基于双重注意力及边缘约束的人体解析方法

刘俊婧，郑宛露，王少荣
北京林业大学信息学院 北京 100083

摘要：单人人体解析任务假设图像中仅包含一个人类实例，通过人体部位之间的关系建模提取人体特征。由高分辨率保持方法和上下文信息嵌入模块构成的基础人体解析模型缺乏对全局依赖关系的建模，在类别信息的区分上也存在不足。一方面，为了解决卷积操作导致的局部感受野的问题，本文引入了双重注意力模块，该模块通过在空间和通道维度上对语义依赖进行建模，以自适应地整合局部特征及远程下上文信息；另一方面，为了增强模型区分相邻部分的能力，提高不同类别边界的解析准确度，本文在人体解析分支之外引入了辅助性的边缘约束分支，为人体解析分支提供各部位的边缘信息。实验结果表明，基于双重注意力及边缘约束的人体解析模型在 LIP 上能够实现近 55%的 mIoU，在 LIP_B 上能够实现近 85%的 mIoU。

关键词：人体解析；注意力机制；边缘检测

Human Parsing Method Based on Dual Attention and Edge Constraints

Liu JunJing, Zheng Wanlu, Wang Shaorong
School of Information Science and Technology of Beijing Forest University, Beijing 100083

Abstract: The single person human parsing task assumes that the image only contains one human instance, and extracts human features through modeling the relationships between human parts. The basic human parsing model composed of high-resolution preservation methods and contextual information embedding modules lacks modeling of global dependencies, and also has shortcomings in distinguishing category information. On the one hand, in order to solve the issues of local receptive fields caused by convolution operations, this paper introduces a dual attention module, which models semantic dependencies in spatial and channel dimensions to adaptively integrate local features and remote contextual information; On the other hand, in order to enhance the model's ability to distinguish adjacent human parts and improve the accuracy of boundary analysis for different categories, this paper introduces auxiliary edge constraint branches outside the human parsing branch, providing edge information for each part of the human parsing branch. The experimental results show that the human parsing model based on dual attention and edge constraints can achieve nearly 55% mIoU on LIP, and nearly 85% mIoU on LIP_B.

Keywords: Human Parsing; Attention Mechanism; Edge Detection

1 引言

人体解析属于像素级的细粒度分割任务，旨在将图像或视频中的人划分为多个像素级的语义部分（例如头部、胳膊等人体部位和服装等）。与一般的分割任务相比，人体解析任务的挑战在于对一个人的每个身体部位进行精细区分，这大大增加了解析的难度。

深度学习的快速发展带动人体解析算法的显著进步。目前已经开发了多种方法[1-6]来实现像素级的人体解析，基于高分辨率保持的方法和基于上下文信息嵌入的方法是两种主流方法[7]：

1）高分辨率保持的方法。由于深度神经网络连续的空间池化和卷积操作，使得特征图的分辨率显著降低，从而导致细节信息的丢失。为了解决这一问题，一些方法尝试通过获取高分辨率的特征来恢复所需要的细节信息。最直接的方法是减少下采样操作[8]，例如使用较小的步幅或减少池化操作的使用，通过限制特征图的空间下采样程度，在一定程度上保持较高的分辨率。这种方法能够保留细微的空间细节，但也可能导致计算资源和内存消耗的增加。还有的方法[3]在编码器中捕获高级语义信息，然后在解码器中则通过融合低级特征图来恢复细节和空间信息。这种方法的优势在于可以充分利用深度神经网络的表达能力和语义理解能力，同时保持较高的分辨率，因此在人体解析任务得到了广泛的使用。

2）上下文信息嵌入的方法。得益于 CNN 架构和训练技术的发展，上下文感知的方法受到越来越多的关注，这些方法通过将更多的环境信息引入模型，使其能够更好地理解当前输入的内容，并做出更准确的预测和推理，提高了人体解析的性能。特征金字塔是丰富上下文信息的有效方法之一，其通过构建不同尺度的特征图来捕获物体在不同尺度上的信息。金字塔池化模块（PPM）[9] 通过多比例融合的方式建立不同尺度的池化模块捕获全局信息，最低比例的映射是全局尺度的池化操作，高比例的映射则将图像分为不同的子区域，形成不同区域的局部信息表示。全局平均池化（GAP）也被广泛用于获取全局信息，实现更加可靠的预测。

