一种上下文敏感的多尺度人脸检测方法

陈龙*，庞彦伟

天津大学电气自动化与信息工程学院，天津 300072

摘要 针对非约束环境下，受姿态、遮挡、尺度变化等因素的影响，密集、分辨率较低的人脸难以检测问题，提出了一种上下文敏感的多尺度人脸检测（CSMS）方法。该方法引入一种结合人脸上下文信息的提取模块，通过有效地融合多感受野特征来丰富目标的判别性信息。从模型结构设计的角度出发，利用多尺度特征提取尺度专门化的特征向量，使人脸检测中尺度变化具有很好的稳健性。在训练阶段采用端到端的学习方式，并引入专注于难分负例样本的训练方法来解决小尺度目标检测中的类间不平衡问题，提高了网络对难例样本的判别能力。实验结果表明，该方法对于非约束环境下的人脸检测具有很好的稳健性，在 Wider Face 数据集上实现了先进的检测效果。

关键词 图像处理；人脸检测；深度卷积神经网络；上下文敏感；多尺度

中图分类号 TP391.4 **文献标识码** A **doi**：10.3788/LOP56.041003

Context-Sensitive Multi-Scale Face Detection

Chen Long*, Pang Yanwei

School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, *China*

Abstract The dense and low-resolution face is difficult to be detected under the influence of attitude, occlusion and scale change. We propose a context-sensitive multi-scale face detection (CSMS) method to solve this problem. First, the CSMS method introduces an extraction module which combines the face context information to enrich the discriminant information by effectively fusing the features of multiple receptive fields. Secondly, from the point of view of model structure design, the CSMS method uses multi-scale features to extract scale-specific feature vectors and achieve the robust scale variety in face detection. In the training phase, the CSMS method adopts the end-to-end learning method, and introduces the training method focusing on the hard negative examples to solve the class imbalance problem in the small-scale target detection, and improves the ability of the network to distinguish the difficult examples. Experimental results show that the proposed method is robust in unconstrained environments and achieves advanced detection performance on the Wider Face dataset.

Key words image processing; face detection; deep convolution neural network; context sensitivity; multi-scale

OCIS codes 100.2000; 100.5010; 150.1135

1 引 言

人脸检测在身份验证、表情分析[1]、视频监控等现代应用技术中扮演着重要的角色，准确有效地检测到图像中的人脸，对相关领域的发展有着重要的促进作用。但受现实场景中不同的人脸尺寸、光照条件、面部遮挡等因素的影响，人脸检测在非约束条件下依旧存在诸多挑战。在早期工作中，Viola 等[2]利用 AdaBoost 算法和类 Harr 特征来训练一个级联的人脸的分类器。在此工作之后，大多数的改进方案都是将原始图像通道转换为特征通道（梯度直方图通道或 LUV 颜色通道等），然后提取局部特征利用 Adaboost 等分类器进行检测[3-6]，但这些早期工作都依赖于手工标注特征并且需要单独优化系统中的每个组件。

近年来，随着卷积神经网络（CNN，Convolutional Neural Network）的突破性发展，从图像分类到目标检测都取得了很大的成功，其中双

收稿日期：2018-08-08；**修回日期**：2018-09-01；**录用日期**：2018-09-05

基金项目：国家自然科学基金（61632081）

* **E-mail**：longchen@tju.edu.cn

阶段检测方法[R-CNN(Regions with CNN features)[7]、Fast R-CNN[8]、Faster R-CNN[9]]分别利用两个子网络完成候选框的提取与分类,准确率较高但速度较慢;单阶段检测方法[YOLO(You Only Look Once)[10]、SSD(Single Shot multiBox Detector)[11]、Focal Loss[12]]利用一个网络同时进行候选框的提取与分类,速度快但准确率不如前者,因此,探索一种有效的特征提取机制来提升单阶段网络的检测性能显得尤为重要。得益于强大的深度卷积网络以及端到端的优化方式,基于 CNN 的人脸检测器实现了更好的检测效果。Yang 等[13]以 Faceness 为面部属性训练了一系列 CNN,用以检测部分遮挡的人脸。Zhang 等[14]提出使用多任务 CNN 同时解决人脸检测与人脸校准。Sun 等[15]、Zhu 等[16]以及 Wan 等[17]分别利用 Faster R-CNN 及其与难例挖掘相结合的方式,在人脸检测性能上实现了进一步的提升。

相比于其他的方法,基于候选框的检测方法在复杂的场景中更加稳健且其检测所需的时间与目标数量无关。但根据 Huang 等[18]的阐述,随着物体尺度的减小,基于候选框的检测器的性能会随之下降。细节纹理信息能够反映目标视觉性质及空间拓扑关系[19-21],Hu 等[22]对人脸检测中的上下文信息进行了探索,使用类似于区域建议网络(RPN, Region Proposal Network)[9]来直接检测人脸,并引入图像金字塔来表达网络结构的多尺度特征,最后针对不同的人脸尺度训练不同的分类器。SSH (Single Stage Headless)[23]和 Face-MagNet[24]方法均使用单阶段架构来实现人脸检测,针对人脸的多尺度问题设计候选框提取与分类策略,但未针对尺寸较小的难例样本检测提出行之有效的结构与方法。

针对上述问题以及在人脸检测任务中存在的挑战,本文提出一种上下文敏感的多尺度人脸检测(CSMS)模型。该方法运用多感受野、多语义层级融合的上下文提取模型,利用跳连接结构与空洞卷积来增强层级之间的信息传递与全局性的特征表达。同时采用单阶段的检测方法,通过权重共享结合不同语义层级的多尺度特征,实现对不同尺度人脸的针对性检测,能够在保证检测效果的同时减少了权重的参数量。在训练阶段,同时利用 Focal Loss[11]和不同尺度间特定比例的正负样例来解决小目标的类别不平衡问题,提高了检测网络对小尺度人脸的判别能力。实验结果证明,该方法能够对不同场景下的人脸进行有效的检测。

