ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 37,No. 5,Sep. 2022,pp. 1070－1083
DOI:10. 16337/j. 1004-9037. 2022. 05. 012

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于轻型尺度自适应深度网络的低清人脸检测算法

胡洪明[1]，邵文泽[1]，李金叶[1]，葛　琦[1]，邓海松[2]

(1. 南京邮电大学通信与信息工程学院，南京 210003;2. 南京审计大学统计与数据科学学院，南京 211815)

摘　要：由于安防设备硬件条件等因素制约，在视频监控场景下的低清人脸检测中注重模型在检测精度、速度以及占用内存大小等方面的权衡已然是必须考虑的问题。针对此问题，将可变形卷积(Deformable convolution，DC)和Lambda层进行融合，提出一种轻型尺度自适应深度网络的低清人脸检测模型DLFace。首先借鉴RetinaFace算法，使用改进后的深度可分离卷积能够有效防止训练过程中信息丢失；其次将改进后的可变形卷积引入骨干网络和SSH (Single stage headless) 检测模块，通过增强感受野适应人脸多因素的变化；最后在骨干网络高层引入Lambda层，有效挖掘语义和位置信息，形成更加丰富的特征表示。在WiderFace数据集上的实验结果表明，DLFace实现了性能和速度的平衡，在不同场景下均验证了DLFace的优越性，表明DLFace能较好地适用于视频监控场景下的低清人脸检测任务。

关键词：人脸检测；可变形卷积；轻量化；多尺度特征融合

中图分类号：TN911.73　　**文献标志码**：A

Low-Resolution Face Detection Based on Light-Weight Scale-Adaptive Convolutional Neural Networks

HU Hongming[1], SHAO Wenze[1], LI Jinye[1], GE Qi[1], DENG Haisong[2]

(1. College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China; 2. College of Statistics and Mathematics, Nanjing Audit University, Nanjing 211815, China)

Abstract: As for low-resolution face detection in real-world video surveillance, achieving balance in terms of speed, accuracy, and memory consumption is of great importance due to the hardware constraints. Towards the problem, inspired by the more recent RetinaFace this paper proposes a light-weight scale-adaptive deep face detection model, termed as DLFace. Firstly, the improved depthwise separable convolution can effectively prevent information loss during training. Secondly, the improved deformable convolution is introduced into the backbone network and single stage headless (SSH) face detector, so as to enlarge the receptive field while also to adapt to facial changes such as expression, pose and so on. Finally, a Lambda layer is introduced in the high level of the backbone network, attempting to effectively explore the semantic and location information to form a richer representation of facial features. Experimental results on the WiderFace dataset show that DLFace has achieved a comparable or even better performance than existing light-weight face detection methods. Meanwhile, DLFace also achieves a better performance balance than most of previous methods in prediction efficiency and effectiveness.

基金项目：国家自然科学基金(61771250, 61972213, 11901299)。
收稿日期：2021-09-22；**修订日期**：2021-12-28

Key words: face detection; deformable convolution; lightweight; multi-scale feature fusion

引 言

在非约束的视频监控等场景下,摄像头采集到的人脸影像往往存在尺寸、模糊、遮挡以及光照、表情、姿态和妆容变化等问题,因此真实监控场景下的人脸检测工作非常具有挑战性。对于人脸检测任务,早期依靠手工设计特征的经典方法由于曝光、角度和遮挡等问题干扰,性能往往有限。

近年来,随着深度学习作为人工智能领域快速发展的热点方向,基于深度卷积神经网络的人脸检测研究取得了突破性进展。例如,2016年,Ohn-Bar等[1]在机器学习中常用的增强树算法基础上,通过对弱学习器体系结构的改进,显著提升增强树算法对于人脸检测问题的建模能力。Yang等[2]提出的LDCF+人脸检测算法是目前在WiderFace榜单排名第一的非深度算法。Li等[3]提出级联人脸检测网络Cascade CNN,通过多尺度方式实现图像金字塔,并采用非极大值抑制(Non-maximum suppression,NMS)合并高度重叠的窗口,解决传统方法对角度、光照敏感等问题,但对小脸检测存在瓶颈。2017年,Yang等[4]提出Faceness-Net,采用多个人脸部位分类器实现人脸评分,依据各部位得分情况进行回归得到感兴趣人脸区域,最后利用特征提取网络得到人脸检测结果。Zhang等[5]提出MTCNN检测算法,借鉴Cascade CNN思想,设计3个级联网络,采用多任务训练方式,有效提升检测效率。Zhu等[6]提出CMS-RCNN,基于多尺度卷积神经网络,利用人脸及人体周围特征实现小脸检测。Hu等[7]从人脸上下文信息、人脸图像分辨率和尺度不变性3个方面对小脸检测展开深入研究,并提出HR算法,为后续小目标检测提供了有效参考。Najibi等[8]提出SSH(Single stage headless)算法,其最大优势在于尺度不相关性,通过对卷积神经网络不同阶段输出层设计3个分支,每个分支只需进行类似的过程进行检测和分类,并创新性地提出了上下文模块,有效获取更大感受野。Zhang等[9]提出S3FD人脸检测算法,有效解决了SSD(Single shot multibox detector)对于小目标不够鲁棒的问题。Tang等[10]提出了PyramidBox检测算法,采用基于锚框的图像上下文方法,能较好地提取模糊、遮挡的小脸特征,并设计低层特征金字塔网络融合图像上下文和人脸特征。腾讯优图提出DSFD算法[11],通过构建新的特征增强模块,在有限的特征图内学到更多上下文和语义信息。2019年,著名的InsightFace团队提出RetinaFace算法[12],利用联合监督和自监督的多任务损失函数实现对多尺度人脸的像素级定位,骨干网络借鉴RetinaNet结构[13],并采用特征金字塔实现多尺度特征融合,添加了SSH网络检测模块,对多尺度目标效果显著。

然而,目前大多数人脸检测算法并不能直接部署到端侧。例如,若将WiderFace榜单上成绩排名前5的DSFD模型(内存大小约460 MB)[11]部署在安防监控嵌入式设备上,设备大概率会死机。因此,注重模型在检测精度、速度以及占用内存大小等方面的权衡显得尤为重要。相较于级联和两阶段人脸检测算法,S3FD[9]、RetinaFace[12]等单阶段人脸检测算法是基于锚框实现分类和回归,有望在检测速度和性能上取得平衡,是目前人脸检测算法优化的主流方向。

