文章编号:1671-4598(2025)11-0073-10 DOI:10.16526/j.cnki.11-4762/tp.2025.11.009 中图分类号:TP391 文献标识码:A

基于自适应跨维加权网络的轻量型人体姿态检测

王 力,谷日涵
(中国民航大学 电子信息与自动化学院,天津 300300)

摘要:人体姿态检测的核心是准确检测人体关键点,由于高分辨率网络存在着一定局限性,对此,提出一种自适应跨维加权高分辨率网络;针对网络跨维信息交互不足的问题,采用跨维分离卷积方法提取信息,实现了在空间和通道之间有效的信息交换;针对关键点定位不精确的问题,采用自适应上下文建模方法,通过自适应变换和输入特征的空间加权,增强了网络捕捉复杂空间关系的能力,使得网络能够提取多尺度上下文信息并建立跨维度依赖关系,从而在不增加计算复杂度的情况下提高了准确性;此外,还引入了坐标注意力机制融合来自不同分支和规模的特征,使检测准确性得到进一步提升;经 COCO 和 MPII 数据集实验测试,与主流轻量型网络相比,自适应跨维加权高分辨率网络性能更好,兼顾了效率与精度。

关键词:人体姿态估计;轻量型网络;注意力机制;分割卷积;特征融合

Lightweight Human Posture Detection Based on Adaptive Cross-dimensional Weighted Networks

WANG Li, GU Rihan
(School of Electric Information and Automation, Civil Aviation University of China, Tianjin 300300, China)

Abstract: The core of human posture detection is to accurately detect the key point of the human body. Due to the limitation of high-resolution networks, an adaptive cross-dimensional weighting high-resolution network (ACW-HRNet) is proposed. In response to insufficient cross dimensional information intersection in the network, a cross dimensional separation convolution method is adopted to extract information, achieving effective information exchange between space and channels. For the unclear positioning of key points, an adaptive context modeling (ACM) is adopted to enhance the network's capacity to capture complex spatial relationships through adaptive transformation and spatial weighting of input features. This approach enables the network to extract multi-scale contextual information and establish cross-dimensional dependencies, thus improving accuracy without increasing computational complexity. Moreover, a coordinate attention mechanism is introduced to fuse the multi-branch and multi-scale features, further enhancing detection accuracy. Through experimental tests on the COCO and MPII datasets, compared with the mainstream lightweight networks, the ACW-HRNet has better performance and balances efficiency and accuracy.

Keywords: human posture estimation; lightweight networks; attention mechanism; segmentation convolution; feature fusion

0 引言

在计算机视觉领域,二维人体姿态估计一直是重要且极具挑战性的问题,具有广泛的应用场景,例如人体动作识别、人机交互、虚拟现实、视频监控、人体轨迹

收稿日期:2024-10-10; **修回日期**:2024-11-19。
基金项目:中央高校基本科研基金项目(3122018S003);民航局安全能力建设基金项目([2024]28);民航局安全能力建设基金项目([2023]50)。
作者简介:王 力(1973-),男,博士,教授。
通讯作者:谷日涵(1998-),男,硕士研究生。
引用格式:王 力,谷日涵.基于自适应跨维加权网络的轻量型人体姿态检测[J].计算机测量与控制,2025,33(11):73-82.

跟踪等。它的主要目的是根据输入图像或视频帧准确预测人体关键点的空间位置，包括关节，如肘部和膝盖。在人体姿态检测中，高分辨率表示已被证明可以提高模型性能[1-2]。但也带来了更高的计算复杂性。在设计姿态检测模型时，平衡计算效率和高性能是一个关键的考虑因素。

高分辨率网络（HRNet）[3]在人体姿势检测方面表现出卓越的性能。然而，其高度的复杂性为在资源受限的设备上实现带来了挑战。学者们最近的研究工作[4-7]一直致力于提高网络效率。虽然小型 HRNet 可以减小 HRNet 的深度和宽度。但是，这种降低会导致性能大幅下降。Lite-HRNet[4]表明，在轻量级高分辨率网络中，可以通过用高效的卷积模块替换标准的 1×1 卷积来实现网络复杂性和准确性之间的平衡。然而，这种独立于输入的固定网络结构在一定程度上限制了模型的性能。Dite-HRNet[5]在动态轻量级块中引入嵌入的动态分割卷积，从而进一步提高网络计算效率。因此，将研究集中在设计与输入相关的自适应模块上，认识到特定操作可能会在网络内的不同位置和不同的输入大小下产生不同的效果[6]。

近年来，特别是在人类姿态检测的背景下，在卷积权重的生成和聚合中建立空间或通道维度之间相互依赖关系的能力已经受益匪浅[7-8]。最近的研究[9-10]利用通道加权、空间加权或两者的组合来提高神经网络的性能。CBAM（Convolutional Block Attention Module）[10]中，注意力权重是通过通道和空间注意力生成的，并应用于特征以对其进行优化。然而，这些方法没有考虑到跨维度信息交换对性能的潜在优势。因此，需要解决如何更有效地在空间和通道维度之间建立跨维度的相互依赖关系，以提高轻量级高分辨率网络的性能。

为了解决上述问题，引入了一种自适应跨维加权高分辨率网络（ACW-HRNet）。设计了一个自适应跨维加权（ACW，adaptive cross-dimensional weighting）模块来提高计算效率。具体来说，ACW 模块包括自适应上下文建模（ACM，adaptive context modeling）模块和跨维分割卷积（CSC，cross-dimensional split convolution）模块，使网络能够提取多尺度信息并在空间和通道维度之间建立相互依赖关系。ACM 包含自适应分辨率加权（ARW，adaptive resolution weighting）和自适应空间加权（ASW，adaptive spatial weighting），从而通过保留丰富的要素表示来增强网络捕获要素空间关系的能力。CSC 模块提取多尺度信息并促进跨维度信息交互。此外，使用超参数微调复杂性和准确性之间的权衡。这些高效的模块显著提高 ACW-HRNet 的性能。

1 轻量型人体姿态检测网络

最近的研究[11]见证了人类姿态检测的重大进步。早期的方法主要依赖于手工制作的功能和传统的机器学习算法。然而，随着卷积神经网络的采用，不仅准确性得到了提高，鲁棒性也得到了提高，使其更适用于实际应用。

1.1 2D 人体姿势姿态检测

二维（2D）人体姿势检测有两种主要方法：自上而下的方法和自下而上的框架。自上而下的框架从人体检测器开始，用于识别图像中的个体，然后是单人姿势检测。相比之下，自下而上的框架最初采用端到端方法来定位所有关键点，并根据相应的个体对它们进行分组。近年来，由于网络更强大、更深入，2D 人体姿势检测方法的性能有了显著的改进，但在资源受限的设备上实现仍然具有挑战性。