2 CSMS 网络

CSMS 方法采用基于候选框的单阶段网络模型结构,能够对整个网络从特征提取、候选框选择,到候选框的分类与回归实现端到端的训练。相比于两阶段的目标检测器,本研究方法降低了网络训练的复杂度,减少了深度卷积网络的推理时间,降低了网络的内存消耗。CSMS 方法将不同语义层级的特征进行有效地融合,对不同尺度的人脸分别进行针对性检测,实现网络模型的尺度不变性;设计了上下文敏感模块来多元化语义特征的感受野,有效地丰富目标的背景特征,适用于对较小人脸的检测。

2.1 尺度不变性网络设计

在非约束的环境下,图像中的人脸尺度有着很大的变化范围。Hu 等[22]提出在检测网络的推理之前,人为地将不同尺度的图像作为输入,并分别对其进行前向推理运算,最后将不同尺度图像中人脸的检测结果进行合并,这种方法能够实现网络的多尺度性能,但其带来的计算量与网络参数量较大,推理时间较长。本研究根据人脸目标高度在图像中所占的像素数目将人脸划分为小、中、大 3 个尺寸等级[25]。受 SSD[11]启发,本研究提出:从网络结构设计的角度实现尺度不变性,利用高语义层级、分辨率低的特征用于检测较大尺度的人脸,利用低语义层级、分辨率大的特征用于检测较小尺度的人脸,并采用不同的检测模块 D_s、D_m、D_l 从主干网络中不同深度的语义层级来对特定尺度范围内的人脸进行训练与检测。这些检测模块有着不同的分辨率步长,分别用于检测小、中、大尺寸的人脸。

本研究的整体网络结构图如图 1 所示,采用通用基础网络[例如 VGG(Visual Geometry Group)-16[26]]用于网络中前期的特征提取任务。1)对于大尺度的人脸检测,将 Conv5 层经过步长为 2 的最大值池化(pooling)后再经过上下文敏感模块(Context-sensitive Module),以提取候选目标的多感受野特征,再利用检测模型 D_l 进行检测,其分辨率为原始图像的 1/32 且语义级别较高,有利于较大人脸的分类与定位。对于中等尺度的人脸,本研究直接利用分辨率步长为 16 的 Conv5 层输出的特征向量经过上下文敏感模块和检测模型 D_m 对人脸进行检测,相比于用来检测大尺度人脸的特征,该特征具有更高的分辨率(1/16),以及更加丰富的细节信息。

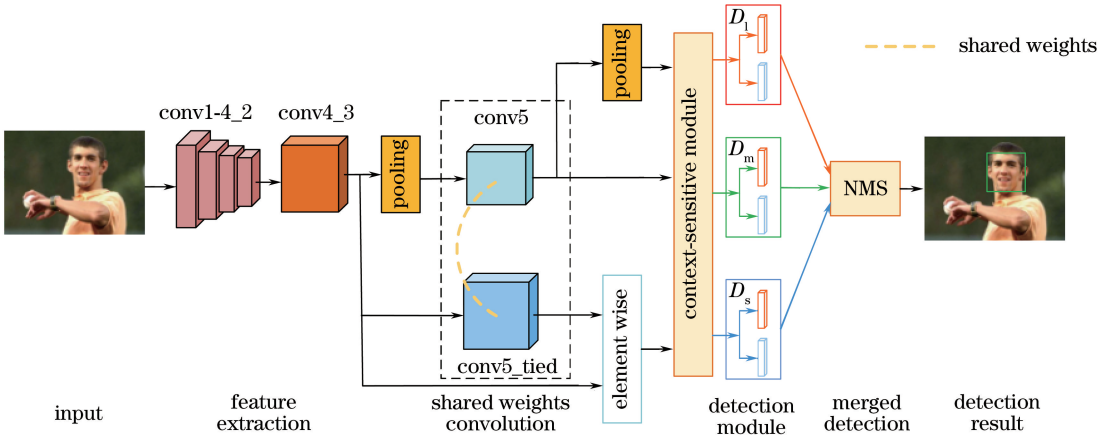

图 1 CSMS 网络结构

Fig. 1 Structure of CSMS network

针对于小尺度人脸的检测，本研究将 Conv4_3 层的输出与 Conv5_shared 层的输出进行融合，一方面保持了特征图的分辨率，另一方面融合了多语义级别和多感受野的特征。本研究中采用权重共享[27]的方法(Shared weights convolution)，直接对 Conv4_3 的特征向量进行 Conv5_shared 层的卷积操作，如图 1 中所示，相比于 FPN(Feature Pyramid Networks)[28]中 Decoder 的形式，在保持分辨率并增加感受野的同时可以获取更多的细节信息，提高检测模块 D_s 对较小人脸的检测性能。

在网络模型的推理阶段，不同检测模型的所有预测结果都经过非极大值抑制(NMS，Non-Maximum Suppression)的操作去除掉重叠较多的目标候选框，并生成最后的检测结果。

