文章编号：2095-2163(2021)12-0133-06 中图分类号：TP391.41 文献标志码：A

基于通道与空间特征选择融合的人体姿态检测

杨洪智，丁学明，姬建林
（上海理工大学 光电信息与计算机工程学院，上海 200093）

摘 要：针对人体姿态检测过程中多尺度特征表达不充分的问题，本文利用选择性卷积核网络（Selective Kernel Networks，SKNet）的思想，提出了一种通道与空间特征选择融合模块，并应用于高分辨率网络，从而在多尺度特征融合过程中进行关键信息选择，不仅提高了多尺度特征表达，同时保留了原有多尺度特征融合交换不同特征信息的优点。实验结果表明，在高分辨率网络中加入通道与空间特征并联模块后，进一步提高了人体姿态检测的精度。在两种不同网络深度的模型中，姿态关键点预测的平均准确率分别有 0.6%和 0.7%的精度提升，最后通过网络推理过程可视化，进一步分析了该模块在卷积过程中起到的作用。

关键词：人体姿态检测；高分辨率网络；特征选择融合模块

Human pose estimation based on channel and spatial feature selection and fusion

YANG Hongzhi, DING Xueming, JI Jianlin
(School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

【Abstract】 In order to overcome the inadequate expression of multiscale features in the process of multiscale feature fusion for human key point detection, in this paper, a channel and feature selection fusion module is proposed and applied to a Deep High-Resolution network to select key information in the process of multiscale feature fusion, which is inspired by the ideas for Selective Kernel Networks (SKNet). This not only improves the expression of multiscale features, but also preserves the advantages of the original multiscale feature fusion in exchanging different feature information. The experimental results show that the accuracy of human posture detection is further improved by adding the parallel module of channel and spatial feature in the Deep High-Resolution network. In the two models with different network depths, the average accuracy of pose key point prediction is improved by 0.6% and 0.7% respectively. Finally, the role of this module in convolution process is further analyzed through the visualization of network inference process.

【Key words】 human pose estimation; Deep High-Resolution networks; feature selection and fusion

0 引 言

2D 人体姿态检测是继深度学习快速发展背景下一个新兴的领域，当前处于计算机视觉领域研究的重点，可以应用在动作识别、异常行为检测、体育健身指导、步态分析等多个领域。相比于传统的基于语义分割的人像分割技术，人体姿态检测能更好的反映人体的关节部分在空间上的位置关系。

与传统的图像分类，目标检测任务不同，2D 人体姿态检测和语义分割都属于像素级感知的预测任务，而前两种任务为特征级感知任务。像素级感知预测任务的最大特点是网络对图像的空间感知力要强，才能准确的赋予像素点的任务属性，进而进行像素分割，或者热图回归。此类任务在研究的初期是利用图像分类的骨干网络作为特征提取模块，如 ResNet、VGG 等主流主干网络，依次降低空间分辨率，并提高通道分辨率进行特征编码。为满足像素级感知任务的特征输出，通常需要设计相应的解码器，包括上采样和反卷积等方式，以恢复空间分辨率[1]。根据这种设计方法，研究人员设计了 Simple Baseline 网络以解决 2D 人体姿态检测任务[2]。而文献[3]认为在卷积过程中保持高分辨率特征将更精细的描绘被检测物体的细节，其实验结果证明这种高分辨率输入输出网络对像素级预测任务准确度具有较高的精度提升。与传统编解码网络不同，高分辨率网络是通过人工设计的多尺度特征融合模块来完成不同感受野之间信息交互，其网络特点是高精度定位和丰富的语义表达能力[4]。

基金项目：国家自然科学基金(61673277)。

作者简介：杨洪智(1996-)，男，硕士研究生，主要研究方向：深度学习和机器视觉；丁学明(1971-)，男，博士，副教授，主要研究方向：智能控制、系统辩识以及嵌入式系统；姬建林(1996-)，男，本科生，主要研究方向：系统辨识、模式识别、图像处理。

