基于多尺度双注意力的人体姿态估计方法研究

马皖宜, 张德平

引用本文

马皖宜, 张德平. 基于多尺度双注意力的人体姿态估计方法研究[J]. 计算机科学, 2022, 49(11A): 220100057-5.

MA Wan-yi, ZHANG De-ping. Study on Human Pose Estimation Based on Multiscale Dual Attention[J]. Computer Science, 2022, 49(11A): 220100057-5.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220100057

基于多尺度双注意力的人体姿态估计方法研究

马皖宜　张德平
南京航空航天大学计算机科学与技术学院　南京 211000
(wanyi_ma@163.com)

摘　要　针对人体姿态估计中人体与背景区分度不高,基于 HRNet 网络的人体姿态估计中重要特征信息利用不完全的问题,利用通道与空间注意力机制,提出了一种基于多尺度双注意力(Multiscale Dual Attention,MDA)的人体姿态估计方法 MDA-HRNet。该方法从通道域和空间域出发,分别设计了结合通道注意力的 Ca-Neck,Ca-Block 模块和结合空间注意力的 Sa-Block 模块,将其融入到高分辨率网络结构中,使网络能够重点关注图像中的人体区域。在 Sa-Block 模块中采用 3×3 和 7×7 的卷积核推导两种不同尺度的空间注意力映射,使网络区分人体特征和背景特征的能力更加显著,从而对人体及其关键点进行准确定位。该方法在 MPII 数据集上进行了实验验证,结果表明 MDA-HRNet 能有效地提高人体姿态估计关节点定位的准确度。

关键词:人体姿态估计;通道注意力;空间注意力;多尺度注意力映射;高分辨率网络
中图法分类号　TP391.41

Study on Human Pose Estimation Based on Multiscale Dual Attention

MA Wan-yi and ZHANG De-ping
School of Computer Science and Technology,Nanjing University of Aeronautics and Astronautics,Nanjing 211000,China

Abstract　In view of the problem of low discrimination between human body and background in human posture estimation,and incomplete utilization of important feature information in human posture estimation based on HRNet,a human posture estimation method MDA-HRNet based on multiscale dual attention is proposed by using channel and spatial attention mechanism. Considering both of the channel domain and spatial domain,the Ca-Neck and Ca-Block modules combined with channel attention and Sa-Block module combined with spatial attention are designed respectively. Then integrating these modules into the high-resolution network structure,so that the network can pay more attention to the human body area in the image. Moreover,in the Sa-Block module,3×3 and 7×7 convolution kernels are adopted to derive two spatial attention maps of different scales,which makes the ability of the network to comprehensively distinguish human features and background features more remarkable,so as to accurately locate the human body and its key points. The proposed method is tested and verified on MPII data set,and the results show that MDA-HRNet can improve the accuracy of joint point location of human posture estimation effectively.

Keywords　Human pose estimation,Channel attention,Spatial attention,Multiscale attention mapping,High resolution network

1　引言

人体姿态估计是研究基于图像的观测数据来恢复关节和躯干姿态的算法或系统,是计算机视觉领域非常具有挑战性和研究意义的方向之一。其中二维姿态估计算法主要通过获取二维图像上人体关键点的位置信息和人体肢干的位置和方向信息,来获得人体关键点坐标与骨骼对应关系,该对应关系的准确度直接影响着人体姿态估计结果好坏[1]。

现阶段流行的人体姿态估计网络模型主要是基于 ResNet[2]、Hourglass[3]、HRNet[4]和生成对抗网络[5]这 4 种骨干神经网络。ResNet 通过引入残差模块以及使用跳跃连接的残差结构,缓解了网络深度增加导致的退化问题;Hourglass 网络能够得到有效的多尺度特征信息,卷积和最大池化将特征降到很低的分辨率,达到最低分辨率时通过上采样重新组合不同尺度下的特征,有利于复杂场景中的人体姿态估计;HRNet 能够在整个网络中维持高分辨率的特征表示,并行连接子网络,重复进行多尺度融合,使预测热图更接近 ground truth;生成对抗网络结合人体骨架先验信息,通过对抗训练来提高关节点预测的正确性,由此提高姿态估计的准确度。其中 HRNet 网络更具优越性,其网络结构能够生成含有丰富语义信息的高分辨率特征图,各子网的卷积层还采用了跳跃连接的方式,包含多个残差单元,可以有效避免过拟合问题。