1.2 轻量级高分辨率网络

HRNet[3]具有多分支结构，通过在并行多分辨率子网之间交换数据来有效地融合多尺度信息。然而，由于其计算复杂性高，实现效率一直是一个挑战。Lite-HRNet[4]引入了一个条件通道加权模块，该模块由通道注意力机制组成，以替代 shuffle 块中计算成本高昂的逐点卷积。Dite-HRNet[5]为卷积核赋予动态属性，提高网络计算效率并增强长距离空间依赖性。

1.3 高效的卷积神经网络

高效的神经架构设计在构建具有较低参数和计算需求的深度网络方面具有很大的前景，MobileNet 等成功模型就是例证，ShuffleNet 等移动网络分解 $k\times k$ 卷积为深度卷积和逐点卷积。移动网络 V2 引入倒置残差和线性瓶颈结构。ShuffleNetV2 介绍了 Channel Shuffle 的概念，为设计轻量级网络提供了有价值的见解。

1.4 条件权重生成

SENet[9]通过对权重进行学习，实现了对特征图通道间相互依赖性的建模，从而增强网络的表示能力。CBAM[10]通过通道和空间注意力生成注意力权重，将这些注意力权重应用于特征以对其进行优化。值得注意的是，虽然 CBAM 中的渠道关注提供了显著的性能改进，但它并没有考虑捕获可能有利于提高性能的跨维度交互。

2 自适应跨维加权高分辨率网络

2.1 ACW-HRNet

ACW-HRNet 网络包括一个高分辨率子网络作为其初始阶段，在并行连接后续子网络的同时逐渐降低分辨率。随后添加的子网的分辨率是前一个子网的一半，通道数的两倍。ACW-HRNet 的网络架构如表 1 所示。第一阶段作为主干模块，由 3×3 stride 为 2 和 shuffle 块的卷积。洗牌块的整体结构如图 1 所示。该模块不仅可以提取特征，还可以有效地压缩特征图的分辨率。每个

后续阶段该结构由一个特征融合模块和两个 ACW 模块构成，这些模块协同工作，以促进跨分辨率信息的交流。网络最终输出的高分辨率表征将用于后续的姿态检测任务。ACW-HRNet 网络结构的直观说明如图 1 所示。为了将 ACW-HRNet 与主流轻量级高分辨率网络进行基准测试，开发了两种变体：ACW-HRNet-18 和 ACW-HRNet-30。这些变体提供类似于 Lite-HRNet 和 Dite-HRNet 的网络深度和宽度。此外，ACW-HRNet 表现出的计算和参数复杂度水平均较高。

表 1　ACW-HRNet 的结构

层数	运算符	输出通道	ACW-HRNet-18	ACW-HRNet-30	分支
主干层	3×3 卷积(×1) Shuffle 模块(×1)	32	1	1	1
二层	ACW 模块(×2) 融合模块(×1)	40,80	2	3	2
三层	ACW 模块(×2) 融合模块(×1)	40,80, 160	4	8	3
四层	ACW 模块(×2) 融合模块(×1)	40,80, 160,320	2	3	4

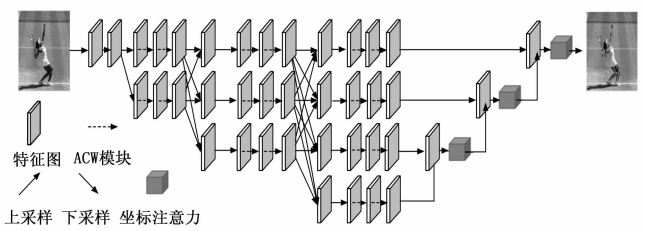

图 1　ACW-HRNet 结构图

2.2　ACW 模块设计

ACW 模块是网络架构的关键组件。最近的研究强调了整合一个广泛的感受野以整合具有不同空间尺度的特征的重要性。这些增强功能使网络能够更好地捕获更精细的特征。基于这些发现，本文设计了 ACW 模块并将其整合到的网络中。ACW 模块的整体结构如图 2(b) 所示，它包括从 ShuffleNetV2 衍生而来的通道分裂、特征连接和通道随机操作。这些操作用于聚合从各个网络图层中提取的不同要素。AC 模块在一半的通道上运行，由一系列层组成，每个层在增强特征表示方面发挥着独特的作用。它包括 3 个主要业务：ARW、CSC 和 ASW。ARW 和 ASW 都是 ACM 方法的实例化。ARW 操作、CSC 操作和 ASW 操作协同工作，以丰富网络内的空间信息。ACW 模块无缝集成到 ACW-HRNet 网络架构中，使网络能够有效地捕获特征关系和更精细的本地特征。

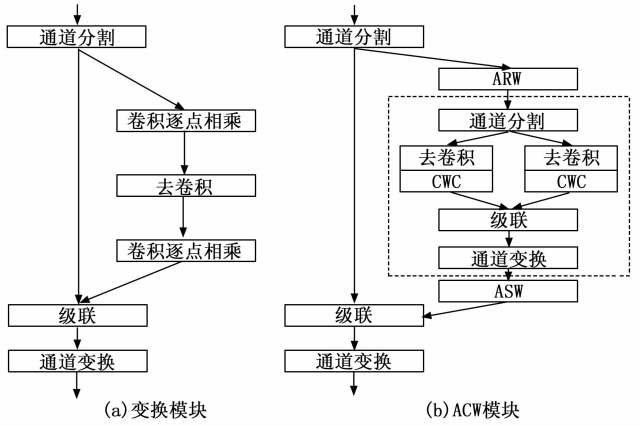

图 2　模块示意图

3　自适应跨维加权高分辨率网络的关键方法

3.1　跨维度拆分卷积

3.1.1　Split-Concat-Shuffle（SCS）模块

SCS 模块是网络设计中的关键组件。在计算机视觉中，使用大型卷积核可以提供更广泛的感受野，从而允许提取广泛的空间信息。然而，在所有网络层中统一使用大型卷积核有其局限性。它可能导致计算和参数复杂性增加，并可能阻碍网络捕获更精细局部特征的能力。为了应对这些挑战，提出了 SCS 模块，该模块旨在通过利用不同大小的卷积核来捕获多尺度空间信息。SCS 模块通过拆分、连接和随机排列特征图来运行。SCS 模块无缝集成到网络架构中，增强了网络提取各种规模空间信息的能力。通过这个模块，在大感受野的好处和捕获更精细的局部特征的需求之间取得了平衡，最终提高了网络的性能。

