面向算力受限边缘环境的双分支多尺度感知人脸检测网络

戚琦，马迎新，王敬宇，孙海峰，廖建新
（北京邮电大学网络与交换国家重点实验室，北京 100876）

摘　要：针对边缘算力受限，难以部署复杂结构的人脸检测深度神经网络的问题，为减少资源消耗，并保证人脸在多尺度变化、遮挡、模糊、光照等复杂场景下的检测精度，提出了多尺度感知的轻量化人脸检测算法。采用改进的人脸残差神经网络作为特征提取网络，并提出双分支浅层特征提取模块，并行分支理解图像多尺度信息，进而由深浅特征融合模块将底层图像信息与高层语义特征融合，配合多尺度感知的训练策略监督多分支学习差异化特征。实验结果表明，所提算法可有效提取多样化的特征，在保持模型高效性和低推理时延的同时，有效提升了算法的精度和稳健性。

关键词：人脸检测；多尺度感知；特征融合；人脸特征分析；深度学习
中图分类号：TP391.41
文献标识码：A
doi: 10.11959/j.issn.1000−436x.2020177

Multi-scale aware dual path network for face detection in resource-constrained edge computing environment

QI Qi, MA Yingxin, WANG Jingyu, SUN Haifeng, LIAO Jianxin
State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China

Abstract: Aiming at the problem that face detectors with complex deep neural structures are difficult to deploy in the resource-constrained edge computing environment, to reduce the resource consumption while maintain the accuracy in complex scenes such as multi-scale face changes, occlusion, blur, and illumination, SDPN(multi-scale aware dual path network) for face detection was proposed. The Face-ResNet (face residual neural network) was improved, and a dual path shallow feature extractor was used to understand the multi-scale information of the image through parallel branches. Then the deep and shallow feature fusion module, a combination of the underlying image information and the high-level semantic feature, was used in conjunction with the multi-scale awareness training strategy to supervise the multi-branch learning discriminating features. The experimental results show that SDPN can extract more diversified features, which effectively improve the accuracy and robustness of face detection while maintaining the efficiency of the model and low inference delay.

Key words: face detection, multi-scale aware, feature fusion, analysis of facial features, deep learning

1　引言

随着物联网时代的到来，边缘计算将与云计算共同推进人工智能的发展。边缘计算的推广与发展可将计算任务从云计算中心迁移到产生源数据的边缘设备上。车联网的车载系统、机器人控制系统、工业应用等边缘环境部署的英伟达开发板 Jetson TX2 和 Jetson Nano 设备，只需几瓦的功

收稿日期：2020−05−13；**修回日期**：2020−07−12
通信作者：王敬宇，wangjingyu@bupt.edu.cn
基金项目：国家重点研发计划基金资助项目（No.2018YFB1800502）；国家自然科学基金资助项目（No.61671079, No.61771068）；北京市自然科学基金资助项目（No.4182041）
Foundation Items: The National Key Research and Development Program of China (No.2018YFB1800502), The National Natural Science Foundation of China (No.61671079, No.61771068), The Beijing Municipal Natural Science Foundation (No.4182041)

耗即可实现每秒进行上万亿次操作。基于深度学习的卷积神经网络已被广泛应用于人脸检测中，并尝试在边缘设备中部署，以减少网络传输时延，加快检测速度。

然而，边缘算力受限于内在或外在因素等条件。主流深度神经网络结构[1-5]趋于复杂，受限于网络复杂度及推理速度，难以在边缘计算实际应用场景中部署，需考虑精简网络结构来保持简洁和高效性。此外，在支撑人脸检测等深度学习任务推理的同时，边缘设备还需要为其他功能预留和共享资源，应尽可能探索轻量级方案，以减少算法所对应网络模型的复杂度和功耗。

围绕人脸检测而设计的轻量化网络算法[6-7]大幅缩减了网络结构来保持推理速度和资源消耗。例如，LFFD（light and fast face detector）[6]去掉了网络中的 BN（batch normalization）层等重要组件以减少网络规模，但未重视网络特征融合上的冗余优化，导致其精度与主流算法相差较多。本文针对主流算法潜在的结构冗余、特征提取网络设计不合理、浅层特征不充分的问题，提出了多尺度感知的轻量化人脸检测算法——SDPN（scale-aware dual path network），在不损失检测精度的前提下优化了网络的特征融合方式，以实现速度与精度的均衡。SDPN 着重改进网络结构，提升特征融合的高效性，基于经典的 ResNet 结构中参数最少的 RetNet18 结构，采用基于双分支的多尺度感知结构，简化了网络层融合中的冗余，实现了高效的人脸检测。

在人脸检测数据集 wider face[8]中，极小人脸（人脸最长边小于 12 像素）占 30%左右，人脸尺度变化范围为 3～1 000 像素。极小、困难人脸以及大范围的人脸尺度变化问题是人脸检测任务中的主要挑战。

图像金字塔方法[1]通过综合同一图像不同缩放系数检测结果来提高多尺度检测能力。这类算法存在对同一张图片不同尺度的冗余计算，不适用于资源有限、对时延有要求的应用场景。

特征金字塔方法[2-5]注重不同层级特征的融合，将中层的高分辨率特征与深层的高语义特征进行融合，兼顾语义与位置信息，得到多尺度的特征图。但这类算法存在以下几点问题，导致精度不如图像金字塔方法。

1) 大多数人脸检测算法使用 ImageNet[9]数据集上预训练得到的图片分类模型，分类尺度大小约为 256 像素，未针对人脸数据中的大尺度范围和极小人脸进行优化。网络浅层主要使用最大池化和下采样卷积操作进行空间信息压缩。由于预训练模型浅层特征不充分，下采样过于频繁，存在极小人脸特征信息丢失、定位不准等问题。