在实际的分析应用中,人体姿态估计算法总是存在无法准确区分人体与背景的问题,同时,在特征提取的过程中,由于各网络的卷积操作是一种局域化操作,因此多分辨率特征融合时不能很好地利用人体区域特征,使特征输出图丢失部分有效信息,影响人体姿态估计准确度。注意力机制以高权重聚焦重要信息,低权重忽略不相关的信息,并根据任务结果反向指导特征图的权重调整,使得在不同的情况下都能选取重要信息,因此它不仅能提升任务完成的高效性,而且具有

基金项目:国防基础科研重点项目(JCKY2020605C003)
This work was supported by the National Defense Basic Scientific Research Key Program(JCKY2020605C003).
通信作者:张德平(depingzhang@163.com)

很好的可扩展性和鲁棒性[6]。注意力机制还能够通过共享重要信息使其他神经元或神经网络进行信息交换，实现重要信息的传递，提升神经网络的学习能力[7]。基于此，本文将注意力机制引入 HRNet 网络模型，结合空间域和通道域的双注意力机制，采用多尺度注意力映射，设计了一种基于多尺度双注意力机制的高分辨率网络的人体姿态估计方法 MDA-HRNet，使网络能够更集中地关注到图像中的人体区域和人体重要关节点，从而提高人体姿态估计准确度。本文的主要贡献如下：

(1)基于通道注意力机制，分别对不同分辨率子网的特征信息进行通道域注意力系数加权，抑制无用信息，重点关注人体关键信息，对多分辨率特征信息进行有效提取和利用。

(2)基于多尺度空间注意力机制对特征图进行自适应特征细化，采用两种不同尺度的注意力映射，利用多尺度信息，综合全局图像区分人体特征和背景特征，对人体及其关键点进行准确定位。

(3)在 MPII 数据集上进行验证，获得 *PCKh*@0.5 评价标准下的平均准确率为 90.5%，较 HRNet 基线模型提升了 0.9%。

2 MDA-HRNet

针对人体姿态估计不能准确区分人体与背景的问题，以及自下而上(Bottom-top)范式的 HRNet 不能充分学习重要特征信息导致人体关键点定位不准确的问题，本文利用通道域和空间域的双注意力机制以及多尺度注意力映射，设计了一种基于多尺度双注意力机制的人体姿态估计方法 MDA-HRNet，其网络结构如图 1 所示。MDA-HRNet 一共分为 4 个阶段(stages)，stage 1 为图像预处理阶段，stage 2－stage 4 为 HRNet 的并行子网。在 stage 2 最高分辨率子网的输入处添加融入通道注意力的 Ca-Neck 模块，在 HRNet 的各个并行子网中添加融入通道注意力的 Ca-Block 模块，在各分辨率特征融合之前添加融合多尺度空间注意力的 Sa-Block 模块，最后将最高分辨率子网的输出特征图用于人体姿态估计。