在RetinaFace算法基础上,本文提出一种面向视频监控场景的轻型低清人脸检测模型DLFace。考虑到监控终端设备内存有限,模型首先使用改进后的深度可分离卷积,有效防止特征提取过程中信息丢失。针对人脸尺寸、模糊、遮挡以及光照、表情、姿态和妆容等因素,模型进一步引入可变形卷积DCNv2[14],通过增加感受野适应人脸多因素的变化。最后,模型在网络高层引入Lambda层[15]有效挖掘语义和位置信息,通过增强局部上下文的关联提升人脸检测算法的整体性能。通过在WiderFace数据集上与代表性人脸检测算法的对比结果有效验证了DLFace在检测精度、检测速度以及占用内存上的均衡性。

1 相关工作

1.1 可变形卷积

在计算机视觉领域,同一目标在不同场景、角度中产生未知的几何形变是检测和分割等问题的一大挑战。对于该问题,Dai等[16]于2017年提出了可变形卷积概念,在传统卷积核上添加位置偏移向量,提升模型对于形变目标的建模能力。

传统卷积通常使用3×3、5×5的规则形状卷积核提取特征,首先在特征图上使用固定大小的感受野R进行采样,然后使用卷积核w对采样点进行加权运算。对于输出特征图中每个位置p_0,其输出特征值$y(p_0)$表达式为

$$y(p_0)=\sum_{p_n\in\Omega} w(p_n)x(p_0+p_n) \tag{1}$$

式中:假设使用3×3卷积核,$\Omega=\{(-1,-1),(-1,0),\cdots,(0,1),(1,1)\}$;$w(p_n)$为该采样位置上的卷积核权重;$x$为输入图像特征图;$p_n$为$\Omega$中位置元素。

相比于传统卷积,可变形卷积对卷积核每个抽样点添加一个位置偏移量Δp_n,不再局限于固定位置采样,表达式为

$$y(p_0)=\sum_{p_n\in\Omega} w(p_n)x(p_0+p_n+\Delta p_n) \tag{2}$$

式中Δp_n为偏置矩阵对应传统卷积感受野的偏移量。图1显示了可变形卷积的实现过程。首先对于输入的特征图,会再经过一个卷积,得到通道数为$2N$的偏置域,偏置域尺寸和输入特征图相同,从偏置域上可以得到每个像素点的偏置矩阵,通过偏置矩阵产生偏移量Δp_n。模型训练过程中,偏移量通过插值算法反向传播进行学习。

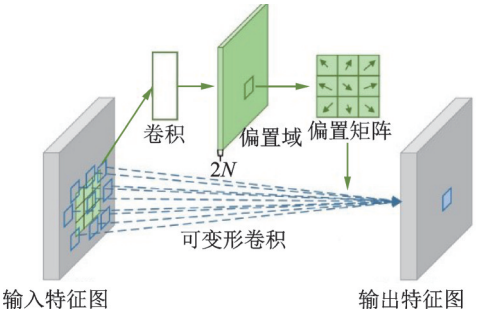

图1 可变形卷积实现过程

Fig.1 Implementation process of deformable convolution

对式(2)进行分析,假设可变形卷积位置偏移量Δp_n为0,可变形卷积等价为传统卷积,此时模型也能实现传统卷积性能。传统卷积加上偏移量的学习后,可变形卷积会根据当前图像内容进行调整,使卷积核能拟合不同物体形状,提取更细致的特征。对于监控摄像头采集的画面,可变形卷积会依据不同人脸形态、大小、尺度变化调整卷积核,有效提升人脸检测精度。特别是当卷积核处理大量密集小脸或是被遮挡的人脸情况时,传统卷积只能对固定大小区域提取特征,这样不可避免地会引入非目标对象信息从而导致误差产生。相比于大脸,小脸分辨率低产生误差的概率更高。而可变形卷积能有效解决该问题,通过提取精细特征提升小脸检测精度。

虽然可变形卷积提取特征的能力更强,但是对于大目标不可避免地也会引入噪声干扰。为此Zhu等[14]提出了可变形卷积DCNv2,通过在偏置矩阵间添加权重,选择性对位置进行偏移,并提出了特征模拟方案指导网络训练,进一步提升可变形卷积拟合效果。DLFace采用的是可变形卷积DCNv2。

1.2 Lambda层

卷积神经网络中低层特征位置细节信息丰富,高层特征语义信息充足。随着层数的加深,位置信息会逐渐丢失,而如何将语义和位置信息建立联系显得尤为重要。针对卷积神经网络的这一缺陷以及

自注意力机制计算量大的问题,谷歌大脑提出了一种新的网络结构Lambda层[15]。Lambda层本质上也是一种注意力机制,相比于自注意力机制[17]需要通过大量计算建立注意力图,其通过巧妙设计绕过这一步,并且在语义和位置信息间建立联系,能够更好地学习到全局、局部特征间的依赖关系。Lambda层通过将上下文信息转换为线性函数λ,并且将这些函数应用于每个输入来获取特征间的依赖关系。其实现过程如图2所示。其中,图2(a)为基于上下文生成λ函数过程。首先m个d维的上下文(Context)元素经过线性映射分别得到m个k维的键$\boldsymbol{K}$和m个v维的值$\boldsymbol{V}$,$\boldsymbol{K}$经过softmax函数$\sigma(\cdot)$归一化,$\boldsymbol{V}$和$\boldsymbol{K}$经过矩阵乘法得到语义映射函数$\boldsymbol{\lambda}^c$,$\boldsymbol{\lambda}^c$固定不变。图2(a)中$\boldsymbol{E}_i$为相对位置编码矩阵,$\boldsymbol{E}_i$分别和$\boldsymbol{V}$相乘得到n个位置映射函数$\boldsymbol{\lambda}_i^p$,将$\boldsymbol{\lambda}_i^p$和$\boldsymbol{\lambda}^c$相加便得到最终的映射函数$\boldsymbol{\lambda}_i$。将图2(a)中得到的映射函数$\boldsymbol{\lambda}_i$应用于图2(b)中每个输入q_i后得到输出$\boldsymbol{y}_i$。上述过程表达式为

$$\boldsymbol{\lambda}^c=\sum_m \overline{\boldsymbol{K}}_m^{\mathrm{T}}\boldsymbol{V}_m \tag{3}$$