2) 主流特征融合方式过多关注中、深网络层的特征多尺度感知融合策略，未考虑浅层多尺度感知结构对人脸检测网络的重要性。

针对上述问题，本文进行了如下改进。

1) 针对主流卷积神经网络（CNN, convolutional neural network）用于人脸检测场景时浅层特征不充分，导致下采样卷积时图像特征偏移、信息采样不均匀的问题，提出了改进的人脸残差神经网络（Face-ResNet, face residual neural network）作为主干网络。

2) 探索了浅层多尺度感知结构对极小、困难人脸特征提取的重要性。提出了高效的双分支浅层特征（DPE, dual path shallow feature extractor）提取模块，通过并行的浅层网络双分支设计，分别从上采样后的图像以及原图中提取浅层特征，在中层进行双分支特征融合，送入后续网络。只在最关键的浅层使用双特征分支获取差异化的特征信息，在后续的网络层中共享 2 个特征融合后的信息，实现了高层网络特征复用，大幅提升了网络对于极小的、难以分辨的人脸特征的识别能力，接近特征金字塔方法的融合结果。

3) 训练过程中，研究了双分支的浅层特征与深层特征的融合方式，提出了深浅特征融合模块，并通过多尺度感知的训练方式使不同的并行分支（浅、深网络层）学习到更加丰富、多样化的信息。

2 相关研究

2.1 基于锚点的检测器

随着深度学习的发展，在越来越多的领域充分证明了卷积神经网络的有效性，以 Faster-RCNN（faster region-based convolutional neural network）[10]为代表的两阶段基于锚点的检测器在目标检测领域成为主流。近年的人脸检测中，基于区域推荐网络（RPN, region proposal network）[11]的多阶段人脸检测器[1,12-14]在精度上取得了质的飞跃，但受限于烦琐的检测流程，速度并不理想，同时网络模块较为复杂，难以在终端部署进行工程应用。最近，基

于锚点的一阶段检测器[2,4-5,15-17]逐渐占据主流，得益于流程的简化，网络的速度也有提升，相对容易在终端应用中部署。

2.2　基于关键点的无锚点检测器

随着 CornerNet[18]的提出，目标检测领域出现了新的无锚点的算法，这类算法受启发于人体关键点检测的算法，通过在网络最后的特征图中直接预测关于物体的几个关键点，以及在并行分支预测物体的大小和偏移等信息，得到关于物体的检测结果。在目标检测领域，这类算法[18-19]相比一阶段、多阶段的基于锚点的算法，更加高效，速度更快，同时呈现出与一阶段基于锚点的检测器相近的精度。在人脸检测中，CSP（center and scale prediction）[20]是基于关键点设计的算法，网络设计进一步简化，由于不需要锚点匹配等流程，速度得到进一步提升。本文的改进借鉴了 CSP 算法的无锚点设计，深入研究了该算法的不足，并提出了新的解决方法。

2.3　基于卷积网络的人脸特征提取方式

基于卷积网络的人脸特征的提取方式主要有图像金字塔和特征金字塔。

如图 1 所示，特征金字塔[2-5]将浅层的高分辨率特征与深层的语义特征相融合，通过不同特征层之间的融合，同时提高了识别准确率和定位精度。不同特征融合网络结构在 FTF（finding tiny face）[1]、MTCNN（multi-task cascaded convolutional network）[13]等为代表的图像金字塔算法中，将同一图片的不同尺度依次送入网络，由非极大值抑制（NMS, non-maximum suppression）算法进行融合，留下置信度比较高的融合结果。本文 SDPN 选择性地进行双分支差异化特征提取，并通过深浅融合以及尺度感知的训练方式，高效提取不同尺度的特征。

3　网络结构

3.1　ResNet 在人脸检测领域的改进

针对 ResNet[21]网络中的特征偏移、下采样不均匀、浅层特征不充分问题，本文提出了改进的 Face-ResNet 结构，并通过对比实验证明了其有效性。

下采样卷积被广泛应用于卷积神经网络，目的是得到更小分辨率的特征图，同时为后续的卷积层提供更大的感受野，获得更加丰富的语义信息。很多主流算法使用步长为 2 的 3×3 卷积核来完成下采样。事实上，这样的操作有 2 个潜在的问题，首先是采样信息不充分问题。如图 2 所示，在一个 3×3 卷积核采样卷积过程中，有 $\frac{1}{4}$ 的点被采样了 4 次和一次，$\frac{1}{2}$ 的点被采样了 2 次（考虑最外层像素，上述采样频率仅为近似结果），这表明卷积过程的信息并没有均匀地被采样并送入后续网络中。尽管网络本身有能力通过学习不同的权重来纠正采样问题，但这会浪费网络的一部分参数来学习纠正问题，是不高效的。同样的现象被 Odena 等[22]发现，称之为棋盘效应。