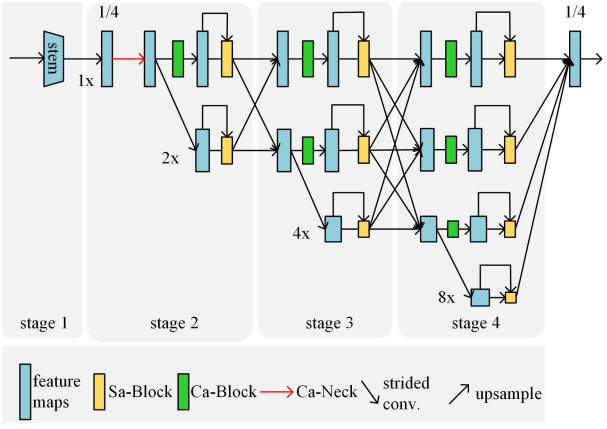

图 1 MDA-HRNet 网络结构

Fig. 1 MDA-HRNet network structure

2.1 Ca-Neck 和 Ca-Block

在卷积神经网络中，通道维度是充分提取特征信息的一个要点，通道注意力机制从各通道之间的依赖性的角度考虑，让网络有选择性地增强信息量大的特征，使得后续处理可以充分利用这些特征，并对无用特征进行抑制。自动驾驶公司 Momenta[8]在 2017 年公布的一种图像识别结构 Squeeze and Excitation Networks(SEN)就是一个具有代表性的通道注意力模型，它聚焦于通道维度，学习通道特征的相关性，提出一种有效的结构单元 SE，能够对各通道的特征响应值进行自适应性调整，强化重要的特征信息，提高模型的准确率。基于此，本文设计了融入通道注意力机制的 Ca-Neck 和 Ca-Block 两个模块，Ca 即为通道注意力(Channel Attention)。Ca 能够捕捉人体图像中局部跨通道特征信息交互，利用特征的通道间的关系来生成通道注意力图。它能对各个通道所对应的空间值进行约束，挤压输入特征图的空间维度，完成去空间化。为了汇集空间信息，Ca-Neck 和 Ca-Block 不仅采用了平均池化(Average Pooling)，还采用了最大池化(Max Pooling)，以获得更精细的通道方向上的注意力。这一经验依据来自 CBAM[9]，即相比单独使用二者之一，同时利用这两种池化得到不同特征，进行特征融合后可以显著提高网络的表达能力。

Ca-Neck 模块的网络结构如图 2 所示，其是基于原 HRNet 网络中的瓶颈模块(Bottleneck)设计而成。经过预处理阶段的特征图 $\boldsymbol{F}\in\mathbb{R}^{H\times W\times C}$ 作为 Ca-Neck 的输入，分别进行空间上的最大池化和平均池化获得两个 $1\times1\times C$ 的通道表示 $\boldsymbol{P}_{\mathrm{CAmax}}(\boldsymbol{F})$ 和 $\boldsymbol{P}_{\mathrm{CAavg}}(\boldsymbol{F})$，通过卷积核大小为 k 的一维卷积，获得两个通道方向注意力，即 ***MaxAttention*** 和 ***AvgAttention***，再使用元素求和合并这两个输出特征向量，得到最终的通道注意力：

$$\begin{aligned}\boldsymbol{\omega}_{\mathrm{ca}}&=\sigma(C1D_k(\boldsymbol{P}_{\mathrm{CAmax}}(\boldsymbol{F}))+C1D_k(\boldsymbol{P}_{\mathrm{CAavg}}(\boldsymbol{F})))\\&=\sigma(\boldsymbol{W}_1(\boldsymbol{F}^{c}_{\max})+\boldsymbol{W}_1(\boldsymbol{F}^{c}_{\mathrm{avg}}))\end{aligned}\tag{1}$$

其中，$C1D$ 表示一维卷积，σ 表示 sigmoid 函数，两个输入共享权重 $\boldsymbol{W}_1$。Ca-Neck 在通道注意力后面还连接了 3 个 3×3 卷积核大小的二维卷积层、Batch Normalization(BN)正则化和 ReLU 操作，以及一个残差连接(包含二维卷积和 BN 操作)。

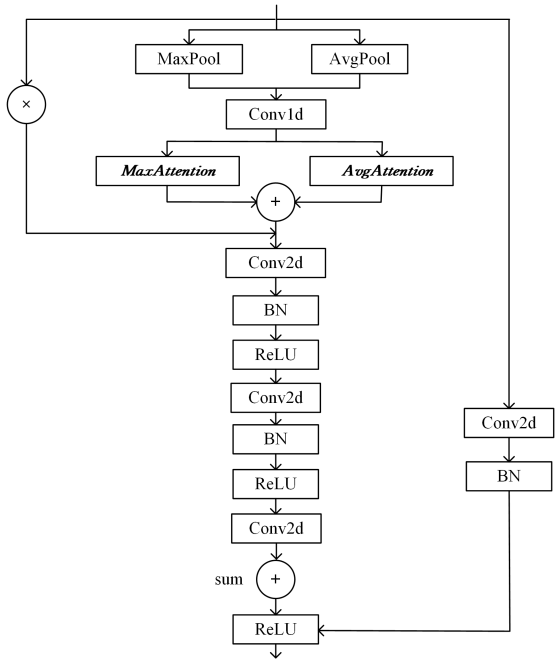

图 2 Ca-Neck 网络结构

Fig. 2 Ca-Neck network structure

将通道注意力机制融入由残差模块构成的 Ca-Block，其网络结构如图 3 所示，是基于原 HRNet 的残差模块(Basicblock)设计的。Ca-Block 模块先通过对输入特征图的不同通道进行注意力系数加权，得到有效的人体特征信息，然后