在 SCS 模块中，初始步骤涉及将通道等分设置成若干组别。随后，在各个通道组别内，同步执行不同内核尺寸的深度卷积运算。此过程可以表示为：

$$\begin{aligned} N &= C/G \\ y_i &= \mathrm{DWConv}(k_i \times k_i \mid N)(x_i) \\ k_i &= 2i+1, i \in [1,G] \end{aligned} \tag{1}$$

其中：x_i 和 y_i 代表深度卷积在 i 处的输入和输出。$\mathrm{DWConv}(k_i \times k_i \mid N)(\cdot)$ 是核大小为 $k_i \times k_i$，通道维度为 N，组间通道总数为 C，通道组数为 G 的深度卷积。

然后，将来自不同组的特征连接起来。为了促进特征图之间的有效信息交换，在拼接后执行 通道转换操作。值得注意的是，SCS 模块的设计方式不会显著增加网络的宽度。它不是简单地添加更多通道，而是将现有通道划分为不同的组，并并行执行各种内核大小的深度卷积。这种方法有助于保持网络的计算效率。计算效率的高低与网络性能的权衡可以通过调整超参数来微调。仔细选择此超参数可以进行优化，同时将计算复杂度保

持在可接受的水平。

SCS 模块的设计使得通过调整几个关键超参数来平衡计算效率和网络性能成为可能。主要超参数包括通道组数 G，以及卷积核大小 k_i。这些参数的选取规则和平衡方法如下：

对于通道组数 G，选取规则为适当增大 G 值会将总通道数分成更多的组，每组进行更小规模的卷积运算，从而减少计算复杂度，提升计算效率。较小的 G 值则会合并更多的通道进行较大规模的卷积，有助于捕捉复杂的特征；平衡方法为可以在浅层使用较大的 G 值（分为更多组）来保持细粒度的特征提取能力，而在深层采用较小的 G 值来捕获更多全局特征。

对于卷积核大小 k_i，选取规则为在层次较浅的组中使用小尺寸核（如 3×3 或 5×5），在较深层使用较大尺寸核（如 7×7），可以在保持感受野的同时，避免在所有组中使用大核带来的高计算开销；平衡方法为核大小的选择应依据任务需求的精度和计算资源。在计算资源受限的情况下，可减少大卷积核的使用频率。

对于通道转换操作，选取规则为通道转换可以通过逐点卷积实现，以控制计算量和信息融合的深度。较低频率的通道转换会降低计算负担，但可能减少特征图的信息交换能力；平衡方法为逐点卷积的频率应适中，以便在提升通道间信息交流和降低计算复杂度之间找到平衡。

在多尺度卷积的组合层面，SCS 模块通过使用不同的卷积核大小，从不同尺度提取特征，提供了多尺度空间信息的捕捉能力。这种方法在感受野和精细特征提取之间实现了平衡。例如，浅层使用小核捕获局部信息，深层使用大核捕捉更广泛的空间关系；在超参数微调层面，可通过实验测试不同 G 和 k_i 组合的性能表现，以找到计算效率和精确度的最佳折中点。在训练过程中逐步增加 G 或减少 k_i，观察其对网络收敛速度和性能的影响，找到适合的参数；在层级化设置层面，为了提高效率，在早期网络层可选择较大的 G 值和较小的 k_i，在后期层选择较小的 G 值和较大的 k_i，以适应不同网络层对细节和全局信息的需求；通过合理调节这些超参数并结合层次化设置和多尺度卷积策略，SCS 模块能够高效地平衡计算复杂度和性能表现。

3.1.2 跨维度权重计算（CWC）模块

如图 3 所示，CWC（Cross-dimensional weight computation）模块由 3 个并行分支组成，每个分支都有不同的作用。其中两个分支负责捕获渠道维度之间的跨维度交互 C 和空间维度 H 或 W。第三个分支结合了通道和空间注意力。每个分支都执行针对其维度对定制的特定操作。然后将 3 个分支的输出相加以取平均值。这种方法保证了不同维度之间的全面信息交换。

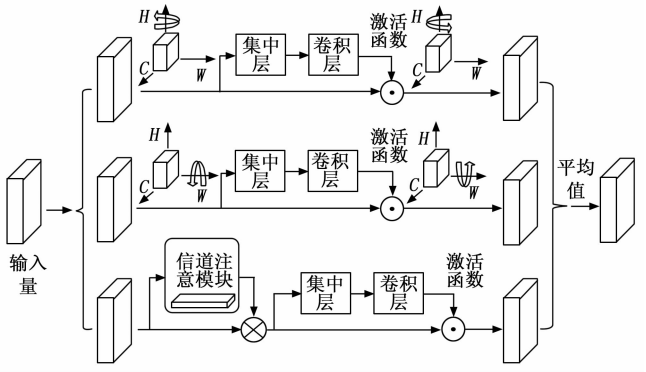

图 3 CWC 模块结构图

这集中层结合了两个跨维度的池化特征，有效地将张量的第一个维度减少到 2。这种减少实现了张量内容的全面表示，同时也降低了深度，从而提高了计算效率。此过程可以表示为：

$$Z-\mathrm{pool}(X)=[\mathrm{Maxpool}_{0d}(X),\mathrm{AvgPool}_{0d}(X)] \quad (2)$$

式中，$0d$ 指沿第一维度执行 Max pooling 和 Average pooling 操作。

将 CWC 定义为一个三分支模块。给定一个输入张量 $X\in R^{C\times H\times W}$ 它首先通过 CWC 模块的每个分支。跨维度加权过程可以表示为：

$$\omega_i=\sigma[\alpha_i(\widetilde{X}_i^{*})],i\in[1,3] \quad (3)$$

$$Y=\frac{1}{3}(\overline{\overline{\widetilde{X}_1\omega_1}}+\overline{\overline{\widetilde{X}_2\omega_2}}+\widetilde{X}_3\omega_3) \quad (4)$$

式中，σ 表示 sigmoid 激活函数，α_i 表示 3×3 跨维权重的 3 个分支中的卷积，ω_i 代表在第 i 个跨维加权分支中获得的注意力权重。$\overline{\overline{\widetilde{X}_1\omega_1}}$ 和 $\overline{\overline{\widetilde{X}_2\omega_2}}$ 表示进行顺时针旋转九十度，以维持输入图形的初始状态（$C\times H\times W$）。

将 SCS 模块与 CWC 模块合并为一个卷积单元的过程需要综合两者的功能：SCS 模块用于捕获多尺度空间信息，CWC 模块则通过跨维度的注意力机制在通道和空间维度上提取特征。合并后的模块通过多尺度卷积和注意力机制，实现对特征的精细调节和更广泛的空间信息捕捉。以下为将 SCS 模块与 CWC 模块合并的具体步骤。