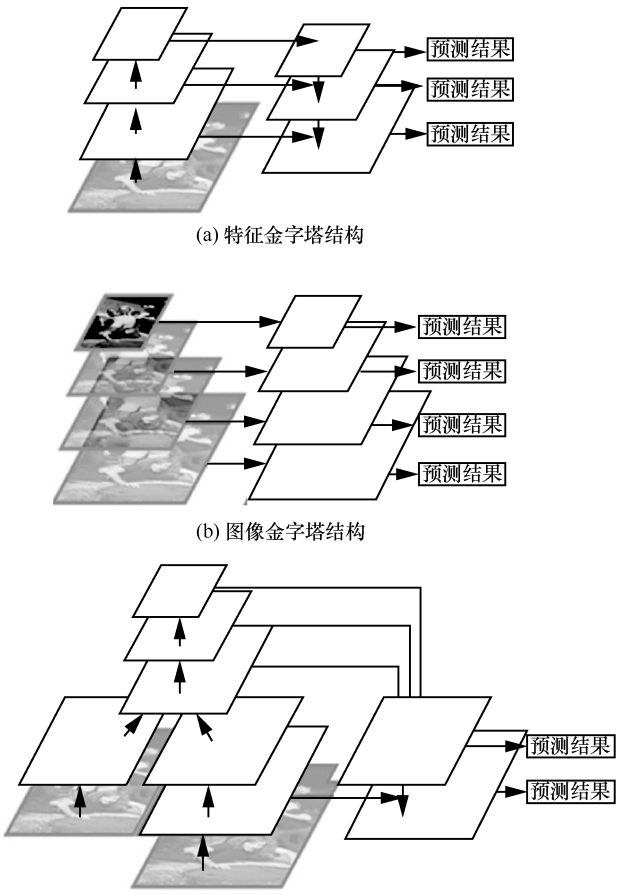


图 1　不同特征融合网络结构

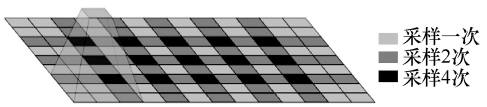
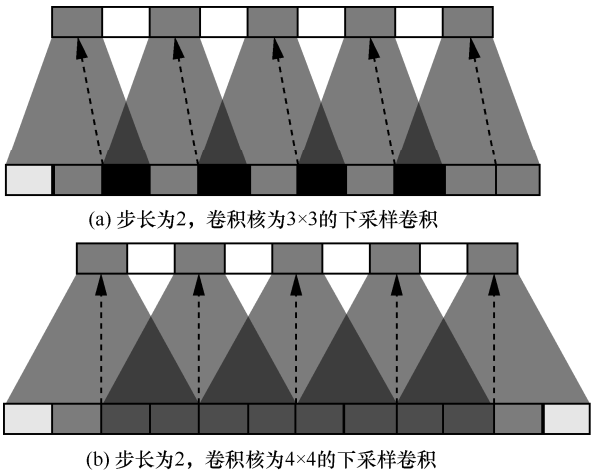


图 2　3×3 卷积核下采样过程

另一个问题是下采样会导致特征图发生偏移，卷积过程中的特征图偏移问题如图 3 所示。下采样的过程中，如果卷积核不能整除步长，特征图的两边不能填充同样多的像素，输出的特征图相对原来位置会发生 0.5 像素的偏移，如图 3(a)所示。



(a) 步长为2，卷积核为3×3的下采样卷积

(b) 步长为2，卷积核为4×4的下采样卷积

图 3　卷积过程中的特征图偏移问题

网络最终输出特征图（缩放倍数为 i）相对原图的特征偏移大小 S_{offset} 为

$$S_{\text{offset}}(i)=\sum_{2^l}^{2^l<i}P_l 2^l \tag{1}$$

其中，P_l 为卷积时的 padding 偏移，l 为网络当前相对于原图的缩放倍数。卷积核为 3×3，步长为 2，标准 ResNet 网络中 $P_l=0.5, i=16, S_{\text{offset}}=7.5$；卷积核为 4×4，步长为 2，ResNet 网络中 $P_l=0, i=16, S_{\text{offset}}=0$。

对于极小人脸，偏移现象会显著影响定位能力。同时随着网络层的加深，分辨率逐渐减小，偏移现象会更加严重。网络需要学习更多的权重来纠正这种偏移现象，但不能完全抵消这种偏移和采样不充分带来的影响。

为了解决这一问题，本文研究了使用不同大小的卷积核进行下采样的效果。如图 3(b)所示，4×4 的卷积核可以使每个点都被充分采样（不考虑边缘点）；同时，由于两边填充了相同数量的像素，没有出现特征图偏移现象。本文替换了网络中降采样卷积层的卷积核，将原网络结构分别替换为 4×4 和 5×5 的卷积核，并进行对比实验，证明了 4×4 卷积核的有效性，详细实验在 4.1 节给出。

同时，本文使用通道更少的残差模块代替 ResNet 中的 7×7 卷积层和 maxpool 层来改善 ResNet 对细小特征的表达能力。SDPN 整体结构如图 4 所示，修改后的 ResNet 主结构由双分支浅层特征提取模块的下分支以及后续的主分支组成。经过上述改进，最终的主干网络结构为 Face-ResNet 结构。

3.2　双分支浅层特征提取模块

人脸检测中，发现极小人脸是一个极具挑战的问题，SNIP（scale normalization for image pyramid）[23] 的研究结果表明，高分辨率预训练模型预测上采样后的低分辨率物体，准确率高于针对低分辨率物体