进行卷积操作，再将初始输入的特征信息与卷积操作后输出的结果通过残差连接（不包含二维卷积和 BN 操作）进行求和，最后进行输出，提取到更有效的通道特征信息。

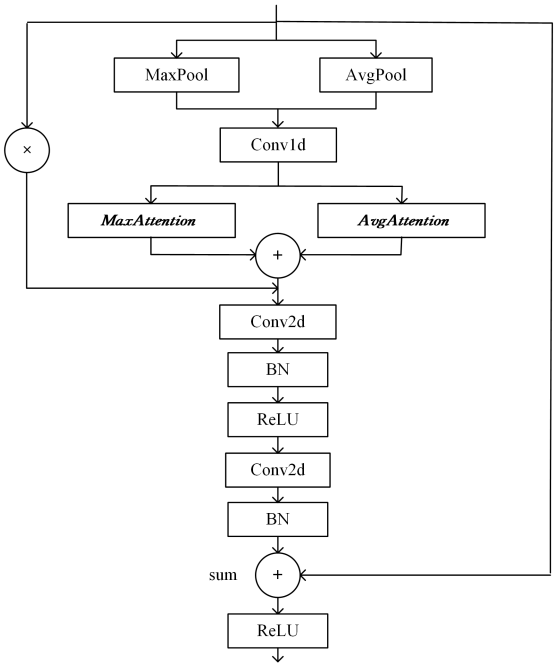

图 3　Ca-Block 网络结构

Fig. 3　Ca-Block network structure

2.2　多尺度 Sa-Block

在图像检测中，图像中不同区域对任务的贡献各不相同。在人体姿态估计任务中，只有与人体相关的区域才最需要被重点关注，进一步才会关注到人体的关键关节点。引入空间注意力的目的是寻找图像中最重要的位置进行处理，也就是人体区域部分及人体关键点。空间注意力对人体图像中所有通道的特征图进行标准化约束，通过 sigmoid 函数获得只含有空间信息的注意力权重，通过注意力加权增强图像上人体特征区域，抑制非人体特征区域，提高人体与背景的区分度，增强网络对人体区域以及人体关键点的定位能力，提高人体姿态估计的准确度。

基于此，本文设计了融入空间注意力机制的 Sa-Block 模块，Sa 即为空间注意力（Spatial Attention）。经典的空间注意力机制，如 Google DeepMind[10] 提出的 Spatial Transformer Network 和 Almahairi 等[11] 提出的 Dynamic Capacity Networks，二者均通过对通道进行平均池化来汇集通道信息。与 Ca-Block 中分别进行空间上的最大池化和平均池化一样，Sa-Block 模块也采用最大池化和平均池化来抑制通道信息，有利于 Sa-Block 模块进行注意力推断。另外，针对图像中人体尺度大小不一导致网络鲁棒性差的问题，借鉴 Inception-Net[12] 网络结构的设计思想，采用不同大小的卷积核推导多个尺度的注意力映射，使网络能够更好地利用多尺度特征信息，综合全局图像来区分人体特征和背景特征，对人体及其关键点进行准确定位。Sa-Block 模块网络结构如图 4 所示。首先对输入特征图进行通道上的最大池化和平均池化，使用元素求和合并这两个输出特征向量，然后采用 3×3 和 7×7 大小的卷积核推导两个尺度的注意力映射，通过平均通道池化融合两个注意力推断结果，再以通道叠加的方式连接起来。