$$\boldsymbol{\lambda}_i^p=\sum_m \boldsymbol{E}_{im}^{\mathrm{T}}\boldsymbol{V}_m \tag{4}$$

$$\boldsymbol{\lambda}_i=\boldsymbol{\lambda}^c+\boldsymbol{\lambda}_i^p \tag{5}$$

$$\boldsymbol{y}_i=q_i\boldsymbol{\lambda}_i=q_i\left(\boldsymbol{\lambda}^c+\boldsymbol{\lambda}_i^p\right)=q_i\left(\sum_m\left(\overline{\boldsymbol{K}}_m+\boldsymbol{E}_{im}\right)^{\mathrm{T}}\boldsymbol{V}_m\right)=\sum_m q_i\left\{\left(\overline{\boldsymbol{K}}_m+\boldsymbol{E}_{im}\right)^{\mathrm{T}}\boldsymbol{V}_m\right\} \tag{6}$$

式中$\overline{\boldsymbol{K}}_m$为经过softmax函数归一化后的$\boldsymbol{K}_m$。相比于自注意力机制,Lambda函数巧妙地将相似度计算过程解耦成2个独立的分量,这样$(\overline{\boldsymbol{K}}_m+\boldsymbol{E}_{im})^{\mathrm{T}}\boldsymbol{V}_m$可以先计算。将$\boldsymbol{V}_m$的维度看作定值,$(\overline{\boldsymbol{K}}_m+\boldsymbol{E}_{im})^{\mathrm{T}}\boldsymbol{V}_m$的计算复杂度为$O(n)$,这样整个映射过程的计算复杂度也为$O(n)$。

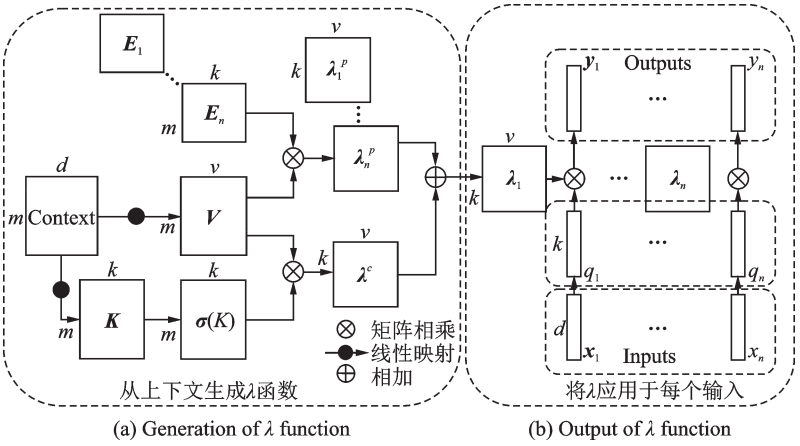

图2 λ函数实现过程

Fig.2 Implementation process of λ function

Lambda层相比自注意力,计算复杂度更低,并且考虑了相对位置信息,通过联合位置和语义信息能够更好地获取全局、局部特征间的依赖关系。对于视频监控场景中的人脸图像出现遮挡、曝光和姿态变化的概率更大,因此将图像中的位置、语义信息联合考虑显得尤为重要。由于网络中高层特征语义信息更丰富,该文模型在骨干网络高层使用Lambda层有效挖掘图像位置和语义信息,提升模型整体性能。

2 本文人脸检测算法

2.1 深度可分离卷积

目前常用的深度卷积神经网络如ResNet,虽然特征表示能力强,但是由于模型复杂度高且训练过于耗时,并不适合视频监控等硬件性能有限的设备。本文基于性能和计算量兼顾的单阶段人脸检测算法RetinaFace,采用MobileNet轻型网络替代ResNet50来提取人脸特征,通过改进深度可分离卷积单元以及改造骨干网络方式,最终在提升视频监控场景下低清人脸检测精度的同时显著降低检测模型的参数量。

2014年Sifre等[18]首次提出了深度可分离卷积,随后便被应用在GoogleNetv2[19]中减少网络前几层的计算量。受此启发,谷歌团队在2017年提出了MobileNet[20]结构,其设计初衷是针对手机等嵌入式设备,核心思想是将卷积核巧妙分解,从而可以有效降低网络参数。如图3所示,其中图3(a)为常用的标准卷积,图3(b,c)分别为深度卷积和点卷积。假设输入和输出特征图的高和宽均为D_K,M和N分别为输入、输出的通道数,输入特征图大小为$D_K \times D_K \times M$,输出特征图大小为$D_K \times D_K \times N$。

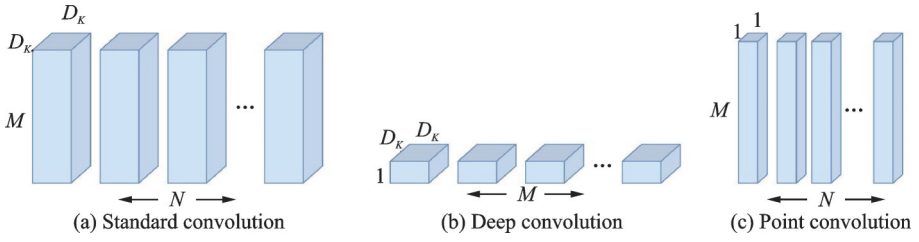

图3 标准卷积和深度可分离卷积图

Fig.3 Standard convolution and deep separable convolution graph

对于标准卷积,假设卷积核步长为1,则其卷积过程可以表示为

$$G_{k,l,n}=\sum_{i,j,m} K_{i,j,m,n} F_{k+i-1,l+j-1,m} \tag{7}$$

该过程计算量为

$$O_1=D_K \times D_K \times M \times N \times D_F \times D_F \tag{8}$$

深度卷积对每个通道使用一种卷积核,卷积过程可以表示为

$$\hat{G}_{k,l,m}=\sum_{i,j} \hat{K}_{i,j,m} F_{k+i-1,l+j-1,m} \tag{9}$$

深度卷积计算量为

$$O_2=D_K \times D_K \times M \times D_F \times D_F \tag{10}$$

深度卷积加上1×1点卷积称为深度可分离卷积,深度可分离卷积计算量为

$$O_3=D_K \times D_K \times M \times D_F \times D_F + M \times N \times D_F \times D_F \tag{11}$$