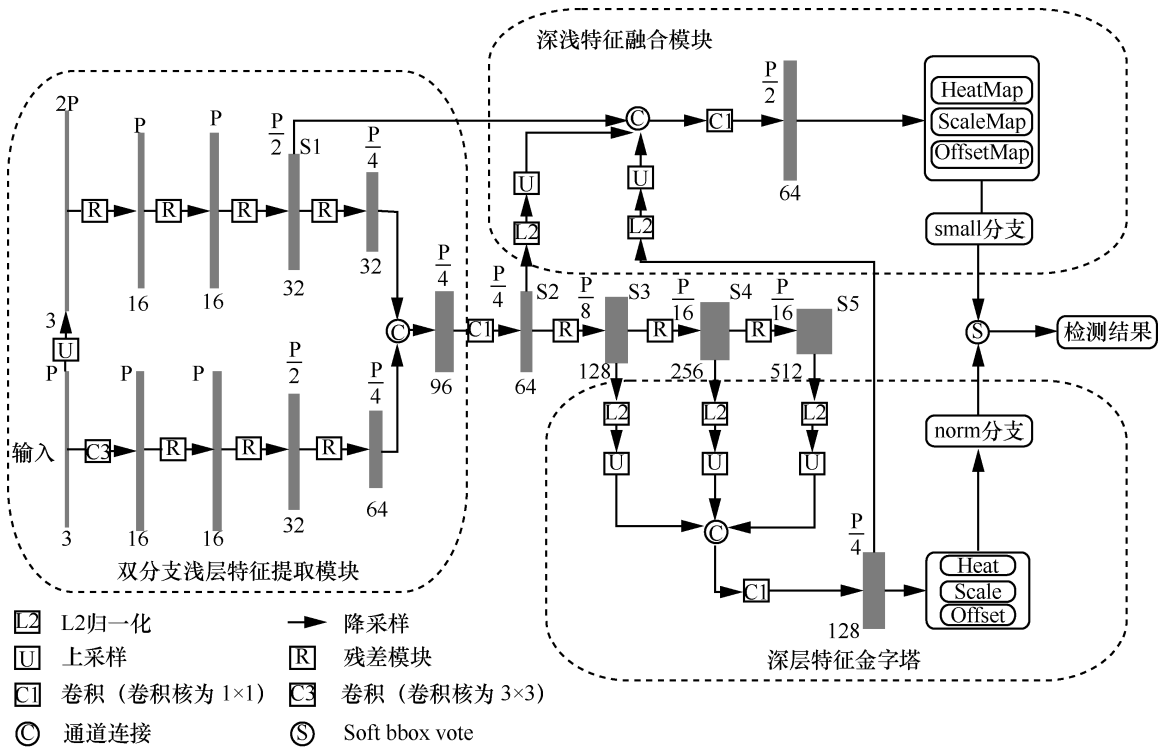


图 4　SDPN 整体结构

设计并训练的网络模型。现有的人脸检测算法使用图像金字塔方法将同一张图片进行不同上采样系数的测试结果进行融合，从而获得更好的检测结果，但是这会牺牲推理时延和计算资源。

本文针对这一问题，设计了双分支浅层特征提取模块来获取图片中不同分辨率的信息，在浅层融合共享网络中的高层语义信息来实现检测精度与速度的均衡。并行于 Face-ResNet 网络的浅层分支，添加了一个更轻量级的高分辨率（HR, high resolution）分支，该分支参数数量和卷积网络层中的通道数只有原分支的$\frac{1}{4}$，使用经过上采样的图片作为输入，主要负责发现更多难以辨认的、细小的细粒度信息。相比原分支，HR 分支末端加入了一个降采样分支，来达到与原分支相同的分辨率。在中层，2 个独立分支融合为一个特征层，送入后续主干网络中，计算式如式(2)~式(4)所示。

$$y_1 = f(x_1)$$
$$y_2 = h(x_2) \tag{2}$$
$$m = j(y_1 \bigcup y_2) \tag{3}$$
$$z = \varphi(m) \tag{4}$$

wider face 数据集中的极小人脸集中在 10~30 像素。在 ResNet 中，网络$\frac{P}{4}$层的神经元感受野大小为 7，仍然能够获取足够丰富的低分辨率极小人脸特征信息；$\frac{P}{2}$层网络深度过浅，无法获取足够的语义信息；后续层由于感受野变大，网络深度增加，浅层低分辨率极小人脸信息丢失较多，同时由于过多的池化和卷积操作将导致位置信息丢失的问题。故选择 P_4 层作为特征融合层。

经过上述双分支采样设计，网络在前向推断过程可以学到更多细小的、难以辨认的特征。在获取更加丰富浅层特征的同时，添加的新分支只带来了轻微的计算消耗。

3.3　深浅特征融合模块

DPE 提取模块的 HR 分支主要负责获取极小的、难以辨认的一些人脸特征，而主分支则负责检测常规人脸特征，为了通过监督信息更好地指导 2 个分支学到不同的特征，本文提出了深浅特征融合模块。深浅特征融合模块由深度特征融合层以及深浅特征融合层两部分构成。深度融合层借鉴自 CSP，选择了图 4 结构中的 S_3～S_5 特征层信息进行融合，对每个特征层使用 L2 正则化处理，通过上采样卷积恢复特征图为输入图像尺寸的$\frac{1}{4}$，检测正常大小的人脸。深浅特征融合层是针对极小人脸专门设计的多尺度特征融合模块，将更浅层分支的特征与深度特征融合层信息进行融合，只负责检测极小物体。经过实验验证，图 4 中 S_1、S_2，以及深度特征层进行融合可以得到最好的结果。为了获得更精准的人脸中心位置信息，输出特征图的尺寸为原图的$\frac{1}{2}$。在训练中，本文采用多尺度感知策略，根据物体的尺度大小分别送入不同的特征融合层进行训练。通过多尺度感知的训练方式，监督信息可以更有效地指导 2 个分支学到更具差异化的特征信息，从而具备了更丰富、全面的特征，提升了模型的准确率和有效性。在实验部分，进一步分析了每个模块对整体性能的影响，并进行了对比实验。

3.4　loss 函数设计

本文基于无锚点目标检测网络 CenterNet[19]的思路提出了多尺度感知的 loss 函数，用于监督不同分支学习差异化的特征。

如图 4 所示，本文算法所采用的卷积网络模型通过 2 个分支分别输出 small 和 norm 特征图，分辨率分别为$\frac{H}{2}\times\frac{W}{2}$和$\frac{H}{4}\times\frac{W}{4}$。

对于 norm 特征图，由于分辨率只有输入的$\frac{1}{4}$，特征图预测的中心位置在原图上是离散化的、非连续的。通过引入离散纠正特征层可以纠正预测的离散化误差，得到连续的中心位置预测结果。norm 特征图包括中心、尺度和离散偏差纠正特征层，分别对应人脸的中心预测、长宽预测，以及离散值偏差的纠正任务。

small 特征图分辨率为原图的$\frac{1}{2}$，与原图较接近。实验证明，离散纠正特征在 small 特征图上不能提升预测精度。故 small 特征图仅包含了中心、尺度特征层。