基于上述两种方法，人体解析网络的基础框架如图 1 所示。输入图片首先经过特征提取网络进行初步的特征提取；然后将网络最后一层的输出馈送到金字塔模块中获得多尺度的上下文信息；最后，将网络的低层特征通过高分辨率保持模块与多尺度信息进行融合，获得最终的解析特征。

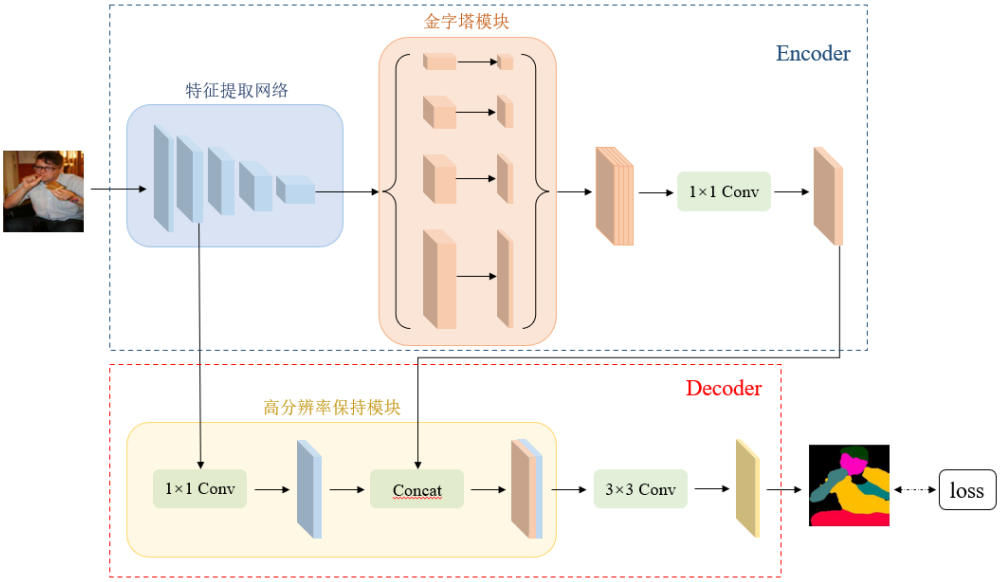

图 1 人体解析网络基础框架

人体解析任务的核心问题是如何对人体结构进行建模。人体拥有高度结构化的层次特

征，各个部分自然地相互作用。大多数解析器都希望显式或隐式地构建这种依赖关系。然而现实情况中存在着类内差异大、类间差异小的问题，例如，不同的上衣的颜色、纹理和形状都会严重影响其外观，导致显著的类内差异，而与皮肤颜色相近的衣服则会使得类间差异较小。这种不协调的类内和类间变化会增加分类器学习决策边界的难度，导致预测中的语义不一致。虽然金字塔模块能够获得多尺度的上下文信息，但其忽略了对远程类间上下文关系的建模。

为了将类别信息嵌入在上下文信息中，本文在金字塔模块之后引入了双重注意力模块（Dual Attention Module，DAM），该模块能够自适应地整合局部特征，其在空间和通道两个维度上分别对位置特征和通道特征进行加权建模，构建全局依赖关系。

除此之外，虽然基于上下文的人体解析方法可以获得丰富的语义和细节信息，但随着网络的加深，这种方法可能会伴随信息衰减。为了补充细节信息，有的方法[10-13]通过引入辅助信息来帮助网络更好地理解部分之间的关系，大大提高了特征的表达能力和人体分析的性能。边缘感知学习可以增强模型区分相邻部分的能力，提高部分边界的精细度。本文采用多任务处理的思想，通过构建双分支结构模型来引入边缘感知监督，为人体解析分支提供辅助信息，强化对类内和类间信息的区分。该分支内还嵌入了一个尺度感知激励模块（Scale-aware Excitation Module，SEM），用于对边缘分支的输入特征进行初步的处理，该模块能够充分利用输入特征丰富的尺寸信息以及细节和语义之间的层次信息。最后将两个分支提取的特征进行融合，促进实现人体的精确解析。