通讯作者：丁学明 Email：xuemingding@ usst.edu.cn

收稿日期：2021-09-13

高分辨率网络多尺度表征融合的方式是经过一系列恢复或降低分辨率后将两个不同尺度的特征图进行特征相加融合,这种融合方式并未充分利用多尺度特征的优势。2019 年 Li 等人设计了一个选择性核单元(Selective Kernel Networks, SKNet)的构造块,允许每个神经元基于多尺度的输入信息自适应地调整其关注通道,是一种动态选择机制[5]。该模块将分支中的信息引导的 softmax 注意力来融合具有不同核大小的多个分支。这种特征通道选择的思想可以追溯到 2017 年,Hu 等人提出的通道注意力机制 SE 模块[6]。在此基础上 2018 年 S 等人提出了结合通道和空间特征选择的注意力 CBAM(Convolutional Block Attention Module)模块,使得视觉检测类任务预测结果得到较好的精度提升[7]。

针对高分辨率网络多尺度表征融合模块的不足和选择性核单元的优势,本文将其相互结合,设计改进后的高分辨率人体姿态检测网络,不仅能够在特征融合过程中增强特征通道表达,还引入空间注意力的思想,增强特征选择在像素级别的表达,从而提升人体姿态检测性能,优化复杂场景下姿态检测力度不足的问题,并将通道和特征选择两种方式进行串行和并行融合,挑选出适合本任务的结合方式。在实验中发现,模块经过 Softmax 模块后会导致网络特征表达能力弱化。为了补偿被弱化的信息,本文在高分辨网络中加入了一种简单的补偿机制,很好的解决了此问题。

本文将卷积特征选择模块引入高分辨网络特征融合部分,进行多尺度特征选择融合;改进原卷积通道特征选择模块,融合通道与空间卷积特征选择模块(Dual Selective Kernel, DSK),并进行两种融合方式的对比实验;针对特征选择后的特征弱化问题,提出了一种补偿机制,证明了其有效性;特征图可视化分析,更清晰的展现在卷积过程中每层特征图对最终的网络预测做出的贡献。实验结果显示,引入改进的模块后,其网络预测精度有一定的提升,且仅仅增加了很小部分的参数量和计算量。

1 高分辨率网络

深度高分辨网络(Deep High Resolution Net, HRNet)是 Sun 等人于 2019 年提出的强 Baseline 网络,在端到端预测的网络结构中保持了高分辨率特征表示[8]。网络预测从一个高分辨率子网作为第一级开始,逐渐增加高到低分辨率的子网,子网特征图将表示更高维的信息,从而形成更多的级,多个不同分辨率子网并行连接,并在网络卷积到一定深度后进行重复的多尺度融合,使得不同分辨率表征之间相互促进,有效融合了由于感受野不同带来的全局和局部信息[4],网络形状类似于直角三角形网络,如图 1 所示。

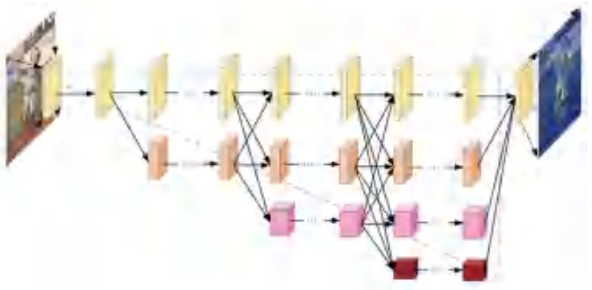

图 1 HRNet 网络结构

Fig. 1 HRNet network structure

特征融合的方式参考级联金字塔的方式,低分辨率特征恢复高分辨特征,采用一次上采样,高分辨率降为低分辨率采用 3×3 卷积核,并采用特征元素对应相加的融合方式。

本文在此网络框架的基础上,改进了融合机制,在特征融合过程中加入特征选择模块,将不同语义信息的特征图加权融合,强化对预测结果贡献较大部分元素所表达的信息,从而更好的提升训练效率和预测精度。

2 特征选择模块

卷积是深度学习领域中重要的模块。卷积操作可以提高空间维度的感受野,提取图像多尺度空间信息,多尺度特征融合则是将不同大小卷积核卷积出来的图像信息进行特征融合。SKNet 是一种改善卷积神经网络不同感受野通道特征信息融合的自适应模块,其重点是关注两个不同尺度特征图中通道维度特征,并建立选择性的融合关系。这种模块通过重新激活分配来进行不同尺度特征的选择,因此可应用在各类多尺度特征融合模块中。