中心预测任务可以看作人脸检测网络在图像空间维度的分类任务，其目标是通过分支的中心特征图得到人脸中心点。本文通过交叉熵损失函数来监督网络输出的人脸中心特征图。首先，生成与人脸大小接近的二维高斯热力图作为监督信号。同时，规定人脸中心的小范围区域为正样本，其他为负样本。图像中人脸中心区域与背景点之间在严重

的正负样本不均衡问题，背景点数量远大于中心点。随着模型的收敛，模型倾向于更新易于学习的大量负样本（背景）特征的权重。为了解决样本不均衡问题，本文的交叉熵损失函数引入了 focal loss[24]来优化对困难样本的监督能力。为了避免中心区域附近的点对模型产生人脸和背景分类上的歧义，loss 函数的权重参数 P_{xy} 根据当前点的 M_{xy} 和 Y_{xy} 值来调整，使用 P_{xy} 将这类点在 loss 函数中的权重降低，减少对模型收敛的影响。Y_{xy}=1 表示当前点为中心点。

关于中心点预测的损失函数定义如式(5)所示。O_{xy} 表示中心特征图上点$[x,y]\in\left[\frac{W}{4},\frac{H}{4}\right]$的响应值，其值为 0～1。$M_{xy}$ 表示根据人脸标签生成的高斯热力图在点$[x,y]\in\left[\frac{W}{4},\frac{H}{4}\right]$对应取值，$\sigma_{\mathrm{w}}$、$\sigma_{\mathrm{h}}$ 为与人脸长宽线性相关的参数。依照 CenterNet[19]中的设置 β=4、γ=2。

$$\mathrm{hmloss}=-\frac{1}{K}\sum_{xy}P_{xy}{O_{xy}}^{\gamma}\log\left(O_{xy}\right) \tag{5}$$

$$O_{xy}=\begin{cases}1-O_{xy}, Y_{xy}=1\\ O_{xy}, \text{其他}\end{cases} \tag{6}$$

$$P_{xy}=\begin{cases}1, Y_{xy}=1\\ \left(1-M_{xy}\right)^{\beta}, \text{其他}\end{cases} \tag{7}$$

$$M_{xy}=\mathrm{e}^{-\left(\frac{(x-x_0)^2}{2{\sigma_{\mathrm{w}}}^2}+\frac{(y-y_0)^2}{2{\sigma_{\mathrm{h}}}^2}\right)} \tag{8}$$

预测分支进行回归任务，输出为每个点对应人脸的长和宽的值 s_k，与人脸标签中的长和宽 t_k 进行对比，其中，k=0 表示长，k=1 表示宽。本文通过 $L1$ 损失函数完成监督，如式(9)所示。

$$\mathrm{whloss}=\frac{1}{K}\sum L1\left(s_k,t_k\right) \tag{9}$$

离散误差纠正分支的损失函数与尺度预测分支一致，输出为图像中心点的离散值对应的偏移误差纠正值，offloss 表示离散偏差 loss。

本文人脸检测算法对应模型的 loss 如式(10)所示，small、norm 分支的损失函数分别为 $\mathrm{loss}_{\mathrm{small}}$、$\mathrm{loss}_{\mathrm{norm}}$。其中，$\theta_i$、$\beta_i$、$\gamma_i$ 分别表示训练时不同 loss 对应的权重。

$$\mathrm{loss}=\theta_1\mathrm{loss}_{\mathrm{norm}}+\theta_2\mathrm{loss}_{\mathrm{small}} \tag{10}$$

$$\mathrm{loss}_{\mathrm{norm}}=\alpha_1\mathrm{hmloss}_{\mathrm{norm}}+\beta_1 w\mathrm{hloss}_{\mathrm{norm}}+\gamma_1\mathrm{offloss}_{\mathrm{norm}} \tag{11}$$

$$\mathrm{loss}_{\mathrm{small}}=\alpha_2\mathrm{hmloss}_{\mathrm{small}}+\beta_2\mathrm{whloss}_{\mathrm{small}} \tag{12}$$

4 实验

本文通过 6 组对比实验验证了 SDPN 每个改进方面的有效性，并展示了与主流人脸检测算法的对比结果。

4.1 实验设置

本文实验在 wider face[8]数据集上进行评估，根据人脸的识别难度分别划分为简单、中等、困难这 3 个子集。人脸范围为 3～1 000 像素。实验使用 ImageNet[9]数据集预训练模型权重进行初始化，与主流做法保持一致。在 wider face 训练集上进行训练，采用 Adam 梯度更新算法，初始学习率为 1×10^{-3}，权重衰减设置为 1×10^{-5}，采用动态学习率衰减策略，即当模型的预测损失结果经过 5 轮迭代不下降后，调整学习率为当前学习率的 20%。实验共迭代训练 200 轮，选取评估损失结果最低的模型权重在 wider face 验证集上进行准确率评估。

评估标准如下：在 wider face 的验证集上进行原分辨率测试来评估结果。原分辨率测试是指每个样本仅使用对应图片原始分辨率输入网络进行推理，并且网络仅推理一次，为保持对比的模型统一，均使用 Pytorch 框架复现，测试阶段在一块显存为 12 GB 的 P100 上实验并对比结果。

最终的检测精度是指网络检测结果与真实标签的查准率和召回率相等时对应的精度。

4.2 不同算法对比实验

在对比实验中，首先验证下采样卷积以及浅层参差模块修改的有效性，在此基础上验证了 DPSE 以及深浅特征融合的有效性，共进行 6 组实验。

实验 1 使用 ResNet18 作为主干网络，其余模块与 CSP 检测算法一致。

实验 2 将下采样卷积核更换为 4×4，其他与实验 1 保持一致。

实验 3 将下采样卷积核更换为 5×5，其他与实验 1 保持一致。

实验 4 将浅层网络层更换为全残差结构，其他与实验 2 保持一致。

实验 5 基于实验 4 中的网络结构，加入 DPE 提取模块。

实验 6 基于实验 5 的网络结构，加入深浅特征融合模块，训练过程中采用尺度感知的训练策略，深浅特征层负责检测尺寸小于 32 像素的人脸，深层特征负责检测尺寸大于 12 像素的人脸。