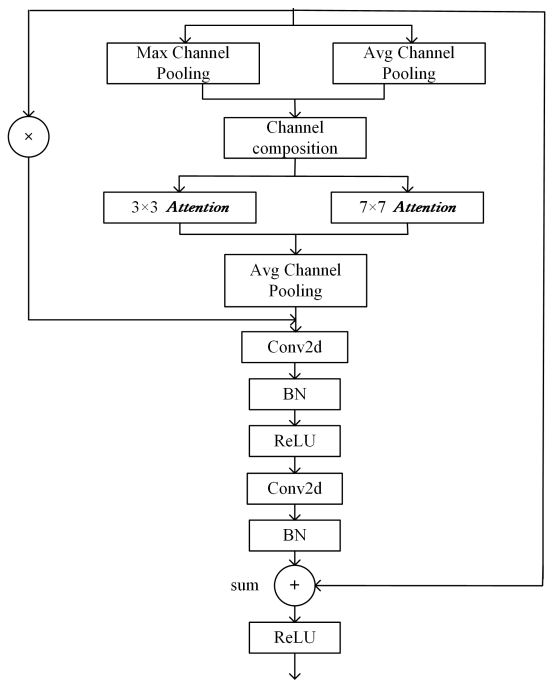

图 4　Sa-Block 网络结构

Fig. 4　Sa-Block network structure

对于第 l 个 Sa-Block 模块的输入特征图 $\boldsymbol{F}_l \in \mathbb{R}^{H\times W\times C}$，经过最大通道池化和平均通道池化分别得到两个 $1\times1\times C$ 的池化表示 $\boldsymbol{P}_{\mathrm{SAmax}}(\boldsymbol{F}_l)$ 和 $\boldsymbol{P}_{\mathrm{SAavg}}(\boldsymbol{F}_l)$，将其通过 3×3 和 7×7 卷积核大小的卷积层融合注意力映射得到 $\boldsymbol{W}_2(\boldsymbol{F}_l)_{3\times3}$ 和 $\boldsymbol{W}_2(\boldsymbol{F}_l)_{7\times7}$：

$$\boldsymbol{W}_2(\boldsymbol{F}_l)_{3\times3}=f_{3\times3}[\boldsymbol{P}_{\mathrm{SAmax}}(\boldsymbol{F}_l);\boldsymbol{P}_{\mathrm{SAavg}}(\boldsymbol{F}_l)] \tag{2}$$

$$\boldsymbol{W}_2(\boldsymbol{F}_l)_{7\times7}=f_{7\times7}[\boldsymbol{P}_{\mathrm{SAmax}}(\boldsymbol{F}_l);\boldsymbol{P}_{\mathrm{SAavg}}(\boldsymbol{F}_l)] \tag{3}$$

其中，$f_{3\times3}$ 和 $f_{7\times7}$ 分别表示卷积核大小为 3×3 和 7×7 的卷积操作。得到两种尺度的注意力映射后，再通过一次平均通道池化来融合这两个结果，并通过 sigmoid 函数将结果规范化到[0，1]。

$$\boldsymbol{\omega}_{\mathrm{sa}}=\sigma(\boldsymbol{P}_{\mathrm{SAavg}}(\boldsymbol{W}_2(\boldsymbol{F}_l)_{3\times3};\boldsymbol{W}_2(\boldsymbol{F}_l)_{7\times7})) \tag{4}$$

其中，σ 表示 sigmoid 激活函数。与 Ca-Block 模块相似，Sa-Block 在空间注意力后连接了 2 个 3×3 卷积核大小的二维卷积层、BN 操作和 ReLU 操作以及一个残差连接（不包含二维卷积和 BN 操作）。Sa-Block 通过多尺度空间注意力的作用，使整个网络能集中关注图像中的人体区域和人体关键点，从而提高人体姿态估计的准确度。

3　实验与分析

3.1　实验设置

本实验在 Linux 内核的 Ubuntu 18.04 版本的操作系统、python 版本为 3.6.5 和 pytorch 版本为 1.4.0+cu101 以及 1 个 NVIDIA Tesla T4 GPU 组成的服务器上完成。在实验中，选择 Adam 优化器来优化模型，初始学习率为 0.001。

本实验采用 MPII 数据集进行训练和测试，该数据集是应用比较广泛的一个基准数据集，有 25 000 张带标注的图片和 40 000 多人的实例，包含单人与多人图像。在每个检测人体上有 16 个标注的关键点，分别是右脚踝、右膝盖、右髋关节、左髋关节、左膝盖、左脚踝、骨盆、胸部、上颈、头顶、右手腕、右手肘、右肩、左肩、左手肘、左手腕。