综上所述，本文的主要贡献有以下几点：

1）引入了双重注意力模块 DAM，该模块通过聚合空间维度和通道维度的远程上下文信息，既强调了特征图中的远程依赖关系，又强调了与类别有关的语义依赖关系，实现了特征图的动态更新。

2）增加了边缘检测分支，该分支包含了一个尺度感知激励模块 SEM，促使解码器获得更加丰富的初始特征信息，并将挖掘到的边缘信息作为人体解析分支的辅助特征。

3）实验证明本文提出的基于双重注意力及边缘约束的人体解析网络中各模块的有效性，并在 LIP 数据集上取得了优异的性能。

2 相关工作

2.1 语义分割

人体解析是一项细粒度的语义分割任务，因此其使用的方法与语义分割任务中的相似。全卷积神经 FCN[14]对整个图像进行完全卷积，以生成每个像素的标签。一些基于 FCN 的方法[4, 5, 15-20]取得了很好的性能。然而，受到卷积层结构的限制，FCN 提供的上下文信息不足，留下了改进的空间。许多研究人员开始利用编码器-解码器结构，通过下采样提取特征，然后使用上采样将其恢复到原始分辨率。为了扩大感受野，另一种结构 DeepLab[21]设计了空洞卷积，迫使网络感知更大的区域，减少预测误差。Xia 等[22]提出了分层自动收缩网络（Hierarchical Auto-Zoom Net，HAZN），HAZN 可以自适应地将预测的图像区域缩放到适当的尺度以改进解析结果。PCNet[23]进一步研究了自适应上下文特征，并通过提出的部位类模块。关系聚合模块和关系分散模块挖掘人体部位的相关语义，进一步捕获具有代表性的全局上下文。

2.2 人体解析

为了完成像素级的分割，人体解析任务需要丰富的特征表达。结合人体检测、人体姿态估计和人体边缘等任务进行辅助监督，可以大大提高特征的表达能力。Liu等[24]将人体姿

态估计模块与基于MRF的颜色类别推断模块和超像素类别分类模块相结合，以解析图像中的时尚单品。与此同时。Gong等[13]开发了一种新颖的自监督结构敏感学习方法，该方法通过计算解析图中相应区域的中心点，生成关节点和关节点标签，且模型在训练中不需要特别标记人体关节。除此之外，考虑到在图像中，低频区域是语义相似的区域，高频区域通常是语义转换比较大的区域。因此，利用人体边缘来描述人体前景和背景变换的区域，具有区域分异性。CE2P[12]是典型的带有边缘感知的上下文嵌入框架，其利用边缘感知框架生成人体轮廓特征，细化人体部位边界。由于其出色的性能和可扩展性，该工作已成为后续许多工作的基石。Zhang等[25]提出了关联解析机（CorrPM），研究人体语义边界和关键点位置如何共同提高网络解析能力。CDGNet[26]采用水平方向和垂直方向累计的人体解析标签作为监督，学习人体部位的位置分布，并通过注意力机制将其加权到全局特征中，实现了精确的部位关系建模。

3 方法

3.1 网络概述

如图 2 所示为双重注意力及边缘约束（Dual Attention and Edge Constraints，DAEC）的人体解析网络的总体框架，该框架除基线网络中所包含的特征提取网络、金字塔池化模块、以及高分辨率保持结构外，还引入了双重注意力模块（Dual Attention Module，DAM）以及边缘约束分支（Edge Constraint Branch，ECB）来帮助网络实现更加精细的人体解析。其中，本文使用 ResNet101 作为特征提取网络。