实验结果如表 1 所示。与实验 1 相比，实验 2 在 3 个数据集上的检测精度均有明显提升；而实验 3 与实验 2 相比提升效果并不明显。在 4×4 卷积核的基础上，将浅层网络部分更换为全残差结构设计，可以看到网络在困难数据集上效果提升非常显著，充分证明使用残差模块替换网络中的池化操作，可以进一步提升网络的浅层特征表达能力，加入浅层残差模块后，网络在困难样本中的极小人脸、遮挡、模糊等干扰情况下检测精度有明显提升。根据实验 4 与实验 5 的对比结果，可以看到引入的 DPE 提取模块能够显著提升算法对于极小、困难人脸的辨识度，准确率提升了 0.025。实验 5 与实验 6 的对比表明，深浅特征融合模块可以提升模型困难样本的准确率 0.06 左右，证明了深浅特征融合的有效性。与实验 5 的结果对比，实验 6 的简单子集和中等子集检测精度分别有 0.003 和 0.004 的下降，这是由于深浅特征融合带来了少量的检测负样本，导致检测精度轻微下降，与困难样本准确率提升 0.06 相比影响并不明显，本文选择实验 6 中的网络模型结构为本文算法对应的最终网络模型。

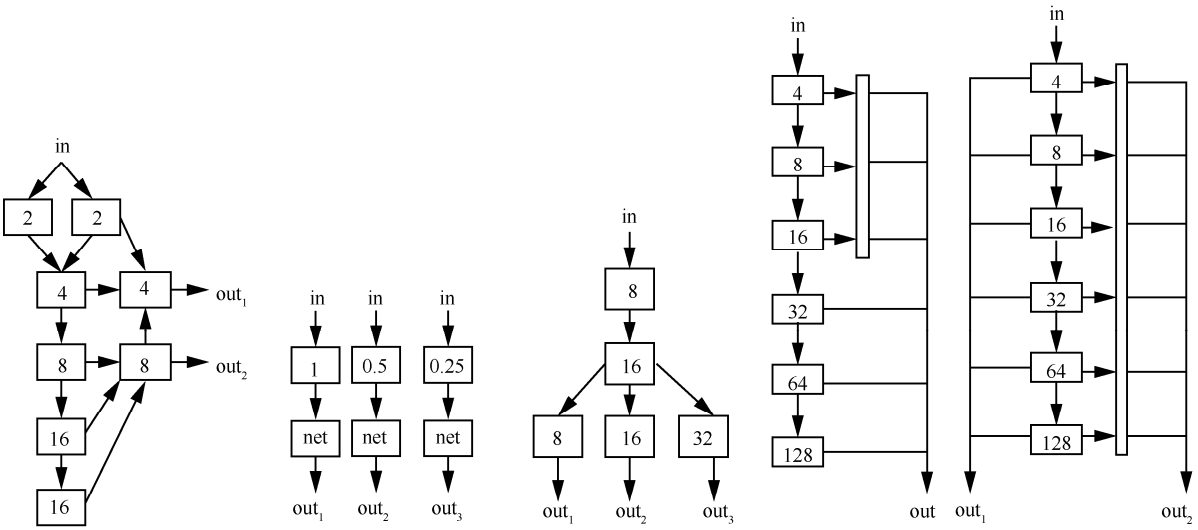
4.3　与主流算法对比实验

本文参考相关文献[1-5]，使用 Pytorch 框架复现了最近几年深度学习中的主流模型。图 5 表示不同网络模型针对多尺度的特征融合方式。

FTF 算法[1]为经典的图像金字塔融合算法。准确率较高的 DSFD（dual shot face detector）算法[3]与 PramidBox 算法[5]注重网络中深层之间的交互融合，加入了很多层间语义增强以及融合模块。SSH（single stage headless）[4]为保持推理速度和减小资源消耗，使用较为简单的特征融合策略，从网络不同层提取不同的检测结果进行融合。本文 SDPN 算法中简化了过深的冗余特征，在浅层（特征图缩放系数为 2）使用 DPE 丰富特征，并通过深浅特征融合模块进行融合，分别得到极小与正常人脸预测结果。SDPN 模型采用实验 6 的网络结构。测试过程中统计了每个网络模型的推理速度、模型参数量以及检测精度，其他实验细节与 4.2 节实验基本保持一致。测试结果如表 2 所示。反映算法查准率与召回率关系的 PR（precision recall）曲线如图 6 所示。

表 1　　**实验结果**

实验	下采样卷积核	浅层残差结构	DPE 提取模块	深浅特征融合	检测精度		
					简单	中等	困难
1	3×3	×	×	×	0.927	0.922	0.673
2	4×4	×	×	×	0.933	0.925	0.674
3	5×5	×	×	×	0.934	0.925	0.677
4	4×4	√	×	×	0.940	0.928	0.767
5	4×4	√	√	×	0.940	0.931	0.792
6	4×4	√	√	√	0.937	0.927	0.849



(a) SDPN　　(b) FTF　　(c) SSH　　(d) PyramidBox　　(e) DSFD

图 5　不同网络模型针对多尺度的特征融合方式

表 2　　**与主流算法原分辨率测试结果对比**

模型	检测精度			推理速度/(frame· s^{-1})	模型参数量/MB
	简单	中等	困难		
FTF[1]	0.819	0.797	0.638	5.3	114.54
DSFD[3]	0.949	0.936	0.845	1.7	457.99
PyramidBox[5]	0.945	0.934	0.850	3.8	256.61
SSH[4]	0.919	0.907	0.814	3.8	75.41
SDPN	0.937	0.927	0.849	6.7	65.06