为了方便与其他方法进行对比，本实验将 MPII 数据集图像以人体骨盆为中心进行裁切，将图像尺寸重新裁剪为固定大小 256×256，将人体检测框调整为固定的宽高比 4:3，以方便网络训练。针对 MPII 数据集中存在的一些不完整的人体图像的问题，对训练图像进行了数据增强操作，包括对数据集进行随机旋转[−45°,45°]、随机缩放规模[0.65,1.35]和随机翻转操作等。

3.2 实验设置

本实验的验证方法采用 *PCKh*[13] 评价标准，该标准针对图像中所有人体执行明确的边界限制操作，以得到检测准确率。具体来说，就是给定边界框内的候选区域，通过控制阈值 r 的不同标准来获得不同的检测准确率。一般情况下选择 $r=0.5$（用 *PCKh*@0.5 表示，其中 r 越小表示评价标准越严格），计算预测关键点与相应 ground truth 关键点之间的归一化距离小于 r 的占比情况。*PCKh* 主要是针对 MPII 数据集，以头部长度（Head Length）作为归一化参考，采用欧氏距离，将该人体的头部框尺度作为归一化其他部位的距离。如果检测的人体关键点与 ground truth 的距离没有超过阈值范围，则判定该检测结果正确。人体中第 k 个关键点的 *PCKh* 值为：

$$PCKh(k)=\frac{1}{N}\sum_{i=1}^{N}\begin{cases}1, & |T_k^i-\widetilde{T}_k^i|\leqslant rS_h\\ 0, & |T_k^i-\widetilde{T}_k^i|\geqslant rS_h\end{cases} \tag{5}$$

其中，N 表示目标图片总量，T_k^i 表示第 i 张图片中的第 k 个人体关键点的 ground truth 数值，$\widetilde{T}_k^i$ 表示该关键点的预测结果，S_h 表示头部框尺度，将 ground truth 与预测结果之间的距离与 $0.5S_h$ 进行比较，若前者小于或等于后者，则判定该检测结果正确。

3.3 实验结果与分析

本节对比了我们提出的 MDA-HRNet 与其他先进的人体姿态估计方法的实验结果，并在 MPII 验证集上进行消融实验，对比了原 HRNet 网络、融入通道域与空间域双注意力的高分辨率网络和融入多尺度双注意力的高分辨率网络的实验结果。最后，将 MDA-HRNet 方法进行可视化实验结果展示和分析，直观展示了该方法的预测过程和测试效果。表 1 列出了 MDA-HRNet 方法与其他流行网络在 MPII 数据集上的验证结果对比，结果显示，在 256×256 大小的输入下，MDA-HRNet 在 *PCKh*@0.5 的标准下获得了 90.5 的准确率得分，比原 HRNet 的得分 89.6 提高了 0.9 个百分点，比其他的流行网络的估计结果都更好。另外从实验结果也可以看出，MDA-HRNet 在身体各部位的检测效果也较好，表明 MDA-HRNet 提高了人体姿态估计关键点检测的准确度。

表 1 MPII 验证集上的实验结果对比（*PCKh*@0.5）

Table 1 Comparison of experimental results on MPII verification set(*PCKh*@0.5)

method	*Hea.*	*Sho.*	*Elb.*	*Wri.*	*Hip.*	*Kne.*	*Ank.*	*Total*
8-stage Hourglass[3]	96.5	96.0	90.3	85.4	88.8	85.0	81.9	89.2
CPN[14]	96.5	96.0	90.4	86.0	89.5	85.2	82.3	89.6
PRM[15]	96.8	96.0	90.4	86.0	89.5	85.2	82.3	89.6
DLCM[16]	95.6	95.9	**90.7**	**86.5**	**89.9**	86.6	82.5	89.8
DeeperCut[17]	95.6	95.9	**90.7**	**86.5**	**89.9**	86.6	82.5	89.8
SimpleBaseline[18]	97.0	95.9	90.3	85.0	89.2	85.3	81.3	89.6
HRNet	97.1	95.5	89.8	85.4	88.5	85.5	81.9	89.6
MDA-HRNet	**97.4**	**96.2**	90.6	86.4	89.3	**86.8**	**83.3**	**90.5**