将标准卷积和深度可分离卷积计算量进行比较,可以得出

$$\frac{D_K \times D_K \times M \times D_F \times D_F + M \times N \times D_F \times D_F}{D_K \times D_K \times M \times N \times D_F \times D_F}=\frac{1}{N}+\frac{1}{D_K^2} \tag{12}$$

MobileNet中的深度卷积主要使用3×3的卷积核,因此D_K为3,通道数N一般在几百,$1/N$可忽略不计,因此深度可分离卷积相比标准卷积,其计算量可以降低至1/9左右。在此基础上引入宽度因子α可以进一步优化模型计算量,其表达式为

$$\frac{D_K\times D_K\times \alpha M\times D_F\times D_F+\alpha M\times \alpha N\times D_F\times D_F}{D_K\times D_K\times M\times N\times D_F\times D_F}=\frac{\alpha}{N}+\frac{\alpha^2}{D_K^2} \tag{13}$$

本文模型中α设置为0.25,引入宽度因子后,模型计算量进一步降低。为了降低过拟合风险,同时避免梯度消失影响,MobileNet网络在深度可分离卷积中两次使用了ReLU激活函数。但使用ReLU也有风险,如果网络在前向传导过程中有个大梯度使得权重更新很大,此时神经元对于所有输入都会给出负值,当负值经过ReLU激活函数后输出都为零,这样导致神经元无法更新参数,网络便会丢失信息。

本文对MobileNet中的深度可分离卷积进行改进,即使用LeakyReLU,在ReLU的负半区间引入泄露(Leaky)值,使得ReLU对于负值输入始终保持小的梯度。改进前后深度可分离卷积的计算量近乎相同,但由于使用LeakyReLU,有效降低模型训练过程中信息丢失。图4为改进前后深度可分离卷积对比。

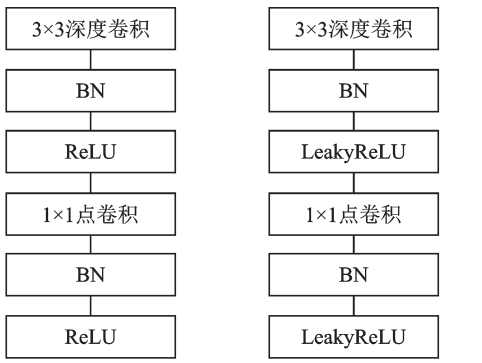

图4 改进前后的深度可分离卷积

Fig.4 Deep separable convolution before and after improvment

2.2 轻型低清人脸检测网络模型

本文提出的轻型低清人脸检测模型DLFace,其最大的特点就是尺度自适应。相比于MTCNN人脸检测算法,需要多次输入不同尺寸图片进行预测最后经过非极大值抑制,DLFace对于特征金字塔(Feature Pyramid network,FPN)和SSH检测模块只需一次输入即可完成预测。DLFace模型结构如图5所示。

本文通过多次实验后将原MobileNet第3、7、13层深度可分离卷积替换为可变形卷积DCNv2,分别在各阶段加强对于不同尺度人脸特征的精细化提取。若将原MobileNet第1层3×3卷积替换为DCNv2,此时模型计算量过大导致训练非常耗时;若将原MobileNet第2、6、12层步长为2的深度可分离卷积替换为DCNv2,实际训练过程中出现过拟合现象。通过多次实验后将原MobileNet第11层深度可分离卷积替换为Lambda Layer,有效挖掘图像位置和语义信息。

DLFace网络中输入为尺寸640×640的3通道图像。在骨干网络中,Conv_bn指带有批量归一化的卷积,其步长为2,卷积核大小为3×3,Conv_dw表示深度可分离卷积。骨干网络依据图片输出尺寸可以分为3个阶段:第1阶段为人脸图片经过特征提取后输出尺寸为80×80的64通道特征图;第2阶段为通道特征图经过特征提取后输出尺寸为40×40的128通道特征图;第3阶段为通道特征图经过特征提取后输出尺寸为20×20的256通道特征图。

在卷积神经网络提取特征过程中,高层特征图语义信息更丰富,对于检测大脸更有优势,低层特征图位置信息更多,对于检测小脸更有优势。为了使模型更好地检测不同尺寸人脸,该文将3个阶段的输出分别经过特征金字塔模块,将高层特征通过上采样和低层特征融合。首先使用1×1卷积降低输入维度,然后将第2、第3阶段输出特征图分别经过2倍上采样,保证结果和第1、第2阶段特征图尺寸相同。为了降低上采样后图片出现的混叠现象,于是将融合后的特征图经过3×3卷积,最后将输出的3个特征图分别输入对应的SSH检测模块。

图6为SSH检测器结构,其设计思想来源于人脸检测算法SSH,通过1个3×3卷积和1个上下文模块进一步加强特征提取。这里上下文模块在使用串联和并联的3×3卷积降低通道数的同时增大感受野,提升特征提取精度。之后将不同通道数的特征进行拼接并输入LeakyReLU激活函数,有效降低特