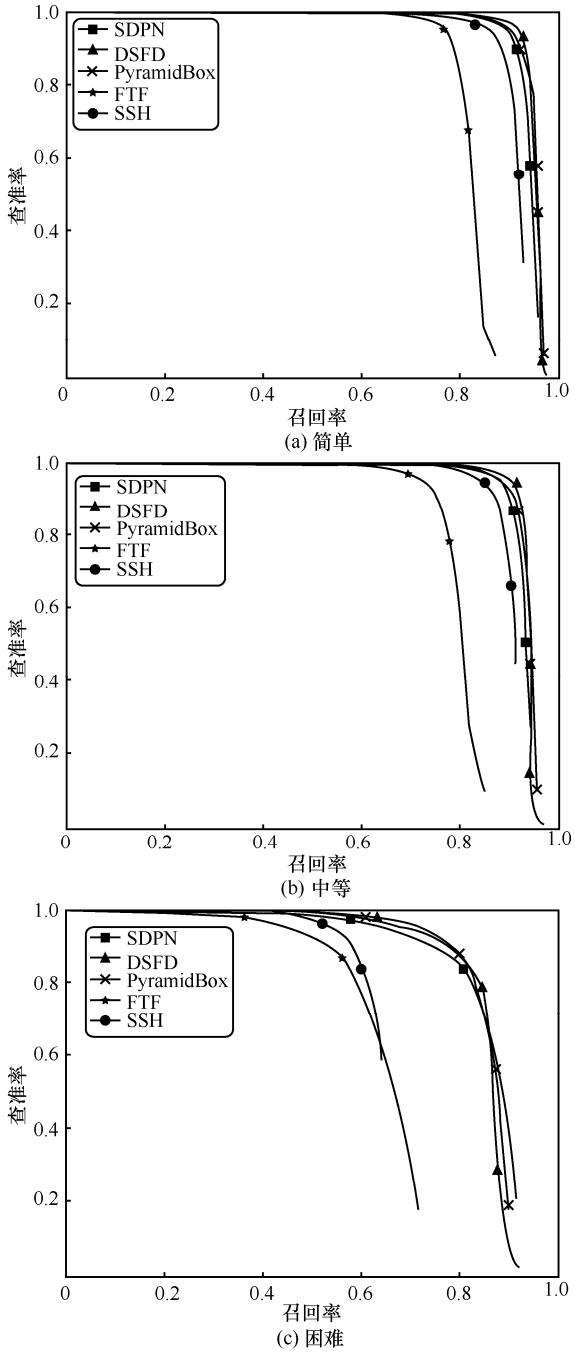


图 6　不同类别验证集上 PR 曲线

由表 2 可知，在 wider face 数据集上，使用修改后的 Face-ResNet18 网络作为主干网络，单尺度测试条件下简单子集检测精度为 0.937，中等子集检测精度为 0.927，困难子集检测精度为 0.849。相比主流算法中效果较好的 DSFD 算法，本文 SDPN 模型的测试速度为其 3.94 倍，同时在简单子集和中等子集上仅有微小差距，在困难子集上表现更好，这体现了算法的稳健性。推理速度如图 8 所示，相比主流算法中速度最快的 FTF 算法，速度依然保持 26%左右的领先，同时在简单、中等、困难子集上表现均优于 FTF 算法。在对比算法中，本文算法的模型参数量最小，为 65.06 MB，体现了算法的轻量化特点。

本文算法对应的模型收敛曲线如图 7 所示，根据收敛曲线，模型训练迭代到第 10 轮之前，学习率较高，同时由于加载了 ImageNet 预训练权重进行初始化，loss 函数收敛很快，在迭代 10 轮后，开始缓慢收敛，在 180 轮左右达到拟合。

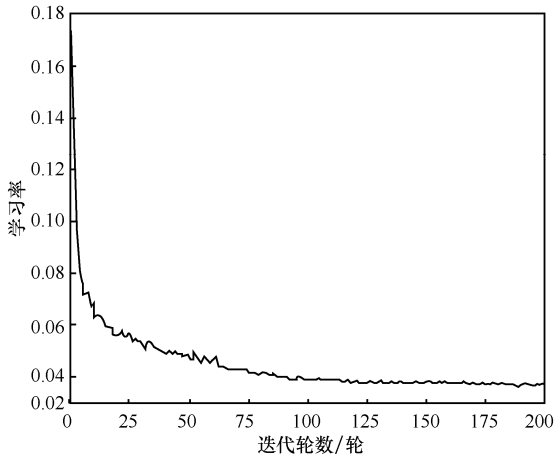


图 7　本文模型收敛曲线

与主流轻量化网络模型 LFFD[6]、Faceboxes[7] 进行对比，选择与 LFFD 实验平台 Titan xp 接近的 Nvidia-P100 单显卡进行时延与速度测试，输入

表 3　　与主流轻量化网络模型对比

模型	检测精度			时延/ms	速度/(frame· s^{-1})
	简单	中等	困难		
SDPN	0.937	0.927	0.849	55.8	17.9
LFFD[6]	0.896	0.865	0.770	31.27	31.98
FaceBoxes3.2x[7]	0.791	0.794	0.715	25.37	39.41

为 1080P 分辨率，Nvidia-P100 和 Titan xp 同样具有 3 840 个 cuda 核心，其单精度浮点性能略低于 Titan xp，测试结果如表 3 所示。SDPN 算法速度为 17.9 frame/s，LFFD 算法速度为 31.98 frame/s。结合对比 SDPN 算法与 LFFD 算法检测精度，简单子集分别为 0.937 和 0.896，中等子集分别为 0.927 和 0.865，困难子集分别为 0.849 和 0.715。可以看出，本文的边缘人脸检测网络在基本保持实时性的同时，实现了较高的人脸检测精度。