2018 年 S 等人提出的 CBAM 注意力模块,将特征激活从通道选择机制扩展到空间选择机制,进一步提升网络预测精度。针对人体姿态检测任务,本文将 SKNet 模块的通道特征选择扩展到空间特征选择,同时设计两种不同的结合方式,并在消融对比实验中选择对结果预测较好的结合方式。

本文将改进后的通道与特征选择融合模块应用在深度高分辨率网络各个阶段特征融合阶段,以提升网络整体预测性能。图 2 结构展现了高分辨率网络 Stage 2 阶段加入空间和通道特征选择模块融合过程。

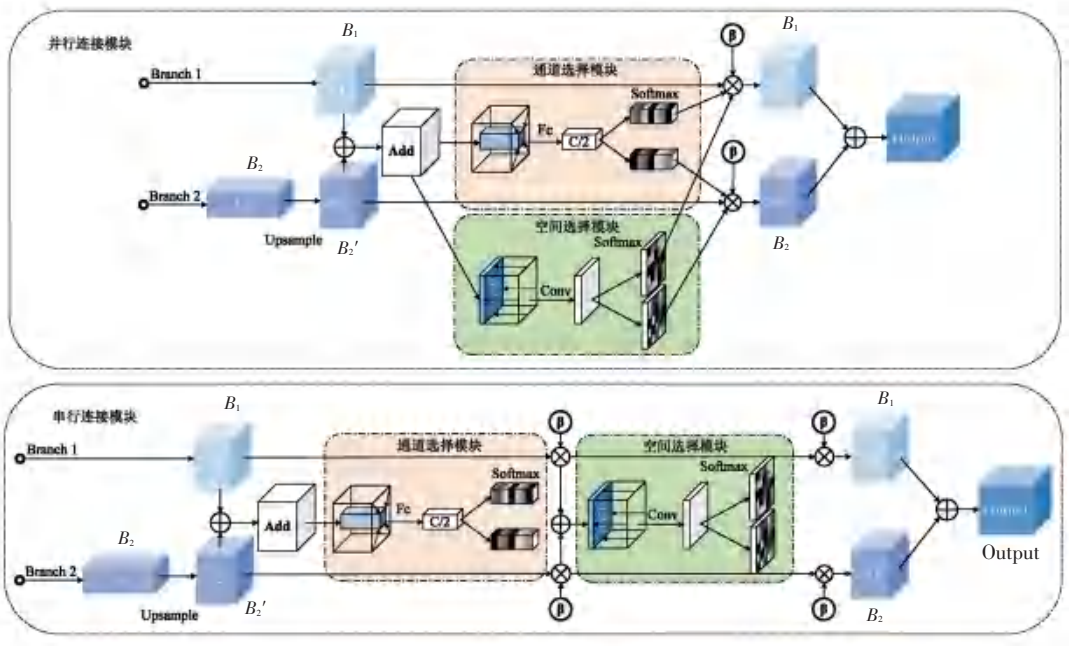

图 2　高分辨率网络 Stage 2 阶段加入空间和通道特征选择模块融合过程

Fig. 2　Adding fusion process of spatial and channel feature selection module to stage 2 of Deep-High Resolution network

3　融合通道与空间特征选择模块的高分辨率网络

SKNet 网络分为特征压缩层和特征激励层。压缩层是将卷积特征图信息进行降维，研究指出深度神经网络更偏好低维信息，且通过在 ImageNet 上不同频率的特征提取实验中证明了这一点，因此最简单的全局平均池化（global average pooling，GAP）是最好的特征降维方式[9]。针对空间和通道的两种取平均方式如式（1）和式（2）所示：

$$\boldsymbol{O}_{ch}^{'}(x)=F_{sq}^{ch}(x)=\frac{1}{h\times w}\sum_{i=1}^{h}\sum_{j=1}^{w}x(i,j) \tag{1}$$

$$\boldsymbol{O}_{sp}^{'}(x)=F_{sq}^{sp}(x)=\frac{1}{c}\sum_{i=1}^{c}y(i) \tag{2}$$

式中，$\boldsymbol{O}_{ch}^{'}(x)$ 和 $\boldsymbol{O}_{sp}^{'}(x)$ 表示通道维度和空间维度 GAP 后的降维矩阵；F_{sq} 表示压缩函数；x 为二维特征图（空间信息维度）组成的集合；y 为通道特征图（通道信息维度）组成的集合；h、w 和 c 分别表示特征图空间信息的高、宽以及通道长度信息。