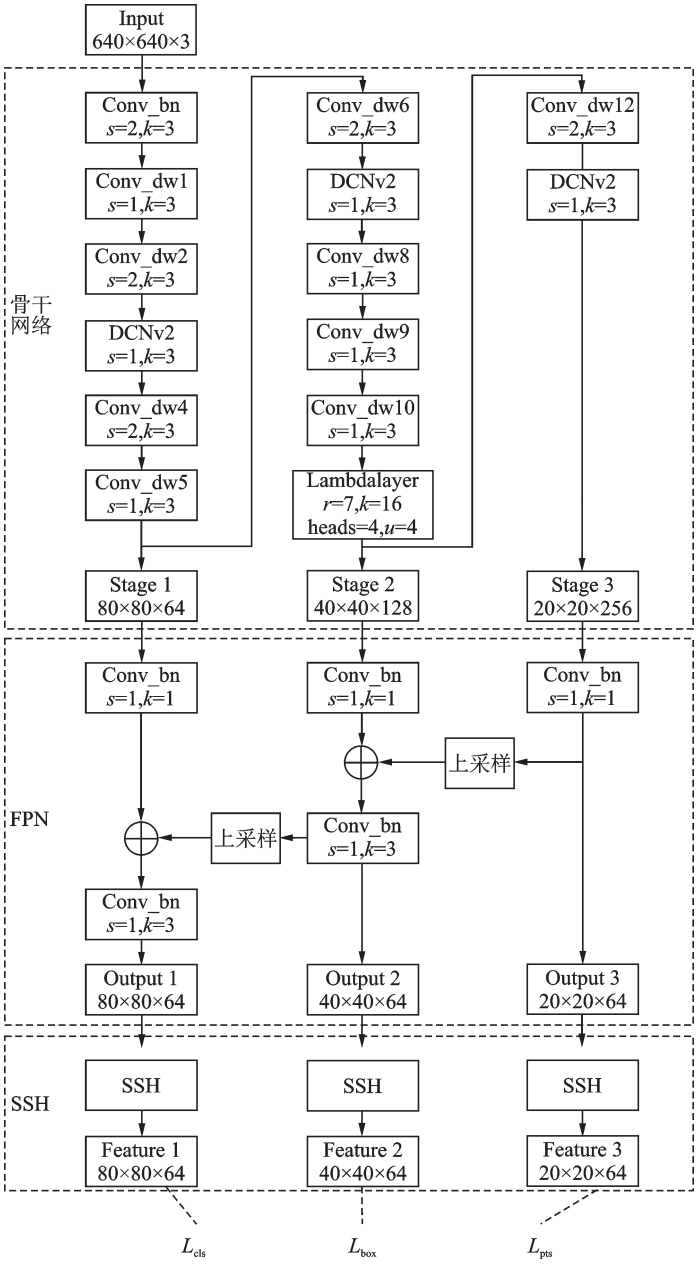

图5 DLFace网络结构

Fig.5 Network structure of DLFace

征提取过程中的信息丢失。相比于原RetinaFace中的SSH检测模块，本文在LeakyReLU激活函数后添加了可变形卷积DCNv2，进一步精细化提取特征，有效提升小脸检测精度，最终经过SSH检测模块输出3个不同尺寸特征图。

在经过SSH检测器后得到7个输出特征图。如图7所示，为了便于后续转换，这里将训练批次数量Batchsize加上，即32×64×80×80表示32张64通道的特征图，每张尺寸为80×80。首先经过1×1卷积及维度转换后得到3类通道和尺寸不同的特征图，再经过依次尺寸转换以及通道拼接即得到用于计

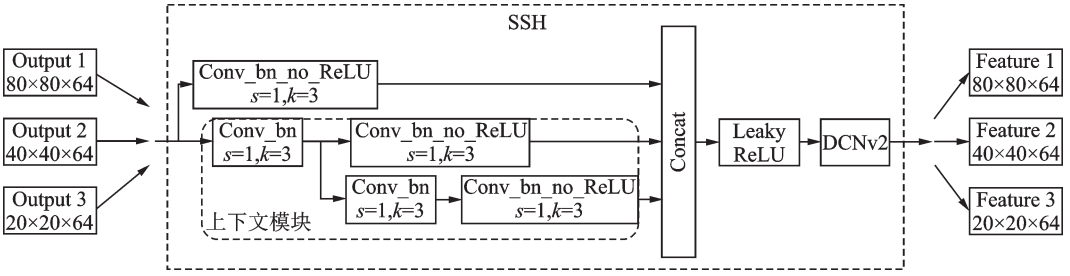

图6 DLFace中的SSH检测模块

Fig.6 SSH detection module in DLFace

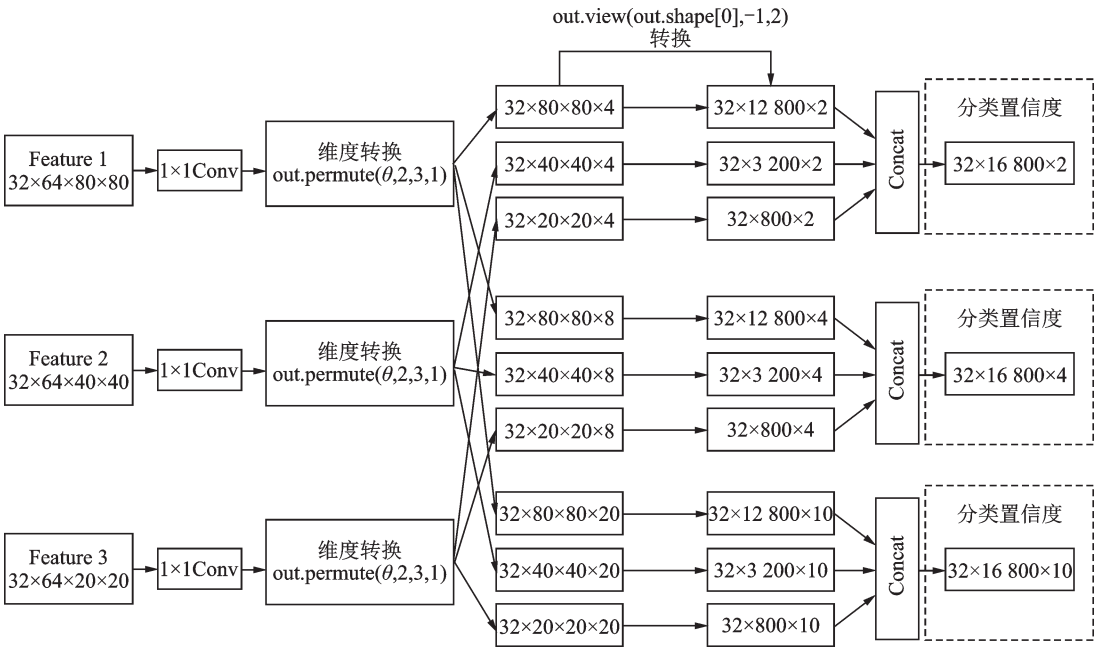

图7 输出特征图转换过程

Fig.7 Transformation process of output features

算分类损失的特征图。根据同样步骤,可以得到计算人脸框定位损失和关键点标记损失的特征图。

2.3 损失函数和训练方法

DLFace损失函数可表述为多任务损失函数,如式(14)所示,其中λ_1和λ_2为平衡3个损失间的平衡因子,实验中分别设置为0.25和0.1,N为每批次输入样本数,N_1为每批次图像中人脸关键点总数。

$$L=\frac{1}{N}\sum_i L_{\text{cls}}\left(p_i,p_i^*\right)+\frac{\lambda_1}{N}\sum_i p_i^* L_{\text{box}}\left(t_i,t_i^*\right)+\frac{\lambda_2}{N_1}\sum_i p_i^* L_{\text{pts}}\left(l_i,l_i^*\right) \tag{14}$$