5　结束语

本文首先研究了将 ResNet 用于人脸检测中的一些潜在问题，改进了下采样卷积操作，并将 ResNet 中浅层的 7×7 卷积层和 maxpool 层替换为全残差模块，针对每一个改进进行了对比实验，充分证明了改进后的 Face-ResNet 的有效性。在此基础上，提出了兼顾正常样本与极小、遮挡、模糊等困难样本特征的双分支浅层特征提取模块，显著提升了网络的浅层表达能力，获取了更加全面、更具细粒度的特征信息。基于上述改进，深浅特征融合模块通过尺度感知的训练策略，更加有效地监督网络中的不同分支学习的差异化信息，显著提升了网络在数据集不同场景下的准确率。

当前人脸检测中的主流特征融合方法主要包括特征金字塔形式的特征融合，以及类似 DSFD 形式对网络层进行特征增强，得到双路径的特征增强融合信息。

与 DSFD 中在整个网络计算过程中设计双路径特征增强结构不同，本文的双分支仅在最关键的浅层（P_4 层之前）并行提取特征。使 2 个分支与不同的深层结构进行融合连接，针对性地监督不同的分支学到差异化的特征信息。并通过人脸尺度感知的训练策略，使小人脸检测分支专注于局部细节特征，主分支则关注全局特征。差异化的特征信息在 P_4 层融合，丰富了人脸特征信息。从信息熵的角度来看，差异化的训练方式指导不同分支学习不同特征，使网络获得了更充分的特征信息。区别于以往的特征融合方式，对网络的不同层进行差异化的监督训练，本文提出的双分支结构可以针对不同的分支进行差异化的训练，从而得到更丰富的多尺度信息。

与主流算法对比，基于 DPE 的多尺度感知的人脸检测网络 SDPN 实现了速度与精度的均衡，在保留主流网络精度的同时，充分优化了网络结构、资源占用以及推理速度，保持了轻量化、简洁、稳健的边缘计算友好特性。同时本文算法对应的模型为端到端的无锚点设计网络，简单高效，更加易于部署到实时边缘计算应用中，实现高精度的人脸检测。

参考文献：

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

[2] CHI C, ZHANG S, XING J, et al. Selective refinement network for high performance face detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2019: 8231-8238.

[3] LI J, WANG Y, WANG C, et al. DSFD: dual shot face detector[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 5060-5069.

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

[5] TANG X, DU D K, HE Z, et al. Pyramidbox: a contextassisted single shot face detector[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2018: 797-813.

[6] HE Y, XU D, WU L, et al. LFFd: a light and fast face detector for edge devices[J]. arXiv Preprint, arXiv: 1904.10633, 2019.

[7] ZHANG S, ZHU X, LEI Z, et al. Faceboxes: a CPU real-time face detector with high accuracy[C]//2017 IEEE International Joint Conference on Biometrics. Piscataway: IEEE Press, 2017: 1-9.

[8] 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. Piscataway: IEEE Press, 2016: 5525-5533.

[9] DENG J, DONG W, SOCHER R, et al. ImageNet: a largescale hierarchical image database[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern. Piscataway: IEEE Press, 2009: 248-255.

[10] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//Neural Information Processing Systems. Massachusetts: MIT Press, 2015: 91-99.

[11] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern. Piscataway: IEEE Press, 2017: 2117-2125.

[12] HUANG L, YANG Y, DENG Y, et al. Densebox: unifying landmark localization with end to end object detection[J]. arXiv Preprint, arXiv: 1509.04874, 2015.

[13] 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.

[14] JIANG H, LEARNED-MILLER E. Face detection with the faster R-CNN[C]//2017 12th IEEE International Conference on Automatic Face & Gesture Recognition. Piscataway: IEEE Press, 2017: 650-657.

[15] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2016: 21-37.

[16] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 779-788.

[17] ZHANG S, ZHU X, LEI Z, et al. S3FD: single shot scaleinvariant face detector[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 192-201.

[18] LAW H, DENG J. Cornernet: detecting objects as paired keypoints[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2018: 734-750.

[19] DUAN, KAIWEN, et al. Centernet: keypoint triplets for object detection[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 6569-6578.

[20] LIU W, LIAO S, REN W, et al. High-level semantic feature detection: a new perspective for pedestrian detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 5187-5196.

[21] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 770-778.

[22] ODENA A, DUMOULIN V, OLAH C. Deconvolution and checkerboard artifacts[J]. Distill, 2016, 1(10):e3.

[23] SINGH B, DAVIS L S. An analysis of scale invariance in object detection SNIP[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 3578-3587.

[24] LIN T, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE International Conference on Computer vision. Piscataway: IEEE Press, 2017: 2980-2988.

[作者简介]

戚琦（1982– ），女，河北廊坊人，博士，北京邮电大学副教授、博士生导师，主要研究方向为智能边缘计算、轻量级神经网络、业务网络智能化等。

马迎新（1996– ），男，山西临汾人，北京邮电大学硕士生，主要研究方向为深度学习、计算机视觉、人脸检测与识别。

王敬宇（1978– ），男，吉林长春人，博士，北京邮电大学教授、博士生导师，主要研究方向为智能网络、人工智能、计算机视觉、深度学习、多媒体通信等。

孙海峰（1989– ），男，天津人，博士，北京邮电大学讲师、硕士生导师，主要研究方向为人工智能、机器视觉、自然语言处理、深度学习等。

廖建新（1965– ），男，四川宜宾人，博士，北京邮电大学“长江学者”特聘教授、博士生导师，主要研究方向为移动通信网络、业务网络化、人工智能、多媒体业务等。