2.2 上下文敏感模块

在小目标的检测任务中，由于在目标上能探索的信息较少，该任务从根本上具有挑战性。直观地来看，人类视觉中目标周围的背景信息有助于正确地分类尺寸较小的人脸。因此，小目标的检测任务也需要利用超出目标范围的图像信息来辅助检测，这也经常被称为上下文信息。在原始的基于候选框的检测器中，增加上下文信息通常的方法是通过在候选框周围扩大滑动窗口来增大感受野，以获取上下文信息。为减小引入较大卷积核带来的参数增长，SSH 通过堆叠相同的小卷积核来等效较大的卷积核，再利用不同步长的卷积预测模型实现感受野的增加。而本研究利用跳连接的结构，设计了一种参数量更少、感受野尺度更多的上下文敏感模块，相比于直接扩大卷积层中的滑动窗口与简单地堆叠小的卷积核[见图 2(a)]，该方法能够在保持大感受野范围的同时有效地节约参数量，同时也继承了跳连接结构的优势，加强不同语义层级之间的特征传递。

上下文敏感模块：受 DenseNet[29]启发，本研究设计了一个上下文敏感模块，其具体结构如图 2(b)中所示。在上下文提取模块中，借鉴了 DenseNet 的稠密连接的思想，同时简化其模块结构，去除了 BatchNorm、Scale 等操作，直接采用跨层之间的跳连接来传递卷积层之间的信息，其结构设计如图 2(b)中所示。这样的模块设计：1) 通过跨连接的方式实现深层与浅层之间的最短路路径连接，减轻了深度卷积网络中的梯度消失问题；2) 加强了不同卷积层之间的特征传递，最大限度地利用了不同语义层级的特征，同时增强了网络的特征表达能力。此外，本研究方法在卷积核参数量和等效卷积核数量

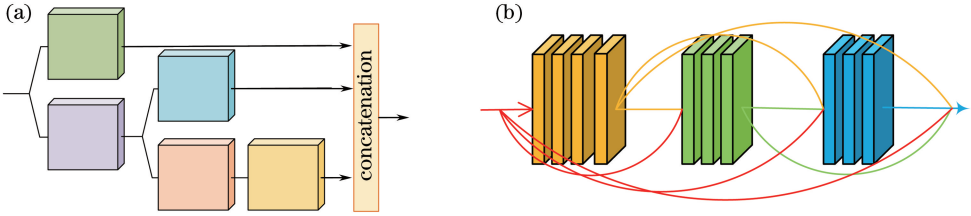

图 2 上下文提取模块。(a)堆叠模型；(b)上下文敏感模块

Fig. 2 Context extraction module. (a)Stacked model; (b) context-sensitive module

的对比结果如表 1 中所示(在不考虑通道数的情况下,其中 $n*k_s\times k_s$ 表示 n 个卷积核为 $k_s\times k_s$ 的参数量)。从表 1 中能够看出,本研究方法既实现了利用多感受野来丰富目标上下文信息的目的,又在一定程度上减少了特征提取模块的参数,同时有效地丰富了不同的卷积层特征,增强了该模块的特征表达能力。因此,本研究在检测网络中采用图 2(b)中所示的结构。

表 1 上下文提取模块对比

Table 1 Comparison of context extraction module

Model	Stacked model	CSMS context-sensitive module
Number of parameters	$5*3\times3$	$3*3\times3$
Equivalent layer	3×3, 5×5, 7×7	$3*3\times3$, $3*5\times5$, 7×7

全局性的特征表达:现代图像分类网络通过连续的池化层和下采样层来整合多尺度的上下文信息和全局性特征,减小特征图的分辨率,进而得到对全局预测的输出[26],但随之带来细节信息的损失。若不使用池化层来降低分辨率,在卷积核较小的情况下,感受野也会很小,使得能够提取的有效信息也较少。因此,在文献[30]和[31]中提出利用空洞卷积(膨胀卷积、带孔卷积、Dilated Convolution)来实现在不增加参数量与计算量的同时增大感受野,获取更多的上下文信息。空洞卷积算子在过去被称为“带有膨胀滤波器的卷积”,它在一种小波分解算法中扮演着重要的角色[32]。Yu 等[30]将 $*_l$ 膨胀的卷积或 l-dilated 卷积定义为

$$(F*_l k_s)(\boldsymbol{q})=\sum_{\boldsymbol{s}+l\boldsymbol{t}=\boldsymbol{q}}F(\boldsymbol{s})k_s(\boldsymbol{t}),\tag{1}$$

式中:F 为离散函数且 $F:\mathbf{Z}^2\to\mathbf{R}$,$\Omega_r=[-r,r]^2$,$k_s:\Omega_r\to\mathbf{R}$ 为离散滤波器大小 $(2r+1)^2$,$\boldsymbol{q}$、$\boldsymbol{s}$、$\boldsymbol{t}$ 分别代表输出特征、输入特征、卷积核的位置向量;对于 n-dilated 空洞卷积的卷积核大小 k_d 定义为

$$k_d=2*n+1。\tag{2}$$

相比于常规的大卷积核滤波器,空洞卷积能够在不增加额外参数量或计算量的情况下有效地扩大滤波器的感受野,更好地获得全局信息,有助于提高对全局预测的输出。

针对增强候选目标的全局性信息,本研究在主干网络中利用空洞卷积来提升不同检测分支的全局性特征表达。为不增加网络参数,直接将 Conv5 层中的线性卷积替换为空洞卷积,以此在不降低特征图分辨率的前提下增大网络的感受野。对于提升小尺度目标的检测性能,在现有方法中的一般做法为:1) 直接提高输入图像的分辨率,人为地丰富小目标所包含的特征,提高了检测性能,但也增加了预测网络的计算量;2) 先提升特征向量的语义级别,其间利用池化操作来增大特征点的感受野,然后对高语义的特征图进行上采样来提高其分辨率,或再将其与浅层的特征进行融合。这种方法虽然通过上采样恢复了特征图的大小,但终究会丢失掉很多无法恢复的细节信息。在本研究中,针对于较小尺寸人脸的检测分支,利用权重共享机制,在保持 Conv4_3 层的输出特征分辨率的情况下,利用 Conv5 层的权重进行特征的提取,将 Conv5 层与 Conv4 层输出的特征图进行融合,其结构设计如图 3 所示。这样既能利用空洞卷积对目标进行大感受野的特征提取,又能尽可能地保留目标特征中的细节信息,同时将 Conv4_3 的细粒度特征与共享权重层 Conv5_shared 的全局性特征相融合,又能够缓解由空洞卷积带来的栅格效应问题,增强候选目标的判别性信息,实现较小尺寸的人脸的有效检测。