降维操作将通道和空间维度上的像素级数据压缩为一个实数，表示了特征图的低频信息，基于这个低频的低维信息可以进一步进行特征激励操作。

在激励操作之前添加一个中间特征，以便更精确的特征自适应选择。对于通道特征采用全连接层降维来提升效率，并采用 *Relu* 函数激活，对于空间特征用一个 1 × 1 卷积来提升效率，并加入批归一化层（batch normalizing，BN），可加快网络训练收敛、控制梯度爆炸，并防止梯度消失及过拟合[10]，公式（3）和公式（4）如下：

$$O_{ch}^{''}(x)=F_{fc}^{ch}(x)=\delta(W_1O_{ch}^{'}(x)) \tag{3}$$

$$O_{sp}^{''}(x)=F_{\mathrm{cov}}^{sp}(x)=BN(W_2O_{sp}^{'}(x)) \tag{4}$$

式中，$O_{ch}^{'}$ 及 $O_{sq}^{'}$ 表示经过中间层的特征输出；W_1 及 W_2 为卷积权重；δ 为 *Relu* 激活函数；BN 为批归一化层；F_{fc} 为全连接函数；F_{cov} 为 1 × 1 卷积核的卷积层。

经过中间层后便是激励层，会生成两个不同的激励层去激活对应卷积分支的特征图，这里的空间和通道特征激活函数均采用原 SKNet 的 *Softmax* 激活函数。这种操作存在一个问题，即特征选择参数与原特征对应相乘后达到了重分配的效果，但是同样会弱化网络表达，导致训练缓慢。因此，在 *Softmax* 激活层后，添加一个设计的补偿系数，其大小为当前融合特征分支个数，激活后的输出用式（5）和式（6）表示：

$$\sum_{i=0}^{Bra}O_{ch}(x)=\beta\sum_{i=0}^{Bra}\sigma(W_3O_{ch}^{''}) \tag{5}$$

$$\sum_{i=0}^{Bra}O_{sq}(x)=\beta\sum_{i=0}^{Bra}\sigma(W_4O_{ch}^{''}) \tag{6}$$

式中，O_{ch} 及 O_{sq} 表示特征选择模块的最终输出；*Bra* 为 *HRNet* 网络不同阶段融合的特征图分支条数；σ 为 *Softmax* 函数；β 为补偿系数，其数值为当前融合特征分支个数。

将两种特征选择模块采用两种不同方式进行连接。采用并行连接的方式输出为式(7)：

$$\sum_{i=0}^{Bar} O(x) = \sum_{i=0}^{Bar} O_{sq} \times \sum_{i=0}^{Bar} O_{ch} \tag{7}$$

采用串行连接的方式输出为式(8)：

$$\sum_{i=0}^{Bar} O(x) = \sum_{i=0}^{Bar} O_{sq}\left(\sum_{i=0}^{Bar} O_{ch}\right) \tag{8}$$

4 实验结果与分析

本实验基于 pytorch 深度学习框架进行网络搭建，使用的计算机 CPU 为两颗 Xeon E5 2678v3，内存为 128G，显卡为 NVIDIA GeForce RTX3090，操作系统环境为 64 位 Ubuntu 18.04，训练及测试数据集采用 coco2017 数据集。

4.1 COCO2017 数据集

4.1.1 数据集简介

COCO 数据集由微软提出，包含超过 20 万张图像和 25 万个人的实例，这些实例标记了 17 个关键点[11]。划分为训练集及测试集，训练集 coco2017train 包括 118 287 张训练图像，测试集 coco2017val，包含了 5 000 张标注图像。

4.1.2 评价指标

COCO2017 数据集的标准评价指标为目标关键点相似度(object keypoint similarity, OKS)，公式(9)可通过下式表示：

$$OKS = \frac{\sum_i \exp(-d_i^2/2s^2k_i^2)\delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \tag{9}$$