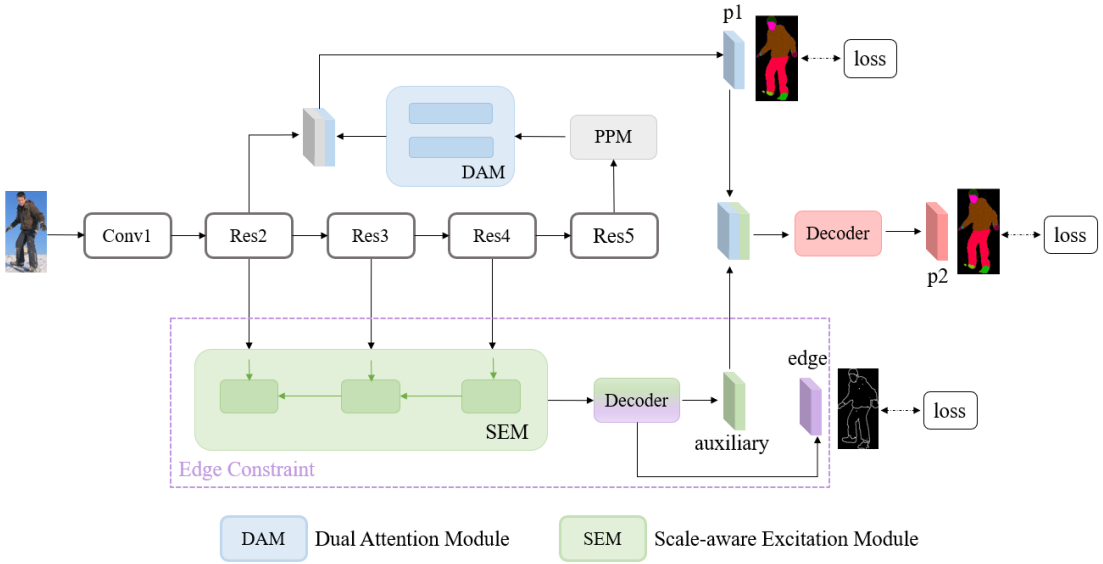

图 2 基于双重注意力及边缘约束的人体解析模型

具体来说，首先利用特征提取网络对给定的输入图像进行初步的特征提取，得到每个阶段的特征Res_i, $i \in \{1,2,3,4,5\}$。随后，人体解析分支将 Res5 输入到金字塔池化模块 PPM 中，通过多比例的特征融合提取多尺度的特征信息。其输出被送入到双重注意力模块 DAM 中进一步通过空间和通道维度的注意力编码进行动态更新，然后与来自Res_2的高分辨率特征融合，得到粗解析特征p_1。与此同时，边缘约束分支首先将Res_2、Res_3、Res_4馈送到尺度感知激励模块（Scale-aware Excitation Module，SEM）中，其输出一方面通过边缘解码作为边缘预测特征$edge$；另一方面，通过解码得到一个新特征auxiliary，其作为辅助特征与粗解

析特征p_1融合，然后解码作为最终的人体解析特征p_2。

3.2 双重注意力模块

PPM 采用分层的策略，将原始行人特征变换成四个不同比例的新特征，最低比例的映射是全局尺度的池化，高比例的映射则将图像分为不同的子区域，形成不同区域的信息表示。考虑到网络深度的增加可能会带来一些信息的丢失，因此将 ResNet101 的第一个残差块的特征Res_2与 PPM 特征融合，来补充高分辨率的细节信息。