特征分组与多尺度卷积（基于 SCS 模块）：给定特征张量 $X\in R^{C\times H\times W}$，首先沿通道方向将 X 等分为 G 组，每组通道数为 $N=C/G$；在每个组内，应用不同的卷积核尺寸进行深度卷积操作，满足公式 $k_i=2i+1$，$i\in[1,G]$，实现多尺度卷积，从而捕获多尺度空间特征；得到的特征输出为 $y_i=DWConv(k_i\times k_i\mid N)(x_i)$，其中 $DWConv$ 是深度卷积，k_i 为不同卷积核尺寸，生成具有多尺度特征的子输出。

多分支跨维度注意力机制（基于 CWC 模块）：在得到多尺度特征图后，将其传递到 CWC 模块的 3 个并

行分支中，每个分支进行不同的跨维度和注意力计算。第一分支（空间维度交互）：将特征图沿 H 轴旋转 90 度，维度变为 $W\times H\times C$，然后执行跨维度池化（Max Pooling 和 Average Pooling）操作，这一步旨在利用空间维度的不同特征信息；第二分支（通道维度交互）：将特征图沿 W 轴旋转 90 度，维度变为 $H\times C\times W$，并进行相同的跨维度池化操作；第三分支（通道与空间注意力结合）：直接应用通道注意力机制，获得注意力权重 ω_i 并施加到特征图上以调整重要特征。

权重计算与特征融合：在每个分支中，使用公式 $\omega_i=\sigma[\alpha_i(\widetilde{X}_i^*)]$，其中 σ 是 sigmoid 激活函数，α_i 表示在每个分支中的 3×3 卷积操作，生成权重 ω_i；对每个分支输出的特征图乘以注意力权重后 ω_i，旋转回原始维度 $C\times H\times W$ 的方向，进行融合：$Y=\frac{1}{3}(\overline{X_1\omega_1}+\overline{X_2\omega_2}+\overline{X_3\omega_3})$。

特征拼接与通道转换：将各组多尺度特征图 y_i 结合在一起并拼接，以生成整合后的特征张量；执行通道转换操作，以便特征图之间的信息能够有效交换，同时保持原始通道数量，保证计算效率。

合并模块具备以下的优势：在多尺度信息捕获方面，SCS 模块的多尺度卷积实现了空间特征的广泛捕获，使模型能适应不同尺度的特征；在跨维度与注意力加权方面，CWC 模块在空间和通道维度间施加注意力机制，使模型能够灵活地关注重要特征；在计算效率方面，合并后的模块通过共享通道和空间维度的特征，避免简单增加网络宽度，保持计算复杂度在可控范围内；这种合并后的卷积模块在计算开销上保持效率，同时通过多尺度和注意力机制，提升了网络对空间和通道维度特征的表达能力，有效平衡了计算效率与网络性能。

3.2　自适应上下文建模

ACM 模块的作用是通过捕获要素的空间关系来增强网络的制图表达功能。

ACM 模块的过程可以表示为：

$$Y=X\odot W_s \tag{5}$$

W_s 表示权重贴图，而⊙表示两个矩阵的元素乘法。

计算 W_s 的过程分为两步：（1）自适应空间池化：这个初始步骤涉及获取上下文掩码。它是通过 1×1 卷积，后跟 SoftMax 层。然后，使用一系列分辨率自适应转换将生成的掩码应用于特征图。这个过程会产生空间上下文特征，如图 4 所示。（2）信息融合：在自适应空间池化之后，下一步涉及进一步整合信息。这是通过两个 1×1 具有非线性激活操作的卷积，然后对生成的特征图进行归一化以增强其表达能力。计算 W_s 的完整过程如下所示：

$$W_s=\mathrm{weight}[X,\mathrm{ASPool}(H,W)(X)] \tag{6}$$

ASPool（H，W）表示自适应空间池化操作，将输入特征 X 汇总到特定输出大小（$H\times W$），（Weight）（·）表示信息融合操作。

为了进一步利用高分辨率网络的性能，将 ACM 转化为 ARW 和 ASW。这在并行多分辨率架构中充分释放了它的潜在优势。

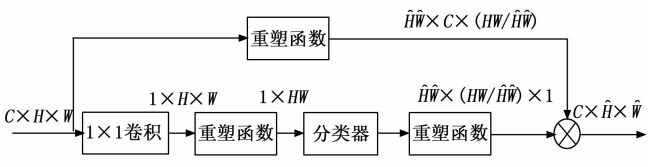

图 4　自适应空间池化过程

3.2.1　自适应分辨率权重

ARW 的实现过程如图 5 所示。在 ACW-HRNet 的第 n 层中，具有不同 n 的分辨率被收集到最低分辨率 $H_m\times W_m$ 然后连接在一起进行信息融合。随后，通过上采样恢复它们的分辨率 $H_m\times{}_m$。之后进行信息融合，用于后续加权操作。在第 n 层，所有特征图都缩小到与各自阶段相关的最小分辨率大小，这个过程可以表示为：

$$X_k\in R^{C\times H\times W}\rightarrow\mathrm{ASPool}(H_m,W_m)\rightarrow \hat{X}_k\in R^{C\times H_m\times W_m} \tag{7}$$

X_k 表示第 k 个最高分辨率的输入张量，$H_m\times W_m$ 代表最小分辨率大小。将所有池化特征连接在一起，进行密集建模，并在信息融合后通过上采样恢复其分辨率，以进行后续的加权操作，如下所示：

$$\overline{X}=f([\hat{X}_1,\cdots,\hat{X}_(n-1),\hat{X}_n]) \tag{8}$$

$$\overline{X}\rightarrow\mathrm{Conv.}\rightarrow\mathrm{Hardswish}\rightarrow\mathrm{Conv.}\rightarrow\mathrm{sigmoid}\rightarrow(W'_1,W'_2,\cdots W'_{n-1}) \tag{9}$$

其中：Conv. 表示 1×1 卷积，W'_n 表示权重映射，Hardswish 表示 Hardswish 激活函数，而 f（·）表示特征串联。对 $n-1$ 个权重映射（W'_1，W'_2，…，W'_{n-1}）的相应解析度进行上采样，输出 W_1，W_2，…，W_{n-1}。每个权重映射都用于对相应分辨率下的原始特征进行加权，最终得出第 k 个分支的输出特征 Y_k。此过程可表示为：