第1项中$L_{\text{cls}}(p_i,p_i^*)$代表人脸分类损失,$p_i$表示预测框中存在人脸的概率,$p_i^*$为真实值,正样本框为1,负样本框为0。分类损失使用交叉熵损失,表达式为

$$L_{\text{cls}}=-p_i^*\lg p_i \tag{15}$$

第2项中$p_i^* L_{\text{box}}(t_i,t_i^*)$代表人脸框回归损失,其中$t_i=\{t_x,t_y,t_w,t_h\}_i$表示与正样本框对应预测框的位置,$t_i^*=\{t_x^*,t_y^*,t_w^*,t_h^*\}_i$表示与正样本框对应真实标注框的位置。对人脸框坐标进行归一化并计算

$L_{box}(t_i, t_i^*)=R(t_i-t_i^*)$,其中$R$表示Smooth L_1损失函数,如式(16)所示。Smooth L_1损失函数能从两方面限制梯度,当预测框和真实框差距过大时,保证梯度不会过大;当预测框和真实框差距很小时,保证梯度足够小。

$$R\left(t_i-t_i^*\right)=\begin{cases}0.5(t_i-t_i^*)^2 & \left|t_i-t_i^*\right|<1\\ \left|t_i-t_i^*\right|-0.5 & \left|t_i-t_i^*\right|\geqslant 1\end{cases} \tag{16}$$

第3项中$p_i^* L_{pts}(l_i, l_i^*)$代表关键点回归损失,其中$l_i=(l_{x1}, l_{y1}, \cdots, l_{x5}, l_{y5})$表示正样本人脸框中5个关键点的预测值,$l_i^*=(l_{x1}^*, l_{y1}^*, \cdots, l_{x5}^*, l_{y5}^*)$表示正样本人脸框中5个关键点的真实值。类似人脸框回归损失,对人脸关键点进行归一化并计算$L_{pts}(l_i, l_i^*)=R(l_i-l_i^*)$。

本文实验在PyTorch深度学习框架下对轻型低清人脸检测模型进行训练,训练集选用WiderFace数据集中的训练数据。优化器使用随机梯度下降法,批量大小(Batch size)设为32,初始学习率设置为0.001,权重衰减设置为0.000 5,在经过第190、220次遍历后,学习率分别下降10倍,最终经过250次遍历训练结束。将训练后得到的模型权重用于验证,验证集选用WiderFace数据集中的验证数据,最后得到模型在验证集简单、中等、困难子集上的验证精度。

3 实验结果与分析

3.1 实验数据集

本文实验采用WiderFace数据集,其包含32 203张图片以及393 703个标注人脸,在面部的尺寸、姿态、遮挡、表情、妆容及光照上都有很大变化。WiderFace数据集涵盖61个不同的事件类别,对于每一个事件类别,随机选取40%用于训练,10%用于验证,50%用于测试,且根据图像中人脸尺寸分为简单、中等、困难3个子集,其人脸分辨率分别大于50像素×50像素、30像素×30像素、10像素×10像素。图8为WiderFace人脸图像示例。

图8 WiderFace数据集图片示例

Fig.8 Pictures in WiderFace dataset

3.2 评价指标

实验中使用的评价指标为预测精准率(Precision)和召回率(Recall)。

(1)预测精准率表示检测出的人脸是正确人脸的概率,如式(17)所示,其中TP表示检测器将检测到的人正确预测为正样本的数目,FP表示检测器将检测到的人错误预测为负样本的数目。

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{17}$$

(2)召回率表示正确检测出的人脸数量占图片人脸总数的比例,如式(18)所示,其中FN表示检测器将未检测到的人错误预测为正样本的数目。

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{18}$$

将预测精准率作为x轴,召回率作为y轴进行做图,得到了PR(Precision-recall)曲线。曲线的2个变量呈负相关,检测器的性能越好,曲线包含的面积越大。平均预测精度(Average precision,AP)即为PR曲线下方面积,检测器的性能越好,AP值越高。

3.3 实验结果

为了验证DLFace的潜在优势,将其与目前性能较优以及模型轻量化的代表性方法进行了综合对比。结果如表1所示,首先从算法骨干网络和模型权重大小方面进行比较。表1中DSFD[11]、RetinaFace(R)[12]人脸检测算法均使用层数较深的深度残差网络ResNet。其中,DSFD使用ResNet152作为骨干网络,通过设计特征增强模块在有限特征图内学习到更多的语义和上下文信息,并使用数据增强进行训练。相比表1中的其他算法,DSFD在WiderFace上的各尺寸人脸均获得了最好结果,但是由于训练网络过深,最终得到的模型权重也很大,不适合直接应用于视频监控等端侧设备。RetinaFace(R)使用ResNet50作为骨干网络,由于借鉴了性能优异的通用目标检测网络RetinaNet[13],并采用FPN[13]和SSH结构加强对多尺度特征的提取,RetinaFace(R)在检测性能上取得了不错的结果。相比于其他采用VGG16作为骨干网络的检测算法,RetinaFace(R)在内存占用上更有优势。

表1 不同人脸检测方法的综合对比

Table 1 Comprehensive comparison of different face detection methods

算法	骨干网络	模型权重大小/MB	检测精度/%		
			简单子集	中等子集	困难子集
DSFD	ResNet152	458	96.6	95.7	90.4
PyramidBox	VGG16	516	96.1	95.0	88.9
RetinaFace(R)	ResNet50	104	95.5	94.0	84.4
S^3FD	VGG16	557	93.7	92.5	85.9
SSH	VGG16	280	93.1	92.1	84.5
HR	VGG16	478	92.5	91.0	80.6
CMS-RCNN	VGG16	—	89.9	87.4	62.4
RetinaFace	MobileNet0.25	1.7	90.7	88.1	73.8
MTCNN	P/R/O Net	11.1	84.9	82.5	59.8
LDCF+	—	—	79.0	76.9	52.2
Faceness	AlexNet	—	71.4	63.4	34.5
本文方法	MobileNet0.25	3.4	92.1	89.7	77.7