其中，d_i 为目标关键点与预测关键点之间的欧式距离；v_i 为关键点坐标；s 表示目标尺度；k_i 为控制衰减的系数。

本文将采用平均精度和召回分数作为评判标准。计算 OKS 在 0.5 的 IOU 准确度 AP^{50}、OKS 在 0.75 的 IOU 准确度 AP^{75}、检测大尺寸图像实例的 IOU 准确度 AP^L 以及检测中等尺寸图像实例的 IOU 准确度 AP^M，并计算所有指标的平均准确度 *mAP* 的平均，最后计算 OKS 在 0.5～0.95 的平均召回率 *AR*。

4.2 输入图像处理

对于人体关键点检测网络固定输入图像长宽比为 4∶3，本文采用尺寸为 256×192 以及 384×288 的图像作为输入，数据增强包括随机旋转([-45,45])随机尺寸([0.65,1.35])，图像的翻转，以及半身数据增强。

4.3 训练策略

本文使用 adam 优化器。Batchsize 设置为 64，设置基本学习速率为 $1\times e^{-3}$，迭代次数为 200 轮，并在第 170 和 190 轮时学习率分别降为 $1\times e^{-4}$ 和 $1\times e^{-5}$。

4.4 结果及其分析

4.4.1 消融实验

实验首先对比了通道和空间特征选择实验对结果的影响，在此基础上添加了串行和并行两种组合方式对结果的影响。消融实验结果见表 1。

表 1 消融实验

Tab. 1 Ablation Experiment

方法	参数/M	*FLOPS*/G	*AP*
HRNet	28.5	7.10	74.4
HRNet+SK(ch)	29.1	7.12	74.6
HRNet+SK(sq)	29.1	7.12	74.7
HRNet+DSK	29.2	7.12	75.0
HRNet+DSK(serial)	29.2	7.12	74.8
HRNet+DSK(parallel)	29.2	7.12	75.0

从实验结果可以看到，通道与空间特征选择融合模块相较于原网络精度均具有一定的提升。单独的通道选择网络提升了 0.2 个百分点，而加入空间选择后提升了 0.3 个百分点，将两种网络互相结合后，进一步提高准确度，其中串行方式提高了 0.4 个百分点，并行方式连接网络精度提升了 0.6 个百分点，证明了加入模块的有效性。

4.4.2 与基准网络对比实验

coco2017val 测试集上与基准网络的对比实验结果见表 2，本实验主要与原 HRNet 网络进行对比，同时还对比了前人设计的基础网络 Hourglass[12]，CPN[13] 以及 Simple Baesline(SBL) 网络，本文训练了 2 个不同层数的网络模型，其骨干网络分别为 HRnet-W32、HRnet-W38，两种网络在结构上近似，但各自网络在卷积层数上有所区别。引入空间与通道特征选择模块后，本文的方法相较于原 HRNet 网络分别有 0.6%和 0.7%的精度提升，且增加的参数量非常少。在多层网络 HRNet-W48 中，网络分别有 0.3%和 0.4%的精度提升。

表 2　coco2017val 测试集上与基准网络的对比实验结果

Tab. 2　Comparative experiments with benchmark results on coco2017val test dataset

方法	骨干网络	输入尺寸	参数	*FLOPS*	*AP*	AP^{50}	AP^{75}	AP^{M}	AP^{L}	*AR*
Hourglass[12]	Hourglass	256 × 192	25.6M	26.2G	66.9	—	—	—	—	—
CPN[13]	ResNet-50	256 × 192	27M	6.2G	68.6	—	—	—	—	—
SBL-50[2]	ResNet-50	256 × 192	34M	8.9G	70.4	88.6	78.3	67.1	77.2	76.3
SBL-101[2]	ResNet-101	256 × 192	53M	12.4G	71.4	89.3	78.9	68.1	78.1	77.1
SBL-152[2]	ResNet-152	256 × 192	68.6M	15.7G	72.0	89.3	79.8	68.7	78.9	77.8
HRnet-W32[8]	HRnet-W32	256 × 192	28.5M	7.1G	74.4	90.5	81.9	70.8	81.0	79.8
HRnet-W48[8]	HRnet-W48	256 × 192	63.6M	14.6G	75.1	90.6	82.2	71.5	81.8	80.4
HRnet-W32+DSK	HRnet-W32	256 × 192	29.2M	7.1G	75.0	90.4	82.2	71.5	81.9	80.4
HRnet-W48+DSK	HRnet-W48	256 × 192	65.0M	14.6G	75.4	90.7	82.4	71.7	82.0	80.5
HRnet-W32[8]	HRnet-W32	384 × 288	28.5M	16.0G	75.8	90.6	82.7	71.9	82.8	81.0
HRnet-W48[8]	HRnet-W48	384 × 288	63.6M	32.9G	76.3	90.8	82.9	72.3	83.4	81.2
HRnet-W32+DSK	HRnet-W32	384 × 288	29.2M	16.0G	76.5	90.6	83.2	72.8	83.6	81.5
HRnet-W48+DSK	HRnet-W48	384 × 288	65.0M	32.9G	**76.7**	**90.8**	**83.3**	72.8	**83.8**	**81.6**