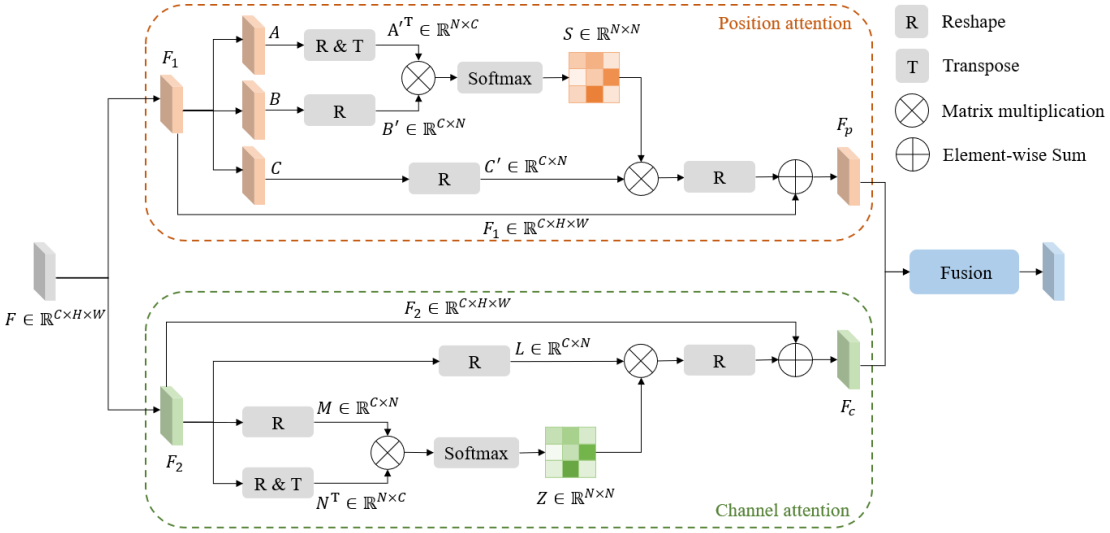

图 3 双重注意力模块

尽管经过金字塔模块后，模型已经获得了不同尺度的全局上下文信息，但由于卷积操作的感受野是局部性的，其仅能捕获局部区域特征的语义相关性，缺少对于远程的语义依赖的建模，导致具有相同标签的像素其对应的特征可能仍会存在一些差异。为了解决这个问题，本文在 PPM 之后引入了双重注意力模块 DAM 进一步处理。如图 3 所示，该模块由位置注意力模块和通道注意力模块通过并联的方式组合而成，输入特征F会复制成两份分别输入到两个模块中。

位置注意力模块致力于捕获和利用图像中不同空间位置之间的关系，增强网络对空间关系的理解能力。位置注意力模块的第一步是生成空间注意矩阵，给定一个维度大小为$\mathbb{R}^{C\times H\times W}$的输入特征$F_1$，将其通过三个分支进行处理，一个分支用于生成新的特征映射A，一个分支用于生成新的特征映射B，最后一个分支用于生成特征映射C，它们的维度均为$\mathbb{R}^{C\times H\times W}$。为方便计算空间注意力图，进一步将$A$的维度重塑为$\mathbb{R}^{C\times N}$得到$A'$，将$B$的维度重塑为 $\mathbb{R}^{C\times N}$得到B'，其中$N=H\times W$。最后应用 softmax 层对A'的转置和B'乘积进行编码，得到空间注意力矩阵 $S\in\ \mathbb{R}^{N\times N}$，其计算过程如公式 3-1 所示：

$$S = softmax(A'^{T}\cdot B) \tag{3-1}$$

S对特征的任意两个像素之间的空间关系进行建模，其中的元素s_{ji}表示位置i对位置j的影响，s_{ji}越大，表示两者的相关性越大。另外，将C重塑为$\mathbb{R}^{C\times N}$得到C'，然后将C'与注意力矩阵S进行矩阵相乘，并将输出特征重塑为 $\mathbb{R}^{C\times H\times W}$。最后将其与原始特征$F_1$执行逐元素求和，得到最后的输出特征$F_p\in\mathbb{R}^{C\times H\times W}$。$F_p$对特征图中所有空间位置进行了加权，其计算过程如公式 3-2 所示：

$$F_p = Reshape(\alpha\ (S\cdot C')) + F_1 \tag{3-2}$$

其中，α 初始化为 0，逐渐学习更多的权重。空间注意力模块使得相似的远程特征实现了相互增强，提高特征图的类内的紧凑性和语义一致性。与此同时，通道注意模块旨在提取通道维度中的上下文信息，特征中的每一个通道代表了对某一类别的相应，通道注意力矩阵揭示了不同通道间的依赖关系，通过增强这些相互依赖的通道，可以优化针对特定语义的特征表达。

通道注意模块的处理过程与位置注意模块类似，只不过是基于原特征直接计算。具体来说，输入特征F_2同样需要通过三个分支分别进行处理，第一个分支和第二个分支将F_2重塑为$\mathbb{R}^{C\times N}$分别得到特征M和N，然后在M与N的转置之间执行矩阵乘法。最后应用 softmax 层对其乘积进行编码，得到通道注意力矩阵$Z \in \mathbb{R}^{C\times C}$，其计算过程如公式 3-3 所示：

$$Z = softmax(N^T \cdot M) \tag{3-3}$$

Z对特征的任意两个像素之间的通道关系进行建模，其中的元素z_{ji} 表示第i个通道对第j个通道的影响，z_{ji}越大，表示两者的相关性越大。另外，第三个分支将F_2重塑为$\mathbb{R}^{C\times N}$得到特征L，然后在L与通道注意矩阵Z之间执行矩阵乘法，并将其输出重塑为$\mathbb{R}^{C\times H\times W}$。最后同样将其与原始特征$F_2$执行逐元素求和，得到最终的输出特征$F_c \in \mathbb{R}^{C\times H\times W}$。$F_c$对特征图中所有通道进行了加权，其计算过程如公式 3-4 所示：