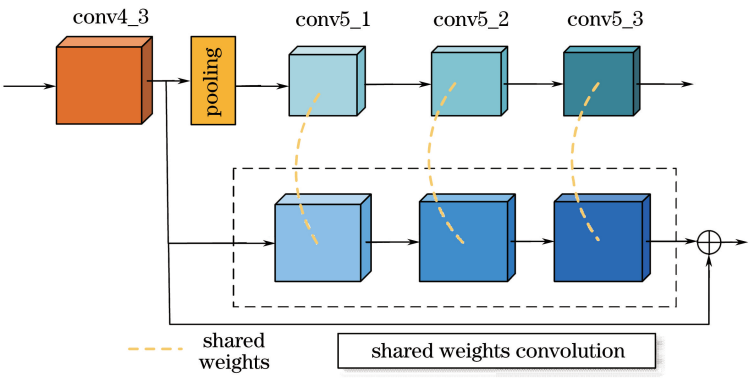

图 3 有效的特征融合结构

Fig. 3 Effective feature fusion structure

2.3 针对小尺度样本的类间不平衡问题

高准确率的物体检测器大多是基于类似 Faster R-CNN 的两阶段检测器，先利用 RPN 生成一系列稀疏的候选框位置来提取潜在的目标位置，再利用卷积神经网络对每个候选位置进行前景与背景的分类。该类方法通过精炼候选框的方式来提高检测准确率，但其检测速度并不能得到实质的提升。相比之下，单阶段的检测器不需要 RPN 提取候选框，直接对网格候选框进行特征提取、分类与回归。这类方法速度较快，但准确率却不如前者，而导致这个问题的原因是在稠密目标检测器的训练阶段存在极端的前景-背景类别不平衡。据此，Focal Loss 重塑标准交叉熵损失，通过降低易分类样本的权重，使模型更加专注于难例样本以及避免大量易分类样例主导模型的优化方向[12]。Focal Loss(FL)定义为

$$L_{\mathrm{F}}(p_{\mathrm{t}})=-\alpha_{\mathrm{t}}(1-p_{\mathrm{t}})^{\gamma}\mathrm{lb}\,p_{\mathrm{t}}, \tag{3}$$

$$p_{\mathrm{t}}=\begin{cases}p, & \text{if } y=1\\ 1-p, & \text{otherwise}\end{cases}, \tag{4}$$

式中：$y\in\{-1,+1\}$ 为类别标签，p 为模型对标签 $y=1$ 计算出的类概率；$p_{\mathrm{t}}\in[0,1]$，其中 $(1-p_{\mathrm{t}})^{\gamma}$ 为调节因子，γ 为可调的聚焦参数，且 $\gamma\geqslant 0$，其作为超参数通过多次的对比实验来进行设置；α_{t} 为加权因子，$\alpha_{\mathrm{t}}\in[0,1]$，其通过类似于定义 p_{t} 的方法来进行定义：当类别为 1 时取 α，当类别为 -1 时取 $1-\alpha$，α 通过与类频率成反比的方法或者通过交叉验证作为超参数来进行设置。

针对小目标的检测特性，想要达到好的检测效果需要在较大分辨率的特征图中进行，又由于目标本身尺寸较小，网络所提取的候选框尺寸也小，导致此部分的检测过程中会出现大量建议候选框，而在其中绝大多部分都是负例样本，故导致正例样本所占比例与之相比较为悬殊，而这也严重影响了正负样本所主导的网络优化方向。在本研究中应用 Focal Loss 来训练检测网络，利用样本自身的难易分程度来控制优化过程中正负样例的损失权重，专注于难分样本对于网络模型的优化，且通过实验对比得到：$\alpha_{\mathrm{t}}=0.5$、$\gamma=1$，具体表达式为

$$L_{\mathrm{F}}(p_{\mathrm{t}})=-0.5(1-p_{\mathrm{t}})\mathrm{lb}\,p_{\mathrm{t}}。 \tag{5}$$

同时对于每个预测模型，本研究针对于不同尺寸的人脸检测分支选出不同比例的正负例样本作为每个迭代中的 mini-batch，对于尺寸最小的人脸目标选择的正负样本比例为 0.5，对于尺寸中等以及较大的人脸目标则选择的正负样本比例为 0.25。

2.4 检测网络的损失函数

本研究中采用多任务损失联合来实现对人脸目标的分类与定位，进而对网络进行优化。网络模型的总损失函数 L 定义为

$$L=\sum_{k}\frac{1}{N_{k}^{(\mathrm{c})}}\sum_{i\in A_k}l_{\mathrm{cls}}(p_i,y_i)+\lambda\sum_{k}\frac{1}{N_{k}^{(\mathrm{r})}}\sum_{i\in A_k}\Gamma(y_i=1)l_{\mathrm{reg}}(\boldsymbol{b}_i,\boldsymbol{t}_i), \tag{6}$$