注：表中“-”意为该网络在数据集上无实验数据。

4.4.3　其他实验

为了展现引入补偿机制后网络训练过程的精度变化，通过实验得到了引入补偿机制后网络的训练精度变化图，如图 3 所示。从图中变化可以看出引入补偿机制后网络开始训练时的精度要比未加入补偿系数后的精度较高，随着训练轮数的增加两者差距逐渐减小，在经过学习率衰减后，两种网络精度都具有小部分跳动，并逐渐趋于平稳。最终加入补偿系数后的网络精度整体要高于未加入补偿后的精度。

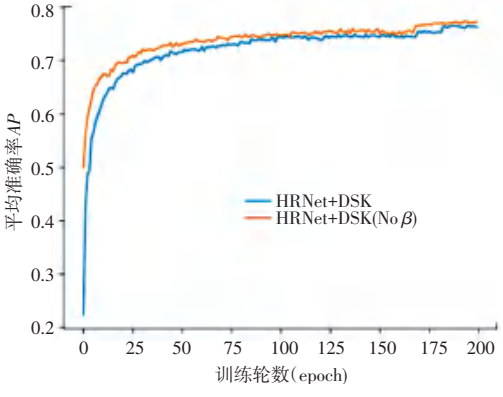

图 3　引入补偿机制后网络的训练精度变化图

Fig. 3　The change of training accuracy after introducing the compensation mechanism

4.4.4　可视化分析

为了更好的展现添加通道与空间特征选择模块对输出特征的影响。本文通过实验展示了改进后的 HRNet 的 stage3 部分网络分出 3 个并行分支过程的末尾处低维高分辨率特征图的特征可视化热图，如图 4 所示。图 4 中深蓝色部分表示网络输出特征中数值较小的区域，即网络不关心区域，而颜色为黄色甚至红色区域表示网络输出特征中数值较大的区域，即网络较为关注区域。从特征热力图中可以看到，加入融合特征选择模块后对人体轮廓集中部分信号加强，并抑制其他不相关部分，使得学习更专注，从而在后续模块中提供更有用的信息。

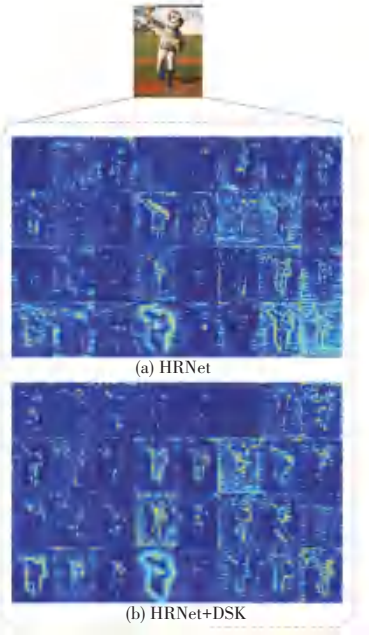

图 4　stage3 阶段特征可视化热图

Fig. 4　Feature visualization with heatmap in stage3

4　结束语

针对人体姿态高分辨率检测网络特征融合过程中不同尺度特征关注不足的问题。本文借鉴 SKNet 的思想，提出了一种结合通道与空间特征选择的高分辨率网络融合模块，并利用两种不同的结合方式，增强了不同尺度特征融合的高效性，并在 coco2017 数据集上验证了改进后的有效性，且额外增加的计算量很小。针对特征选择经过 softmax 输出后特征的表征被削弱，导致训练较慢的现象，提出了一种非常简单有效的参数补偿方法，很好的解决了这个问题。

（下转第 142 页）