 33(7): 139–148. [doi: 10.15888/j.cnki.csa.009567]

12 Wang CY, Yeh IH, Liao HYM. YOLOv9: Learning what you want to learn using programmable gradient information. arXiv:2402.13616, 2024.

13 Ding XH, Zhang XY, Han JG, *et al*. Diverse branch block: Building a convolution as an inception-like unit. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 10881–10890.

14 Wu YX, He KM. Group normalization. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 3–19.

15 Xu W, Wan Y. ELA: Efficient local attention for deep convolutional neural networks. arXiv:2403.01123, 2024.

16 Li SC, Zheng Y, Lu XJ, *et al*. iCartoonFace: A benchmark of cartoon person recognition. arXiv:1907.13394, 2019.

17 Liu W, Anguelov D, Erhan D, *et al*. SSD: Single shot MultiBox detector. Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer, 2016. 21–37.

18 Ren SQ, He KM, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2015. 91–99.

19 Redmon J, Farhadi A. YOLOv3: An incremental improvement. arXiv:1804.02767, 2018.

20 管嘉程, 任红卫, 周宋佳. 基于 YOLOv5 改进的轻量化目标检测. 计算机系统应用, 2023, 32(9): 132–142. [doi: 10.15888/j.cnki.csa.009292]

21 Li CY, Li LL, Jiang HL, *et al*. YOLOv6: A single-stage object detection framework for industrial applications. arXiv:2209.02976, 2022.

22 Zheng ZH, Wang P, Liu W, *et al*. Distance-IoU loss: Faster and better learning for bounding box regression. Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York: AAAI, 2020. 12993–13000.

23 Gevorgyan Z. SIoU loss: More powerful learning for bounding box regression. arXiv:2205.12740, 2022.

24 Tong ZJ, Chen YH, Xu ZW, *et al*. Wise-IoU: Bounding box regression loss with dynamic focusing mechanism. arXiv:2301.10051, 2023.

(校对责编: 张重毅)