式中：i 代表候选框的索引，k 指代网络中检测模块 $D=\{D_{\mathrm{s}},D_{\mathrm{m}},D_{l}\}$ 的索引，l_{cls} 和 l_{reg} 分别代表候选框的分类损失和回归损失，其中 l_{cls} 利用 Focal Loss[12] 来计算，l_{reg} 采用 smooth L1 loss[8] 进行计算。在计算分类损失中，A_k 代表在检测模块 D 中用于分类与回归的候选框集合，p_i 和 y_i 分别指代在检测模块 D 中的第 i 个候选框的类别预测概率及其所分配的 ground-truth 标签，其中当且仅当候选框与 ground-truth 的 IoU(Intersection over Union)大于预设阈值(如 0.5)时才会被分配到相对应的正例标签，反之，若 IoU 小于某个预设阈值(如 0.3)时则被分配为负例标签。对于人脸检测的分类标签只有两类(人脸和背景)：正例样本标签为 1，负例样本标签为 0。$N_{k}^{(\mathrm{c})}$ 是参与分类损失计算的模块 D_k 中候选框的数目，此处作为归一化参数。在计算回归损失中，$N_{k}^{(\mathrm{r})}$ 是 A_k 中 $y_i=1$ 的候选框数量，λ 为作为平衡参数，用于平衡分类损失与回归损失归一化的不平衡。$\boldsymbol{b}_i$ 和 $\boldsymbol{t}_i$ 分别是第 i 个预测候选框的 4 维坐标向量以及匹配的 ground-truth 回归目标。指示函数 $\Gamma(y_i=1)$ 的意思是仅对正例激活回归损失。

3 实验与分析

3.1 实验配置

网络参数初始化：本研究中利用 caffe[33] 框架在 ImageNet 分类网络上进行预训练的 VGG16 网络架构对模型参数进行初始化，但未使用 Conv_fc6 与 Conv_fc7 层的权重参数，对于其他增加的网络层均采用“xavier”的方法[34]进行随机初始化。

训练参数设置：实验在 GPU NVIDIA TITAN X 的图形处理器(GPU)上对检测网络进行训练。本研究对动量和权重衰减采用随机梯度下降算法来对检测网络进行优化，在检测模块中，建议候选框与 ground-truth 的 IoU 小于 0.5 的被标记为正例，IoU 小于 0.3 的则被作为背景标记为负例。在训练阶段中，学习率设置为 0.01，momentum 设置为 0.9，迭

代次数设置为 30000，batchsize 设置为 64；在推理阶段中，NMS 的阈值设置为 0.3。

测试模型：在测试阶段中，CSMS 模型为检测网络在输入不加入图像金字塔时的测试模型，将图像的输入尺寸在不改变原始长宽比的情况下缩放至短边为 1200 pixel，同时长边不超过 1600 pixel。CSMS+Pyramid 模型为检测网络在输入加入图像金字塔时的测试模型，参考 HR(Hybrid-Resolution model)[20]的图像金字塔构成，其中共有 4 个尺度等级：500，800，1200，1600 pixel。

3.2 实验数据

本研究的检测网络在 Wider Face 数据集[25]进行训练与测试，该数据集总计 32203 张图像，共包含 393703 张标注的人脸目标，其中训练集中包含 158989 张(40%)，验证集中包含 39496 张(10%)，其余的则是测试集(50%)。此外，在验证集与测试集中总计包含 60 个场景类别，根据其包含的人脸姿态、尺寸、遮挡等将图像划分为 3 个等级：Easy、Medium 和 Hard，其中 Easy 子集的人脸尺寸最大最容易检测，Hard 的尺寸最小也最难以检测。本研究仅在 Wider Face 的训练集上对检测网络进行训练，在验证集上对训练出的网络模型进行测试与切片实验分析，验证其对不同尺寸、姿态及遮挡程度的人脸检测性能。

3.3 网络模型的切片分析

尺度不变性设计：在 3.1 节中介绍了本研究提出的人脸检测网络，从不同步长的网络层级中利用特有检测模型 $D=\{D_s, D_m, D_l\}$，来实现分别对特定尺度的人脸进行检测。为了更好地分析本研究中尺度不变性设计的有效性，表 2 比较了在该检测网络结构下不同的尺度不变性设计对最终检测效果的影响并给出不同情况下的平均精确度均值(mAP)。首先仅保留从 Conv5 层对特征图进行检测的中间尺度的检测模块 $D=\{D_m\}$，利用单一分支对所有尺度的人脸进行检测，模型记为 CSMS-sglD_m。结果表明，本研究网络中在检测网络不同语义层级的、针对性的检测模块通过专门化的训练与推理，对于实现尺度不变性是一个非常有效的策略。

表 2 尺度不变性设计的 mAP 对比

Table 2 Comparison of mAP for scale invariance design %

Level	CSMS-sglD_m	CSMS
Easy	81.2	**92.6**
Medium	83.6	**91.0**
Hard	39.8	**82.5**

上下文特征提取：在 3.2 节中详细描述了用于提取上下文信息的不同方法，CSMS 利用稠密的跳连接结构，一方面利用较少的参数实现了大感受野的特征提取，增强特征的全局信息；另一方面缩短深层与浅层间的连接路径，同时增强了不同层之间的特征传递与网络的特征表达能力。表 3 中的实验结果显示，与堆叠模型相比，CSMS 对上下文信息的表达能力更强，这也表明了更大的感受野与稠密的跳连接结构对正确分类与定位的重要性。