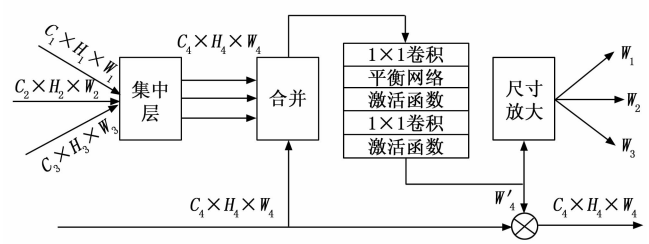

图 5　ARW 模块结构图

$$Y_k = \begin{cases} X_k \odot W_k, 1 \leqslant k \leqslant n-1 \\ X_k \odot W'_n, k = n \end{cases} \tag{10}$$

W_k 表示权重映射（大小为 $C_k \times H_k \times W_k$）。

3.2.2 自适应空间加权（ASW）

ASW 的实现过程如图 6 所示。为了捕获具有相同分辨率的要素之间的空间关系，整个网络都采用了 ASW 操作。当自适应空间池的输出大小为 1×1，它会压缩维度的输入特征 $C \times H \times W$ 到 $C \times 1 \times 1$。随后，提取的特征被输入到一个由两个 1×1 具有非线性激活操作的卷积。此操作将生成从 0 到 1 的权重系数，这些系数将用于后续的加权操作。

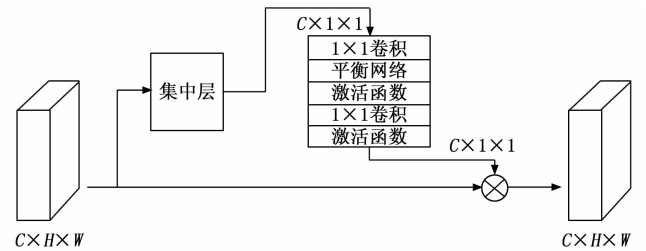

图 6 ASW 模块结构图

ASW 操作旨在捕获系统中具有相同分辨率的要素之间的空间关系，从而提供丰富的上下文信息。另一方面，ARW 操作经过定制，用于捕捉不同分辨率要素间的空间联系。以上两种做法都擅长保留丰富的特征表示，使网络能够提取丰富的上下文信息。它们可以有效地替换逐点卷积，进一步增强网络的功能。

3.3 坐标注意力

坐标注意力（Coordinate Attention）将通道注意力分解为两个特征编码过程，这两个过程沿两个空间方向聚合特征，以使用精确的位置信息来增强感兴趣对象的表示。它的优势在于压缩输入 X 进入二维 H 和 W，分别通过自适应池化操作。具体而言，对于输入 X，我采用两种不同的池化核空间配置（H，1）和（1，W），分别对各个通道沿水平与垂直方向进行编码。输出和宽度为 H 的一个通道的输出可以表示为：

$$\begin{aligned} Y_W &= \sum_{0 \leqslant \omega < W} y(\omega) = \sum_{0 \leqslant \omega < W} \frac{\sum_{0 \leqslant i < H} x(i, w)}{H} \\ Y_H &= \sum_{0 \leqslant h < H} y(h) = \sum_{0 \leqslant h \leqslant H} \frac{\sum_{0 \leqslant i < W} x(h, i)}{W} \end{aligned} \tag{11}$$

$y(w)$ 表示特征向量中每行的平均值，Y_W 表示将这些平均值叠加到特征向量的一列中。接下来 Y_H 和 Y_W 通过旋转特征矩阵进行融合，然后拆分为两个维度的特征，两个维度的输出为 f_H 和 f_W 分别通过 3×3 卷积操作。同时，注意不同维度对特征图的每个通道的重要性，然后，利用这个重要性给每个特征分配一个权重值，这样神经网络就可以专注于位置信息。输出 Y_{block} 可以写成：

$$y(i,j) = x(i,j) * f_H * f_W$$

f_H 和 f_W 表示卷积、池化和 h-swish 激活函数，激活函数公式如下：

$$h\text{-swish}[x] = x \frac{RuLU6(x+3)}{6} \tag{12}$$

4 实验结果与分析验证

4.1 数据集和评估指标

COCO 数据集包含 20 多万张图片和 25 万个人物实例，每个实例都标注了 17 个关键点如图 7 所示。在训练集上训练网络，该数据集包含 57 000 幅图像和 150 000个人物实例。在验证集和测试集上对网络进行评估。评估采用基于对象关键点相似性的平均精度（AP）和平均召回率（AR）分数。为了进一步验证网络，在 MPII 人体姿态数据集上进行了实验。MPII 数据集是由现实场景活动中拍摄的图像构成，约有25 000 张图像，包含 40 000 多个均被标注了 16 个关节点信息的人体目标如图 8 所示。其中 28 000 个目标用来作训练集，11 000个用来作测试集。

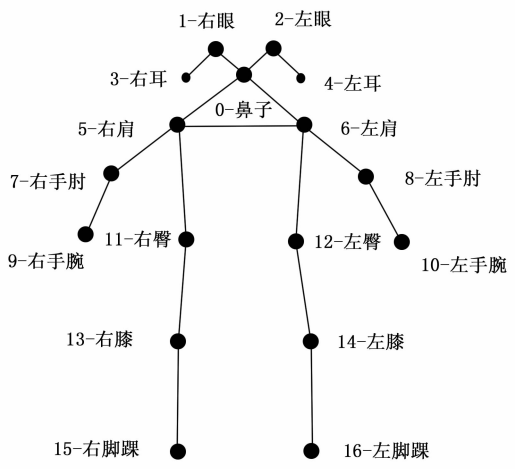

图 7 COCO 骨架结构

为了有效衡量实际关节点和预测关节点位置之间的相似度，使用对象关节点相似度（OKS，object keypoint similarity）作为衡量指标，公式如下：

$$OKS = \frac{\sum_i \exp\left(-\frac{d_i^2}{2s^2 k_i^2}\right)\delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \tag{13}$$