为了验证双注意力机制和多尺度注意力映射的有效性，本文在 MPII 数据集上进行消融实验，其中 CA-HRNet 方法只添加了通道注意力，即只添加了 Ca-Neck 和 Ca-Block；MA-HRNet 方法混合添加了通道注意力和空间注意力两种机制，但在空间注意力机制中只通过 7×7 卷积核大小的卷积操作，得到一种尺度的注意力映射来获得空间注意力权重；而 MDA-HRNet 则通过 3×3 和 7×7 卷积核大小推导两种不同尺度的注意力映射，再通过平均通道池化融合这两种注意力推断结果。

表 2 列出了这两种方法与原 HRNet 验证实验的结果对比。根据实验结果可以看出，虽然 MA-HRNet 相比原 HRNet 检测效果有所提升，但增加了多尺度注意力映射的方法的人体姿态估计效果更好，证明了双注意力机制与多尺度注意力映射的有效性。

表 2 MPII 验证集上的注意力消融实验结果对比

Table 2 Comparison of experimental results of attention ablation on MPII verification set

method	Ca	One-scale Sa	Multiscale Sa	*Hea.*	*Sho.*	*Elb.*	*Wri.*	*Hip.*	*Kne.*	*Ank.*	*Mean*@0.5	*Mean*@0.1
HRNet				97.1	95.5	89.8	85.4	88.5	85.5	81.9	89.6	36.4
CA-HRNet	√			97.1	95.8	89.9	85.6	88.8	85.9	82.1	90.0	36.6
MA-HRNet	√	√		97.1	96.1	90.4	86.3	89.2	86.4	82.8	90.3	36.4
MDA-HRNet	√		√	**97.4**	**96.2**	**90.6**	**86.4**	**89.3**	**86.8**	**83.3**	**90.5**	**37.9**

本文采用了 *PCKh* 的评价标准来可视化实验过程与结果。

如图 5 所示，进行人体姿态估计时，首先得到热图（Heatmap）图像，然后预测人体关节点的坐标，再将热图结果映射到原图像上，经过整合即可得到人体对应的 16 个关节点的位置坐标，实现根据关键点检测完成人体姿态估计的任务。如图 6 所示，对图像上的各种人体姿势，包括不同尺度和不同受遮挡程度的人体图像，MDA-HRNet 都有很好的关键点检测结果，表明了 MDA-HRNet 具有鲁棒性，能准确检测出人体各个关节点位置，完成人体姿态估计。