CMS-RCNN[6]、HR[7]、SSH[8]、S^3FD[9]和PyramidBox[10]检测算法均使用VGG16作为骨干网络,PyramidBox通过设计低层特征金字塔融合上下文和人脸特征,提升模型在各尺度人脸上的性能,但是由于使用VGG16带来更多参数,导致更多内存占用。S^3FD和SSH均是单阶段人脸检测算法,且检测性能相近,但SSH算法在内存占用上更有优势。HR算法是专门针对小脸提出的,通过从尺度不变、图

像分辨率和上下文信息3个方面对小脸检测展开了深入探究，但模型内存也较大，实时性不好。CMS-RCNN是唯一的两阶段人脸检测算法，受两阶段通用目标检测器启发，该方法使用人脸上下文信息加强多尺度人脸检测，相比于级联算法，该方法在当时取得了最好结果。

MTCNN[5]采用图像金字塔和级联网络实现人脸检测。通过多任务训练方式，将滑动窗口替换为卷积，有效提升了模型在各个尺度上的检测精度，但在小尺寸人脸上结果仍不够理想。Faceness[4]采用AlexNet作为骨干网络提取特征，使用人脸部位分类器对人脸进行打分，然后回归得到感兴趣人脸区域，但是该方法在小脸上的结果最不理想。LDCF+[1]是唯一一个非深度方法，通过对弱学习器体系结构改进，进而提升增强树算法的建模能力。相比于性能优异的深度方法，经典机器学习方法受限于在有限的数据上对问题建模，因此模型性能存在瓶颈。

本文提出的DLFace使用改进后的MobileNet0.25作为骨干网络，相比于RetinaFace仅有1.7 MB的模型内存，DLFace的内存占用为3.4 MB，主要是因为引入DCNv2[14]后带来了更多参数量和计算量。但是DLFace相比RetinaFace，检测性能在WiderFace三个子集上分别提升1.4%、1.6%和3.9%，且DLFace的检测结果和使用VGG16作为骨干网络的HR算法最接近。这主要是因为引入Lambda层后，加强了上下文联系，丰富了位置和语义信息，引入DCNv2强化了对于小脸特征的精细化提取。本文提出的方法在内存占用和检测精度上都取得了平衡。

考虑到模型参数量、计算量及推断时间对于全面评价模型十分重要，本文对MTCNN、RetinaFace、RetinaFace(R)、DLFace模型的参数量、运算量、平均推断时间做了对比。如表2所示，FLOPs为每秒执行的浮点运算次数，参数量和运算量越大，意味着模型训练过程越耗时。可以看出，MTCNN由于仅使用3个简单的级联网络，算法的运算量最低。由于引入深度可分离卷积大大降低参数量，RetinaFace在参数量和平均推断时间上更有优势。RetinaFace(R)由于使用ResNet50作为骨干网络在检测结果上取得了明显优势，但是却是以牺牲时间和内存空间为代价。本文提出的DLFace由于引入了DCNv2，参数量和计算量均有小幅增加，但是引入Lambda层在一定程度上降低了计算量，且DLFace的平均推断时间只有66 ms，实时性较好，说明DLFace综合性能最优。

表2 不同算法的参数量、运算量和推断时间对比

Table 2 Comparison of the number of parameters, computation and inference time of different algorithms

算法	参数量/10^3	FLOPs/10^6	单张平均推断时间/ms
MTCNN	0.5	0.77	85
RetinaFace	0.44	1.02	61
RetinaFace(R)	27.29	44.52	240
DLFace	1.36	1.58	66

为了进一步验证文中提出的DLFace模型的合理性，进行了消融实验，结果如表3所示。表3第1组为使用原始MobileNet骨干网络的RetinaFace[12]结果。第2组为基于改进后MobileNet骨干网络实验结果。在WiderFace数据集不同子集上AP值相比第1组网络分别提升了0.1%、0.2%和0，说明Leaky-ReLU可以改善特征提取过程中信息丢失问题。第3组网络在骨干网络中引入了Lambda层，在WiderFace数据集不同子集上的AP值相比第2组网络分别提升了0.6%、0.6%和0.7%，在3种不同尺度人脸上的提升效果相当。这是因为在骨干网络高层引入Lambda层后，有效增强特征上下文间的联系，提升高层特征图中的语义和位置信息，增强了特征图表达能力。第4组网络基于第3组网络，在SSH[8]检测模块中引入了可变形卷积DCNv2，在不同子集上的AP值相比第3组网络分别提升了0.3%、0.7%和1.7%，提升最高的是小尺寸人脸，这是因为特征经过SSH检测模块后再经过DCNv2，能进一步对通道拼接后的特征图精细化提取特征。相比于大尺寸人脸，小尺寸人脸的变化有限，受光照、遮挡等因素的影响更小。因此对于小脸，其拟合程度最好，在该尺度上性能提升最明显。第5组网络基于第4组网

络,在骨干网络中引入了DCNv2,在不同子集上的AP值相比第4组网络分别提升了0.5%、0.3%和1.5%,提升最高的仍为小尺寸人脸,这是因为在骨干网络3个阶段的低层引入DCNv2后,网络提取的不同尺度的特征图更多覆盖到了人脸轮廓上。DCNv2在增大感受野的同时提升对各尺寸人脸特征的拟合能力,相比于大脸,小脸在特征提取过程中引入的干扰更少,因此对于小脸性能提升最多。第5组网络在3种子集图片上的检测精度均达到了最优值,说明本文提出DLFace结构的合理性和有效性。

表3 引入不同结构后在WiderFace上的消融实验

Table 3 Ablation experiments on WiderFace after introducing different structures

ID	I-MobileNet	Lambda层	DCNv2(SSH)	DCNv2(Backbone)	AP/%		
					简单子集	中等子集	困难子集
1	×	×	×	×	90.6	87.9	73.8
2	√	×	×	×	90.7	88.1	73.8
3	√	√	×	×	91.3	88.7	74.5
4	√	√	√	×	91.6	89.4	76.2
5	√	√	√	√	**92.1**	**89.7**	**77.7**

图9为DLFace在中小目标上的检测结果,可以看出DLFace可以有效检测出复杂场景中的人脸目标,并且能很好解决大规模人群中多尺度、大姿态和高遮挡的人脸检测难题,说明DLFace能较好地应用于视频监控场景下的低清人脸检测任务。