i 为关节点的顺序；

d_i 为表示有关真实值和预估值的欧几里得距离同时位于第 i 个关节点；

s 为对象尺度；

k_i 为每个关节点的减幅量的控制系数；

v_i 为人为划分的规则是：可见性值为 0，当关节点不生成时；可见性值为 1，当关节点生成但不可观时；可见性值为 2，当关节点完全可观时。

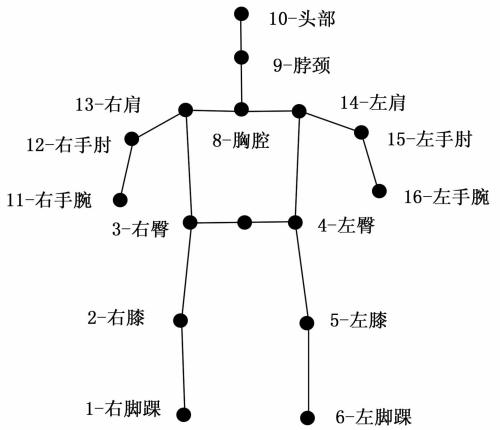

图 8　MPII 骨架结构

δ 为对是否生成关节点进行整体分析，如未生成关节点，即规定 $\delta=0$；如生成了关节点，则规定 $\delta=1$。

以 OKS 这一衡量指标为参考，可以阐述平均准确率 AP^T 这一广泛应用于轻量型人体姿态估计与识别的参考量。AP^T 的计算公式如下：

$$AP^T = \frac{\sum_k^M \delta_k(OKS > T)}{M} \tag{14}$$

T 为临界值；

k 为实验对象顺序；

M 为实例数量总和；

δ_k 为将临界值 T 与对象关节点相似度进行对比，如果对象关节点相似度超过了临界值，则 δ_k 值为 1。

准确度使用正确关键点概率（PCKh）分数进行评估，并根据头部大小进行归一化处理，公式如下所示：

$$PCKh = \frac{\sum_a \delta\left(\frac{d_{ai}}{d_a^h} \leqslant T_c\right)}{\sum_a 1} \tag{15}$$

δ 为条件判断函数，若括号内条件成立，$\delta=1$，否则 $\delta=0$；

c 为第 c 个阈值；

T 为人工设定的阈值；

d_a^h 为 第 α 个人的头部尺度因子；

d_{ai} 为表示有关真实值和预估值的欧几里得距离同时位于第 α 个人体上的第 i 个关节点；

i 为第 i 个关节点；

α 为第 α 个人。

采用的指标为 PCKh@0.5，即 $T_c=0.5$ 时的 PCKh 值。

实验采用的评价标准具体如下：AR 表示 10 个不同节点的平均召回率，方法为在对象关节点相似度为 0.5 到 0.95 之间，每隔 0.05 取一个值，共取 10 个节点。AP 表示在对象关节点相似度从 0.5 到 0.95 之间，每隔 0.05 取一个值，共计 10 个不同节点临界值的平均准确率。AP^M 表示中等实验目标的准确率，AP^{75} 表示在对象关节点相似度为 0.75 时的准确率，AP^{50} 表示在对象关节点相似度为 0.5 时的准确率，AP^L 表示大型实验对象的准确率。

为了方便对改进后的模型进行测试，拟定 3∶4 为人体检测框的宽度和高度比例，按此标准对 COCO 以及 MPII 数据集中的图片预先进行裁剪。MPII 数据集的裁剪尺寸为 256×256，而 COCO 数据集的裁剪尺寸则为 192×256 和 288×384。为了进行输入数据的增强，对 MPII 数据集进行了随机旋转 30 度和 0.3 倍的比例，而 COCO 数据集则进行了随机旋转 45 度和 0.3 倍的放缩。这两个数据集均采用了翻转测试，并对半身数据进行了增强。在训练过程中，使用了 Adam 优化器，初始学习率为 0.001，总共进行 210 轮训练，在第 160 和 190 轮时进行学习率的降低，比例为 0.1。

4.2　实验结果

4.2.1　COCO 验证集

表 2 显示了 ACW-HRNet 与常用方法的比较结果。在输入大小为 256×192 的情况下，ACW-HRNet-18 和 ACW-HRNet-30 的 AP 分别为 66.2 和 69.0，优于主流的轻量级网络。与 MobileNetV2 相比，ACW-HRNet-18 和 ACW-HRNet-30 的性能分别提高了 1.6 和 4.4。与 ShuffleNetV2 相比，ACW-HRNet-30 仅用 25% 的计算量就实现了 9.1 的 AP 增长。在使用 384×288 分辨率作为输入大小时，ACW-HRNet-18 的 AP 得分为 69.9，与 Dite-HRNet-18 相比提高了 0.9。具体来说，ACW-HRNet-30 在输入大小为 384×288 时的 AP 得分最高，达到 72.1，仅用 40% 的计算量就比小型 HRNet 的 AP 值显著提高了 16.1。与大规模姿势检测网络（如 CPN 和 Sample Baseline 相比，ACW-HRNet 以更低的计算复杂度实现了更高的精度。在计算复杂度几乎相同的情况下，ACW-HRNet 优于主流轻量级网络 Lite-HRNet 和 Dite-HRNet，分别提高了 1.7 和 0.6。图 9 显示了 ACW-HRNet-30 在 COCO 上的可视化结果，包括单人、多人和遮挡场景。

4.2.2　COCO 测试集

表 3 显示了自适应跨维加权网络与其他主流网络的比较结果。ACW-HRNet-18 的 AP 值为 67.2，与其他小型网络相比，显示出更高的准确性和效率。与 Lite-HRNet-18 相比，ACW-HRNet-18 的 AP 提高了 0.5，而模型复杂度保持不变。与 MobileNetV2 相比，ACW-HRNet-18 仅用 12% 的计算量就将 AP 值提高了 0.6。随着模型规模的增加，ACW-HRNet-30 的 AP 值为 69.5，与 ShuffleNetV2 相比 AP 提高了 6.6，而计算量仅增加了 25%。虽然本网络性能可能无法与更大的网络相提并论，但它以其卓越的计算效率脱颖而出。