表 3 上下文敏感模块 mAP 对比

Table 3 Comparison of mAP for context-sensitive module %

Level	Stacked model	CSMS
Easy	92.5	**92.6**
Medium	90.8	**91.0**
Hard	82.0	**82.5**

专注于难例样本的训练：如 3.3 节中所描述的，CSMS 在训练阶段采用 Focal Loss 作为分类阶段的损失计算函数，通过候选框的得分来实现对损失权重的动态调整，使得网络对难分的样本更加敏感，同时弱化易分样本对网络优化方向的影响。表 4 中显示：在其他因素不变的情况下，CSMS 的性能与采用 OHEM(Online hard example mining)[35]的检测结果的比较。

表 4 中的实验结果表明在网络的训练阶段利用 Focal Loss 来针对难分样本的优化对提升网络的检测性能具有更好的效果。

表 4 专注于难例样本训练方的 mAP 比较

Table 4 Comparison of mAP for training methods focused on hard sample samples %

Level	OHEM	Focal Loss
Easy	91.9	**92.3**
Medium	90.7	**90.8**
Hard	81.4	**81.6**

有效的特征融合：为实现对较小尺寸人脸更好的检测效果，在 CSMS 中将线性卷积层 Conv4_3 与共享权重层 Conv5_shared 的细粒度信息与全局性信息相融合，其中并未通过 pooling 来增大感受野，而是在 CSMS 中增加权重共享分支，直接对 Conv_4 的特征进行卷积操作，尽可能保留原始的细节信息。设计两组对比实验：1) 采用一般方法先利用 pooling 增大感受野，再将其上采样来恢复分辨率，模型记为 CSMS-pl；2) 在 1)中的 Conv5 层加入空洞卷积，模型记为 CSMS-plDlt，见表 5。

表 5　特征融合的 mAP 比较

Table 5　Comparison of mAP for feature fusion　%

Level	CSMS-pl	CSMS-plDlt	CSMS
Easy	92.3	92.6	**92.6**
Medium	90.8	90.8	**91.0**
Hard	82.0	82.2	**82.5**

相比于增强小目标检测的一般方法，CSMS 在 Hard 上的检测性能提高了 0.5%，实验结果表明，利用空洞卷积来增大感受野与利用权重共享层来实现深层特征与浅层特征的融合，可以有效提高人脸检测性能。

表 6 给出 CSMS 在 Wider Face 验证集中的切片分析，详细比较了不同结构下 CSMS 的检测性能（其中√表示引入该结构，否则不引入）。从表 6 中能够看出，CSMS 相对于 baseline[23] 在 Easy、Medium、Hard 子集中 mAP 分别提高了 0.7%、0.3%、1.1%，实验结果证明了 CSMS 的有效性，尤其对于较小尺寸人脸的检测。

表 6　CSMS 在 Wider Face 测试集中的切片分析

Table 6　Ablation study of CSMS on Wider Face test set

Method	Contribution			mAP / %		
	Focal Loss	Context-sensitive module	Effective feature fusion	Easy	Medium	Hard
Baseline				91.9	90.7	81.4
CSMS	√	√	√	92.3	90.8	81.6
		√	√	92.4	90.8	82.2
			√	92.6	91.0	82.5

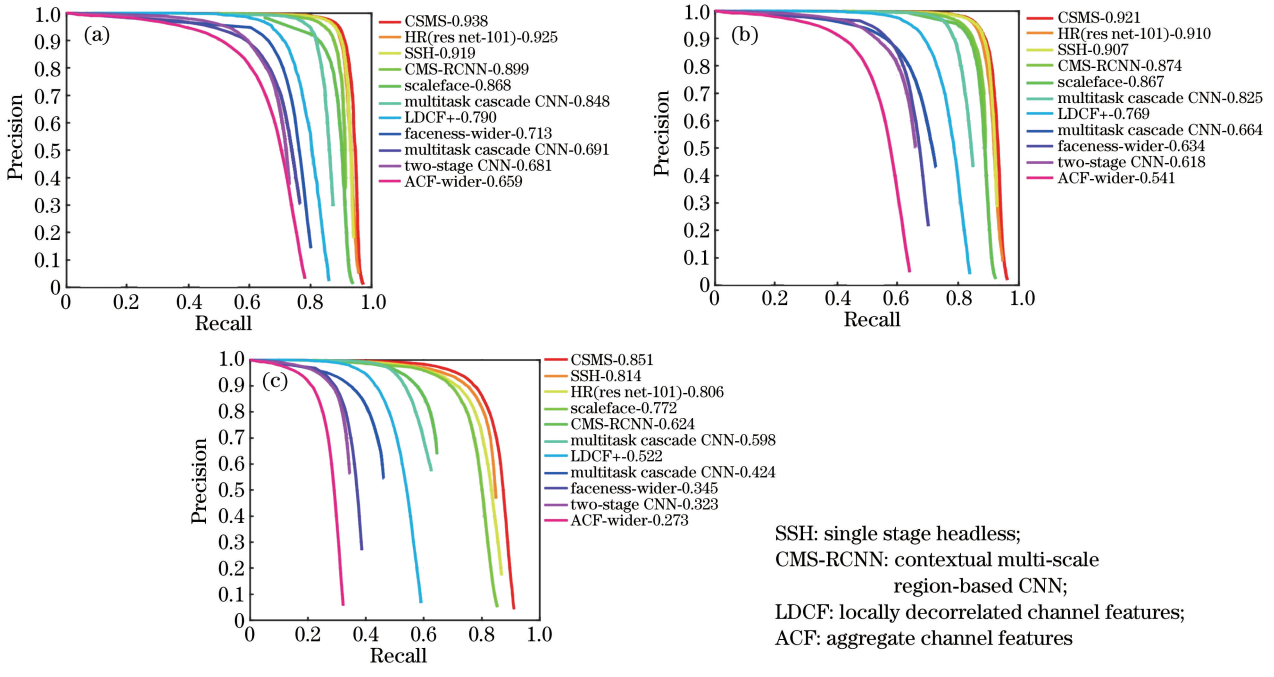

图 4　Wider Face 验证集上准确率-召回率曲线。(a) Easy；(b) Medium；(c) Hard

Fig. 4　Precision-recall curves on Wider Face validation set. (a) Easy; (b) Medium; (c) Hard