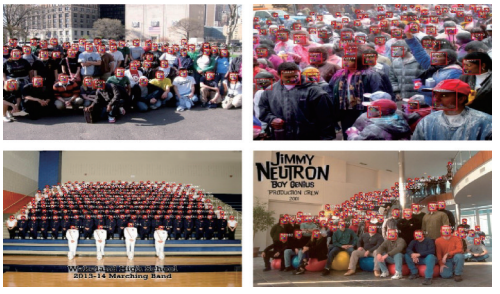

图9 DLFace在中小目标上的检测效果

Fig.9 Detection effect of small and medium targets using DLFace

为了更直观地展示DLFace方法的优势,本文将RetinaFace和DLFace的检测视觉效果进行了对比。图10为RetinaFace和DLFace在密集小脸、大姿态、遮挡、大表情以及光照条件下的对比结果。从图中可以看出,RetinaFace在小脸、侧脸、部分遮挡和表情变化上都取得了不错的结果,但是相比于DLFace还稍显逊色。RetinaFace在密集小脸上仍存在很多漏检,对于姿态、表情变化人脸特征的提取能力有待提升。DLFace由于引入了DCNv2,网络在对人脸特征提取过程中考虑到形态变化,提取的特征图更多覆盖在人脸轮廓上,对于小脸检测性能提升尤为明显。同时,DLFace由于引入Lambda层,在高层特征中同时兼顾语义和位置信息,加强局部上下文之间联系,助力模型整体检测性能提升。

4 结束语

针对视频监控场景下低清人脸检测问题,本文提出轻型尺度自适应低清人脸检测模型DLFace。首先借鉴RetinaFace算法,使用改进后的深度可分离卷积能够防止训练过程中信息丢失;其次将改进后的可变形卷积引入骨干网络和SSH检测模块,通过增强感受野适应人脸多因素变化;最后在骨干网络高层引入Lambda层,有效挖掘语义和位置信息,形成更加准确丰富的特征表示。在WiderFace数据集上的实验结果表明,DLFace实现了性能和速度的平衡,在不同场景下均验证了DLFace的优越性,表明DLFace能较好地适用于视频监控场景下的低清人脸检测任务。

图 10 RetinaFace 和 DLFace 在不同场景下的检测效果

Fig.10 Detection effect in different scenarios using RetinaFace and DLFace

参考文献：

[1] OHN-BAR E, TRIVEDI M M. To boost or not to boost? On the limits of boosted trees for object detection[C]//Proceedings of 2016 23rd International Conference on Pattern Recognition (ICPR). [S.l.]: IEEE, 2016: 3350-3355.

[2] YANG S, LUO P, LOY C C, et al. Wider face: A face detection benchmark[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2016: 5525-5533.

[3] LI H, LIN Z, SHEN X, et al. A convolutional neural network cascade for face detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2015: 5325-5334.

[4] YANG S, LUO P, LOY C C, et al. Faceness-Net: Face detection through deep facial part responses[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(8): 1845-1859.

[5] ZHANG K, ZHANG Z, LI Z, et al. Joint face detection and alignment using multitask cascaded convolutional networks[J]. IEEE Signal Processing Letters, 2016, 23(10): 1499-1503.

[6] ZHU C, ZHENG Y, LU K, et al. CMS-RCNN: Contextual multi-scale region-based CNN for unconstrained face detection

[M]//Deep Learning for Biometrics. Cham: Springer, 2017: 57-79.

[7] HU P, RAMANAN D. Finding tiny faces[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2017: 951-959.

[8] NAJIBI M, SAMANGOUEI P, CHELLAPPA R, et al. SSH: Single stage headless face detector[C]//Proceedings of the IEEE International Conference on Computer Vision. [S.l.]: IEEE, 2017: 4875-4884.

[9] ZHANG S, ZHU X, LEI Z, et al. S3FD: Single shot scale-invariant face detector[C]//Proceedings of the IEEE International Conference on Computer Vision. [S.l.]: IEEE, 2017: 192-201.

[10] TANG X, DU D K, HE Z, et al. Pyramid box: A context-assisted single shot face detector[C]//Proceedings of the European Conference on Computer Vision (ECCV). [S.l.]:[s.n.], 2018: 797-813.

[11] LI J, WANG Y, WANG C, et al. DSFD: Dual shot face detector[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2019: 5060-5069.

[12] DENG J, GUO J, ZHOU Y, et al. Retinaface: Single-stage dense face localisation in the wild[EB/OL].(2019-05-04)[2021-09-10]. https://arxiv.org/abs/1905.00641v2.

[13] 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. [S.l.]: IEEE, 2017: 2980-2988.

[14] ZHU X, HU H, LIN S, et al. Deformable convnets v2: More deformable, better results[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2019: 9308-9316.

[15] BELLO I. Lambdanetworks: Modeling long-range interactions without attention[EB/OL].(2021-02-17)[2021-09-10]. https://arxiv.org/abs/2102.08602.

[16] DAI J, QI H, XIONG Y, et al. Deformable convolutional networks[C]//Proceedings of the IEEE International Conference on Computer Vision. [S.l.]: IEEE, 2017: 764-773.

[17] ZHANG H, GOODFELLOW I, METAXAS D, et al. Self-attention generative adversarial networks[C]//Proceedings of International Conference on Machine Learning. [S.l.]: PMLR, 2019: 7354-7363.

[18] SIFRE L, MALLAT S. Rigid-motion scattering for texture classification[EB/OL]. (2014-03-17)[2021-09-10]. https://arxiv.org/abs/1403.1687.

[19] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]//Proceedings of International Conference on Machine Learning. [S.l.]: PMLR, 2015: 448-456.

[20] HOWARD A G, ZHU M, CHEN B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications [EB/OL]. (2017-04-17)[2021-09-10]. https://arxiv.org/abs/1704.04861.

作者简介:

胡洪明(1994-),男,硕士研究生,研究方向:深度学习、低清人脸检测与识别,E-mail:hminghu@126.com。

邵文泽(1981-),**通信作者**,男,副教授,研究方向:计算成像、计算机视觉、黑箱优化,E-mail:shaowenze@njupt.edu.cn。

李金叶(1997-),女,硕士研究生,研究方向:深度学习、低清人脸幻构与检测。

葛琦(1984-),女,副教授,研究方向:图像处理与视觉理解。

邓海松(1980-),女,副教授,研究方向:黑箱优化与计算成像。

(编辑:张黄群)