表 2 COCO 验证集性能对比

模型	输入尺寸	参数量/M	计算量/G	AP/%	$AP^{0.5}$/%	$AP^{0.75}$/%	AP^{M}/%	AP^{L}/%	AR/%
HRNet[3]	256×192	28.5	7.1	73.4	89.5	80.7	70.2	80.1	78.9
Small HRNet[1]	256×192	1.3	0.5	55.2	83.7	62.4	52.3	61.0	62.1
DY-MobileNetV2[12]	256×192	16.1	1.01	68.2	88.4	76.0	65.0	74.7	74.2
DY-ReLU[13]	256×192	9.0	1.03	68.1	88.5	76.2	64.8	74.3	—
UDP[14]	256×192	28.7	7.1	75.2	92.4	82.9	72.0	80.8	80.4
Dite-HRNer-30[5]	256×192	1.8	0.3	68.3	88.2	76.2	65.5	74.1	74.2
MobileNetV2[15]	256×192	9.6	1.4	64.6	87.4	72.3	61.1	71.2	70.7
8-stage Hourglass[16]	256×192	25.1	14.3	66.9	—	—	—	—	—
Lite-HRNet-18[4]	256×192	1.1	0.2	64.8	86.7	73.0	62.1	70.5	71.2
CPN[17]	256×192	27.0	6.2	68.6	—	—	—	—	—
ShuffleNetV2[18]	256×192	7.6	1.2	59.9	85.4	66.3	56.6	66.2	66.4
Dite-HRNet-18[5]	256×192	1.1	0.2	65.9	87.3	74.0	63.2	71.6	72.1
DARK[19]	128×96	63.6	3.6	71.9	89.1	79.6	69.2	78.0	77.9
Lite-HRNet-30[4]	256×192	1.8	0.3	67.2	88.0	75.0	64.3	73.1	73.3
Simple Baseline[20]	256×192	34.0	8.9	70.4	88.6	78.3	67.1	77.2	76.3
ACW-HRNet-18	256×192	1.2	0.2	66.2	89.6	72.8	64.2	69.3	69.6
ACW-HRNet-30	256×192	1.9	0.3	69.0	90.6	76.1	66.7	72.5	72.1
Small HRNet[1]	384×288	1.3	1.2	56.0	83.8	63.0	52.4	62.6	62.6
MobileNetV2[15]	384×288	9.6	3.3	67.3	87.9	74.3	62.8	74.7	72.9
ShuffleNetV2[18]	384×288	7.6	2.8	63.6	86.5	70.5	59.5	70.7	69.7
Lite-HRNet-18[4]	384×288	1.1	0.4	67.6	87.8	75.0	64.5	73.7	73.7
Lite-HRNet-30[4]	384×288	1.8	0.7	70.4	88.7	77.7	67.5	76.3	76.2
Dite-HRNet-18[5]	384×288	1.1	0.4	69.0	88.0	76.0	65.5	75.5	75.0
Dite-HRNet-30[5]	384×288	1.8	0.7	71.5	88.9	78.2	68.2	77.7	77.2
ACW-HRNet-18	384×288	1.2	0.4	69.9	90.6	77.3	67.0	73.9	72.9
ACW-HRNet-30	384×288	1.9	0.7	72.1	91.6	79.8	69.6	75.9	75.1

表 3 COCO 测试集性能对比

模型	输入尺寸	参数量/M	计算量/G	AP/%	$AP^{0.5}$/%	$AP^{0.75}$/%	AP^{M}/%	AP^{L}/%	AR/%
CPN[17]	384×288	—	—	72.1	91.4	80.0	68.7	77.2	78.5
Simple Baseline[20]	256×192	34.0	8.9	70.0	90.9	77.9	66.8	75.8	75.6
HRNet[3]	384×288	28.5	16.0	74.9	92.5	82.8	71.3	80.9	80.1
Small HRNet[1]	256×192	1.3	0.5	55.2	83.7	62.4	52.3	61.0	62.1
OpenPose[21]	—	—	—	61.8	84.9	67.5	57.1	68.2	66.5
Small HRNet[1]	384×288	1.3	1.2	55.2	85.8	61.4	51.7	61.2	61.5
UDP[14]	384×288	28.7	16.1	76.1	92.5	83.5	72.8	82.0	81.3
Dite-HRNet-30[5]	384×288	1.8	0.7	70.6	90.8	78.2	67.4	76.1	76.4
MobileNetV2[15]	384×288	9.8	3.3	66.8	90.0	74.0	62.6	73.3	72.3
Lite-HRNet-18[4]	384×288	1.1	0.4	66.9	89.4	74.4	64.0	72.2	72.6
ShuffleNetV2[18]	384×288	7.6	2.8	62.9	88.5	69.4	58.9	69.3	68.9
Dite-HRNet-18[5]	384×288	1.1	0.4	68.4	89.9	75.8	65.2	73.8	74.4
DARK[19]	384×288	63.6	32.9	76.2	92.5	83.6	72.5	82.4	81.1
Lite-HRNet-30[4]	384×288	1.8	0.7	69.7	90.7	77.5	66.9	75.0	75.4
ACW-HRNet-18	384×288	1.2	0.4	67.4	89.7	74.7	64.1	72.8	73.2
ACW-HRNet-30	384×288	1.9	0.7	69.5	90.3	77.5	66.5	74.9	75.2

为了生动展示模型的检测效果，对部分检测结果进行了可视化，如图 9 所示。可以观察到，本研究的模型在关节点检测的位置上与标注真值的结果相当接近。结果表明在光照良好场景下对关节点的检测效果较好，在光照不足场景下对手腕及脚踝关节点的检测效果不佳；在单人肢体重合存在遮挡的情况下对关节点的检测效果较好；单人背景模糊时对关节点的检测效果同样较好。多人场景下，四肢存在小部分遮挡时关节点能够得到有效的识别，如在部分遮挡场景下可通过冲浪板边缘手部信息、摩托车驾驶员腿部信息以

及周围特征来有效推测遮挡的关节点位置；肢体存在大面积遮挡时，关节点识别效果一般，在多人物场景下对关节点的识别除膝盖关节点以外均较为准确。以上都表明了实验方法的鲁棒性和稳定性并且均适用于大多数主流应用场景。在目标检测、动作识别、医疗康复等领域提供了参考与支持。

图 9　可视化图

4.2.3　MPII 验证集

表 4 显示 ACW-HRNet 与其他主流网络的比较结果。与 Small HRNet-W16 相比，ACW-HRNet-18 的 PCKh@0.5 提高了 6.5，显示出显著的提升。与 MobileNetV3 相比，ACW-HRNet-18 在 PCKh@0.5 上提高了 2.4，但只使用了 21%的计算量。与 Lite-HRNet-18 相比，ACW-HRNet-18 将 PCKh@0.5 的值提高了 0.6，同时保持了同等的模型复杂度。随着模型规模的增加，ACW-HRNet-30 达到了 87.5 的 PCKh@0.5，以 0.5 的增益超过了主流轻量级网络 Lite-HRNet 与 MobileNetV2 和 Shuffle-NetV2 相比，ACW-HRNet 表现出更高的计算效率。尽管与 Simplebaseline、DCLM、HRNet 和 ViTPose 等大型网络相比，实验模型在性能上有所不足，但其运算量和参数规模却显著减少。

表 4　MPII 验证集性能对比

模型	参数量/M	计算量/G	PCKh
SimpleBaseline[20]	68.6	20.9	91.5
DCLM[22]	15.5	15.6	92.3
HRNet-W32[3]	28.5	9.5	92.3
ViTPose-B	86.0	18.0	93.3
MobileNetV2[15]	9.6	1.97	85.4
MobileNetV3[23]	8.7	1.82	84.3
ShuffleNetV2[18]	7.6	1.70	82.8
Small HRNet-W16[1]	1.3	0.72	80.2
Lite-HRNet-18[4]	1.1	0.2	86.1
Lite-HRNet-18[4]	1.8	0.4	87.0
Dite-HRNet-30[5]	1.1	0.2	87.0
Dite-HRNet-30[5]	1.8	0.4	87.6
ACW-HRNet-18	1.2	0.2	86.7
ACW-HRNet-30	1.9	0.4	87.5