3.4　实验结果对比

本研究在 Wider Face 数据集的验证集上对检测网络模型的检测性能进行测试，并与 HR（ResNet-101）＋ Pyramid[20]、CMS-RCNN（Contextual multi-scale region-based CNN）[16]、Multitask Cascade CNN[14]、ScaleFace[36]、LDCF（Locally Decorrelated Channel Features）[37]、Faceness[14] 和 Multiscale Cascade CNN[25] 相比较。图 4 中展示了测试模型 CSMS 在加入图像金字塔时 Wider Face 验证集上的 Easy、Medium、Hard 3 个部分的准确率-召回率曲线，以及与其他先进检测网络的性能对比，结果表明 CSMS 在保持对大尺寸人脸较高检测性能的同时，很好地提升了对小尺寸人脸的检测性能，从而证明了 CSMS 检测网络结构的有效性。

图 5 中展示了 CSMS 在 Wider Face 验证集中部分定性的检测结果。能够从图中看出无论是对空间分布较为密集的小尺度人脸，还是对于受到姿态、

遮挡、表情等影响的人脸，CSMS 都能够准确地将其检测出来，并给出人脸在图像中的位置。

图 5 定性的检测结果

Fig. 5 Qualitative detection results

表 7 中给出了不同的检测模型在不同配置下的检测性能，从表中能够看出 CSMS 在使用 VGG-16 作为基础网络时的检测性能与 HR＋Pyramid 在使用 ResNet-101 时的检测性能基本持平，但在 Hard 部分提高了 1.9%的准确率。CSMS 与 SSH 相比，不论是否加入图像金字塔结构，在 Easy、Medium、Hard 部分均提高了检测准确率，体现了 CSMS 对人脸检测的尺度不变性，证明了 CSMS 采用的上下文敏感模块与专注于难例的训练方法对于在非约束环境下人脸检测的有效性，很好地提高了检测网络的稳健性。

表 7 检测性能的对比

Table 7 Comparison of detection performance

Method	mAP /%		
	Easy	Medium	Hard
Method in Ref. [15]	86.2	84.4	74.9
Method in Ref. [15]	92.5	91.0	80.6
Method in Ref. [14]	89.9	87.4	62.9
Method in Ref. [17]	91.9	90.7	81.4
Method in Ref. [17]	93.1	92.1	84.5
CSMS (VGG-16)	92.6	91.0	82.5
CSMS(VGG-16)＋Pyramid	93.8	92.1	85.1

4 结 论

本研究提出了一种上下文敏感的多尺度人脸检测方法，简称为 CSMS，着重解决非约束环境中人脸检测中存在的难点，利用一种多尺度的网络结构实现了单阶段的人脸尺度不变性检测；结合上下文提取模块与空洞卷积，增强了网络对人脸上下文信息的学习能力。此外，CSMS 运用权重共享机制，将高语义高分辨率特征与低语义特征进行融合，同时利用 Focal Loss 与尺度专门化的正负样例比使 CSMS 专注于对难分负例样本的优化，有效地提高了对小尺度样本的判别能力。实验结果表明 CSMS 在公共的人脸检测基准上实现了先进的检测性能。

参 考 文 献

[1] Wang L L, Liu J H, Fu X M. Facial expression recognition based on fusion of local features and deep belief network [J]. Laser & Optoelectronics Progress, 2018, 55(1): 011002.
王琳琳, 刘敬浩, 付晓梅. 融合局部特征与深度置信网络的人脸表情识别[J]. 激光与光电子学进展, 2018, 55(1): 011002.

[2] Viola P, Jones M J. Robust real-time face detection [J]. International Journal of Computer Vision, 2004, 57(2): 137-154.

[3] Cao J L, Pang Y W, Li X L. Pedestrian detection inspired by appearance constancy and shape symmetry [J]. IEEE Transactions on Image Processing, 2016, 25(12): 5538-5551.

[4] Dollar P, Appel R, Belongie S, *et al*. Fast feature pyramids for object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(8): 1532-1545.

[5] Zhang S S, Benenson R, Schiele B. Filtered channel features for pedestrian detection [C] // 2015 IEEE Conference on Computer Vision and Pattern Recognition, June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015:1751-1760.

[6] Kong Y P, Liu X, Xie X Q, *et al*. Face liveness detection method based on histogram of oriented gradient [J]. Laser & Optoelectronics Progress, 2018, 55(3): 031009.
孔月萍, 刘霞, 谢心谦, 等. 基于梯度方向直方图的人脸活体检测方法[J]. 激光与光电子学进展, 2018, 55(3): 031009.

[7] Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation [C] // 2014 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York:

IEEE, 2014:580-587.

[8] Girshick R. Fast R-CNN[C]// 2015 IEEE Conference on Computer Vision, December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 1440-1448.

[9] Ren S Q, He K M, Girshick R, *et al*. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[10] Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, real-time object detection[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas NV, USA. New York: IEEE, 2016: 779-788.