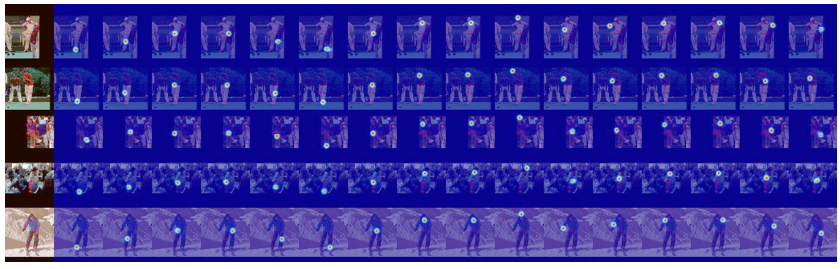

图 5 关键点预测热图示意图

Fig. 5 Schematic diagram of key point prediction heatmap

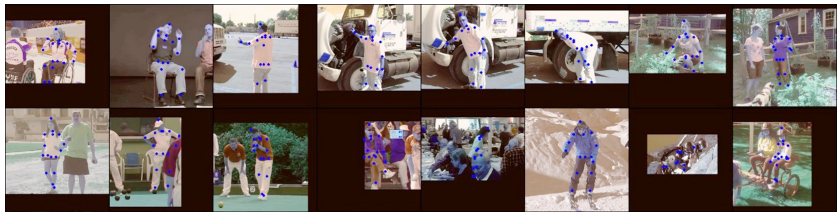

图 6　在 MPII 数据集上的结果可视化

Fig. 6　Visualization of results on MPII dataset

结束语　本文针对人体图像与背景的区分度不高，以及 HRNet 中重要信息不能被充分利用的问题，提出了基于多尺度双注意力的人体姿态估计方法 MDA-HRNet。本文设计的 Ca-Neck，CA-Block 和 SA-Block 模块分别关注通道域和空间域的重要人体特征信息，并且通过空间多尺度注意力映射，使整个网络能更加有效地关注人体区域的特征信息。最后在 MPII 数据集上进行了验证实验和消融实验，验证实验的对比结果表明，当输入图像的大小固定为 256×256 时，相比原 HRNet 网络，MDA-HRNet 的 *PCKh*@0.5 得分提升了 0.9 个百分点，表明 MDA-HRNet 提升了人体姿态估计的准确度；消融实验结果表明了双注意力机制和多尺度注意力映射在人体姿态估计上的有效性。基于神经网络的人体姿态估计的方法在不断提升准确度，因此未来的工作重点是在保证精度的前提下加快检测速度。

参 考 文 献

[1] ZHOU Y，LIU Z Q，ZENG F Z，et al. Survey on Two-dimensional Human Pose Estimation of Deep Learning[J]. Journal of Frontiers of Computer Science and Technology，2021，15(4)：641-657.

[2] 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. 2016：770-778.

[3] NEWELL A，YANG K，JIA D. Stacked Hourglass Networks for Human Pose Estimation[C]// European Conference on Computer Vision. Springer International Publishing，2016.

[4] SUN K，XIAO B，LIU D，et al. Deep High-Resolution Representation Learning for Human Pose Estimation[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019.

[5] GOODFELLOW I J，POUGET-ABADIE J，MIRZA M，et al. Generative Adversarial Networks[J]. Advances in Neural Information Processing Systems，2014，3：2672-2680.

[6] HAO S，LEE D H，ZHAO D. Sequence to sequence learning with attention mechanism for short-term passenger flow prediction in large-scale metro system[J]. Transportation Research Part C：Emerging Technologies，2019，107：287-300.

[7] ZILLICH M，FRINTROP S，PIRRI F，et al. Workshop on attention models in robotics：visual systems for better HRI[C]// Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction. New York：ACM，2014：499-500.

[8] JIE H，LI S，GANG S. Squeeze-and-Excitation Networks[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). IEEE，2018.

[9] WOO S，PARK J，LEE J Y，et al. CBAM：Convolutional Block Attention Module[C]// European Conference on Computer Vision. 2018.

[10] JADERBERG M，SIMONYAN K，ZISSERMAN A. Spatial transformer networks[J]. Advances in Neural Information Processing Systems，2015，28：2017-2025.

[11] ALMAHAIRI A，BALLAS N，COOIJMANS T，et al. Dynamic capacity networks[C]// International Conference on Machine Learning. PMLR，2016：2549-2558.

[12] SZEGEDY C，WEI L，JIA Y，et al. Going deeper with convolutions[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). IEEE，2015.

[13] ANDRILUKA M，PISHCHULIN L，GEHLER P，et al. Human Pose Estimation：New Benchmark and State of the Art Analysis[C]// Computer Vision and Pattern Recognition(CVPR). IEEE，2014.

[14] CHEN Y，WANG Z，PENG Y，et al. Cascaded Pyramid Network for Multi-person Pose Estimation[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). IEEE，2018.

[15] YANG W，LI S，OUYANG W，et al. Learning feature pyramids for human pose estimation[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017：1281-1290.

[16] TANG W，YU P，WU Y. Deeply learned compositional models for human pose estimation[C]// Proceedings of the European Conference on Computer Vision(ECCV). 2018：190-206.

[17] INSAFUTDINOV E，PISHCHULIN L，ANDRES B，et al. Deepercut：A deeper，stronger，and faster multi-person pose estimation model[C]// European Conference on Computer Vision. Cham：Springer，2016：34-50.

[18] XIAO B，WU H，WEI Y. Simple baselines for human pose estimation and tracking[C]// Proceedings of the European Conference on Computer Vision(ECCV). 2018：466-481.

MA Wan-yi，born in 1996，postgraduate，is a member of China Computer Federation. Her main research interests include image processing and artificial intelligence modeling.

ZHANG De-ping，born in 1973，Ph. D，postgraduate supervisor，is a member of China Computer Federation. His main research interests include image processing and artificial intelligence modeling.