4.3　消融实验

在 COCO 验证集上进行了一系列消融实验，输入尺寸为 256×192。

4.3.1　跨维分割卷积和自适应上下文建模

为了进一步验证提出的 CSC 模块和 ACM 方法在 ACW-HRNet-18 中的有效性，对 COCO 验证集进行了消融。最初，使用 Lite-HRNet-18 作为基线，并分别添加网络中的 CSC、ACM 以及 CA 模块。然后将结果与 ACW-HRNet-18 进行比较。表 5 表明，提出的 CSC、ACM 以及 CA 模块为 Lite-HRNet-18 提供了显着的增益，同时只增加了少量的参数量和计算负载。

表 5　COCO 验证集上的消融实验

模型	参数量/M	计算量/M	*AP*
Lite-HRNet-18[4]	1.1	205.2	64.8
With CSC	1.2	227.8	65.5
With ACM	1.09	209.3	65.3
With CA	1.1	212.3	65.8
ACW-HRNet-18	1.21	238.5	66.2

4.3.2　跨维分割卷积中的超参数

为确定超参数在 CSC 中的最佳配置，在 COCO 验证集上进行了四组消融实验。表 6 显示了 ACW-HRNet-18 在各种设置下的结果，其中每行标题中的数字代表从最高分辨率到最低分辨率分支。如在 ACW-HRNet-18 的每个分支中，参数计数和计算负载也会增加。设置设置为 1、1、2 和 4，将其应用于每个水平分支，以实现 CSC 的最佳计算效率。

表 6　不同超参数下的消融实验

超参数	参数量/M	计算量/M	*AP*
1 1 1 1	1.14	229.3	65.5
1 1 2 4	1.21	238.5	66.2
1 2 2 4	1.23	239.8	66.2
1 2 4 4	1.25	241.5	66.5

5　结束语

针对高分辨率网络中网络模块固定、跨维信息交互不足的问题，引入了 ACW-HRNet。引入的 CSC 和 ACM 可以有效地提取多尺度空间信息，并提供充足的跨维信息交换。引入的 CA 可以捕获通道间依赖关系和空间上下文，使网络能够准确定位人体的关键点。由于 CSC 、ACM 以及 CA 的有效性，自适应跨维加权网络在 COCO、MPII 数据集上取得了优异的成绩，优于当

前主流的轻量级网络。随后，考虑将提议的 ACW 模块扩展到其他网络，并进一步优化其性能。

参考文献：

[1] WANG J D, SUN K, CHENG T H, et al. Deep High-Resolution representation learning for visual recognition [C] // IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE, 2021: 3349 - 3364.

[2] PIOTR D, ROSS G, HE K, et al. Feature pyramid networks for object detection [C] // IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Honolulu, 2017: 936 - 944.

[3] SUN K, XIAO B, LIU D, et al. Deep High-Resolution representation learning for human pose estimation [C] // IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Long Beach, 2019: 5686 - 5696.

[4] YU C Q, XIAO B, GAO C X, et al. Lite-HRNet: a lightweight High-Resolution network [C] // IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Nashville, 2021: 10435 - 10445.

[5] ZHANG H, DUN Y J, PEI Y X, et al. HF-HRNet: a simple hardware friendly High-Resolution network [C] // IEEE Transactions on Circuits and Systems for Video Technology. IEEE, 2024: 7699 - 7711.

[6] WANG B, WANG G. A lightweight model for road sign detection based on Improved-YOLOv8 [C] // 2024 9th International Conference on Electronic Technology and Information Science (ICETIS), IEEE, Hangzhou, 2024: 319 - 322.

[7] CAO Y, XU J R, STEPHEN L, et al. GCNet: Non-Local networks meet Squeeze-Excitation networks and beyond [C] // IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), IEEE, Seoul, 2019: 1971 - 1980.

[8] WANG X L, ROSS G, ABHINAV G, et al. Non-local neural networks [C] // IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Salt Lake City, 2018: 7794 - 7803.

[9] HU J, SHEN L, SUN G. Squeeze-and-Excitation networks [C] // IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Salt Lake City , 2018: 7132 - 7141.

[10] SANGGYUN W, JONGCHAN P, JOON-YOUNG L, et al. CBAM: convolutional block attention module [C] // Proceedings of the European Conference on Computer Vision (ECCV), 2018: 3 - 19.

[11] BOWEN C, XIAO B, WANG J D, et al. HigherHRNet: Scale-Aware representation learning for Bottom-Up human pose estimation [C] // IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Seattle, 2020: 5385 - 5394.

[12] ZHAO Z, DONG M. Channel-Spatial dynamic convolution: an exquisite Omni-dimensional dynamic convolution [C] // 8th International Conference on Intelligent Computing and Signal Processing (ICSP), Xi'an, 2023: 1707 - 1712.

[13] CHEN Y P, DAI X Y, LIU M C, et al. Dynamic ReLU [C] // Proceedings of the European Conference on Computer Vision (ECCV), 2020: 351 - 367.

[14] HUANG J J, ZHU Z, GUO F, et al. The devil is in the details: delving into unbiased data processing for human pose estimation [C] // IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Seattle, 2020: 5699 - 5708.

[15] MARK S, ANDREW H, ZHU M L, et al. MobileNetV2: inverted residuals and linear bottlenecks [C] // IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Salt Lake City, 2018: 4510 - 4520.

[16] ALEJANDRO N, YANG K Y, DENG J. Stacked hourglass networks for human pose estimation [C] // Proceedings of the European Conference on Computer Vision (ECCV), 2016: 483 - 499.

[17] CHEN Y L, WANG Z C, PENG Y X, et al. Cascaded pyramid network for Multi-person pose estimation [C] // IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Salt Lake City, 2018: 7103 - 7112.

[18] MA N N, ZHANG X Y, ZHENG H T, et al. ShuffleNet V2: practical guidelines for efficient CNN architecture design [C] // Proceedings of the European Conference on Computer Vision (ECCV), 2018: 116 - 131.

[19] ZHANG F, ZHU X T, DAI H B, et al. Distribution-Aware coordinate representation for human pose estimation [C] // EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Seattle, 2020: 7091 - 7100.

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

[21] CAO Z, TOMAS S, SGIH-EN W, et al. Realtime multi-person 2D pose estimation using part affinity fields [C] // IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Honolulu, 2017: 1302 - 1310.

[22] 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.

[23] ANDREW H, MARK S, CHEN B, et al. Searching for mobileNetV3 [C] // IEEE International Conference on Computer Vision (ICCV), IEEE, Seoul, 2019: 1314 - 1324.