[11] Liu W, Anguelov D, Erhan D, *et al*. SSD: Single shot multiBox detector[M]// Leibe B, Matas J, Sebe N, *et al*. eds. Computer Vision-ECCV 2016. Cham: Springer, 2016: 21-37.

[12] Lin T Y, Goyal P, Girshick R, *et al*. Focal loss for dense object detection[C]// 2017 IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. New York: IEEE, 2017:2999-3007.

[13] Yang S, Luo P, Loy C C, *et al*. From facial parts responses to face detection: A deep learning approach [C]// 2015 IEEE International Conference on Computer Vision, December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 3676-3684.

[14] Zhang K P, Zhang Z P, Li Z F, *et al*. Joint face detection and alignment using multitask cascaded convolutional networks[J]. IEEE Signal Processing Letters, 2016, 23(10): 1499-1503.

[15] Sun X D, Wu P C, Hoi S C H. Face detection using deep learning: An improved faster RCNN approach [J]. Neurocomputing, 2018, 299: 42-50.

[16] Zhu C, Zheng Y, Luu K, *et al*. CMS-RCNN: Contextual multi-scale region-based CNN for unconstrained face detection[M]// Bhanu B, Kumar A. eds. Deep Learning for Biometrics. Advances in Computer Vision and Pattern Recognition. Cham: Springe, 2017: 57-79.

[17] Wan S, Chen Z, Zhang T, *et al*. Bootstrapping face detection with hard negative examples [EB/OL]. (2016-08-07) [2018-08-07] http://arxiv.org/abs/1608.02236.

[18] Huang J, Rathod V, Sun C, *et al*. Speed/accuracy trade-offs for modern convolutional object detectors [C]// 2017 IEEE International Conference on Computer Vision and Pattern Recognition, July, 21-26,2017, Honolulu, HI, USA. New York: IEEE, 2017: 3296-3297.

[19] Deng X Q, Zhu Q B, Huang M. Variety discrimination for single rice seed by integrating spectral, texture and morphological features based on hyperspectral image [J]. Laser & Optoelectronics Progress, 2015, 52(2): 021001.
邓小琴, 朱启兵, 黄敏. 融合光谱、纹理及形态特征的水稻种子品种高光谱图像单粒鉴别[J]. 激光与光电子学进展, 2015, 52(2): 021001.

[20] Hou Z Q, Wang L P, Guo J X, *et al*. An object tracking algorithm based on color, space and texture information[J]. Opto-Electronic Engineering, 2018, 45(5): 39-46.
侯志强, 王利平, 郭建新等. 基于颜色、空间和纹理信息的目标跟踪[J]. 光电工程, 2018, 45(5): 39-46.

[21] Sun Y J, Dong J N, Wang Z F. Estimation of lighting parameters for uniform texture image [J]. Laser & Optoelectronics Progress, 2017, 54(6): 061002.
孙玉娟, 董军宇, 王增锋. 灰度一致纹理图像的光参数估算方法[J]. 激光与光电子学进展, 2017, 54(6): 061002.

[22] Hu P Y, Ramanan D. Finding tiny faces[C]// 2017 IEEE International Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 1522-1530.

[23] Najibi M, Samangouei P, Chellappa R, *et al*. SSH: Single stage headless face detector [C]// IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. New York: IEEE, 2017: 4885-4894.

[24] Samangouei P, Najibi M, Davis L, *et al*. Face-MagNet: Magnifying feature maps to detect small faces [C]// 2018 IEEE Winter Conference on Applications of Computer Vision, March 12-15, 2018, Lake Tahoe, NV, USA. New York: IEEE, 2018: 122-130.

[25] 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, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 5525-5533.

[26] Simonyan K, Zisserman A. Very deep convolutional

networks for large-scale image recognition[EB/OL]. (2014-09-04)[2018-08-07]. https://arxiv.org/abs/1409.1556.

[27] Cao J L, Pang Y W, Li X L. Exploring multi-branch and high-level semantic networks for improving pedestrian detection[EB/OL]. (2018-04-03)[2018-08-07]. http://arxiv.org/abs/1804.00872.

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

[29] Huang G, Liu Z, van der Maaten L, *et al*. Densely connected convolutional networks[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 2261-2269.

[30] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions[EB/OL]. (2015-11-23)[2018-08-07]. http://arxiv.org/abs/1511.07122.

[31] Chen L C, Papandreou G, Kokkinos I, *et al*. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.

[32] Shensa M J. The discrete wavelet transform: Wedding the atrous and Mallat algorithms[J]. IEEE Transactions on Signal Processing, 1992, 40(10): 2464-2482.

[33] Jia Y, Shelhamer E, Donahue J, *et al*. Caffe: Convolutional architecture for fast feature embedding[C]// Proceedings of the 22nd ACM international conference on Multimedia, November 4-7 2014, Dallas, Texas, USA. New York: ACM, 2014: 675-678.

[34] Glorot X, Bengio Y. Understanding the difficulty of training deep feed forward neural networks[J]. Journal of Machine Learning Research, 2010, 9: 249-256.

[35] Shrivastava A, Gupta A, Girshick R. Training region-based object detectors with online hard example mining[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 761-769.

[36] Yang S, Xiong Y, Loy C C, *et al*. Face detection through scale-friendly deep convolutional networks[EB/OL]. (2017-06-09)[2018-08-07]. http://arxiv.org/abs/1706.02863.

[37] Ohn-Bar E, Trivedi M M. To boost or not to boost? On the limits of boosted trees for object detection[C]// IEEE International Conference on Pattern Recognition, December 4-8, 2016, Cancun, Mexico. New York: IEEE, 2016: 3350-3355.