$$F_c = Reshape(\beta\ (Z \cdot L)) + F_2 \tag{3-4}$$

其中，β 初始化为 0，逐渐学习更多的权重。

最后，将两个注意力模块的输出聚合起来，这样双重注意模块既强调了特征图中的远程依赖关系，又强调了与类别有关的语义依赖关系，获得了更好的特征表示。

3.3 融合尺度感知注意模块的边缘约束分支

图像中的低频区域是语义相似的区域，高频区域则通常是语义转换较大的区域。边缘提取网络正是利用了这一特点，通过分析图像中像素之间的变化来检测图像中对象的边缘部分。传统的边缘检测算法通过计算像素之间的梯度来检测像素的变化。进一步，基于学习的边缘检测器结合颜色、亮度和其他线索进行特征表示。然而这些算法在面对特定的任务时仍然需要设计针对性的特征提取策略，泛化性较低。为了解决这一问题，基于深度学习的方法通常使用卷积核来检测图像中的边缘，这些卷积核通常是一些特定的滤波器，可以通过与图像进行卷积操作突出边缘信息。

卷积神经网络旨在获取输入图像的初级特征表示，其中，浅层特征通常包含更多的低级线索，例如边缘和颜色，深层特征则通常包含更多的语义信息，如对象类别。因此，本文选取特征提取网络中的Res_2、Res_3和Res_4特征作为边缘约束分支的输入。边缘约束分支如图 4 所示。

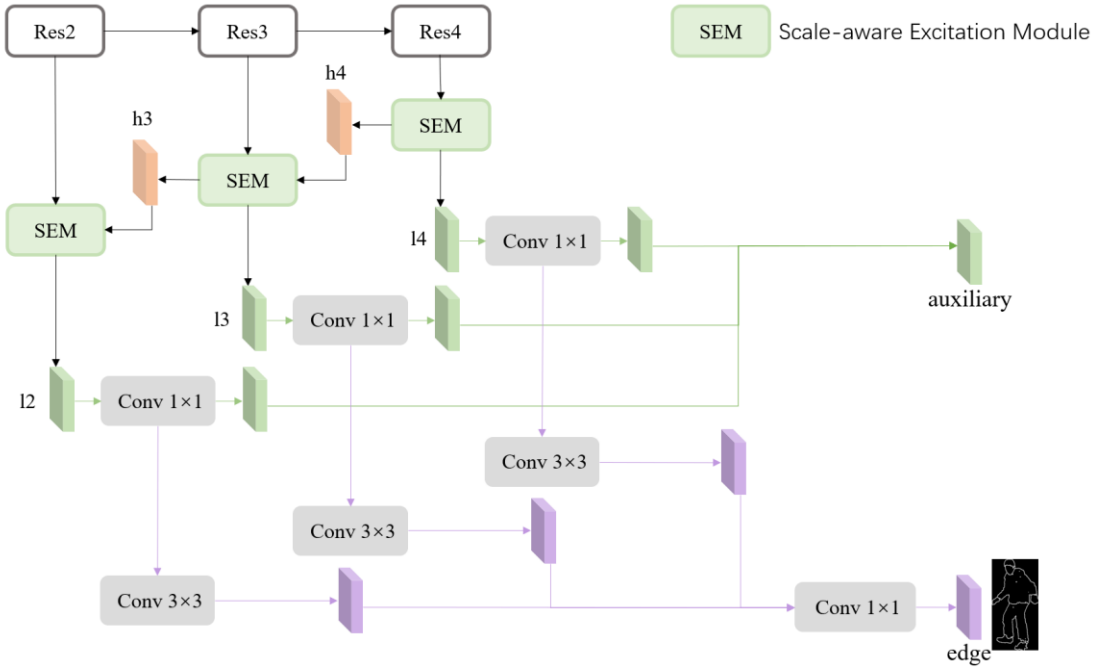

图 4 边缘约束分支

考虑到在不同大小的图像上使用相同卷积核的空洞卷积可以得到不同的局部特征，将这些局部特征进行叠加就可以得到完整的特征表达。基于此，本文引入了包含空洞卷积的尺度感知激励模块 SEM，其结构如图 5 所示。

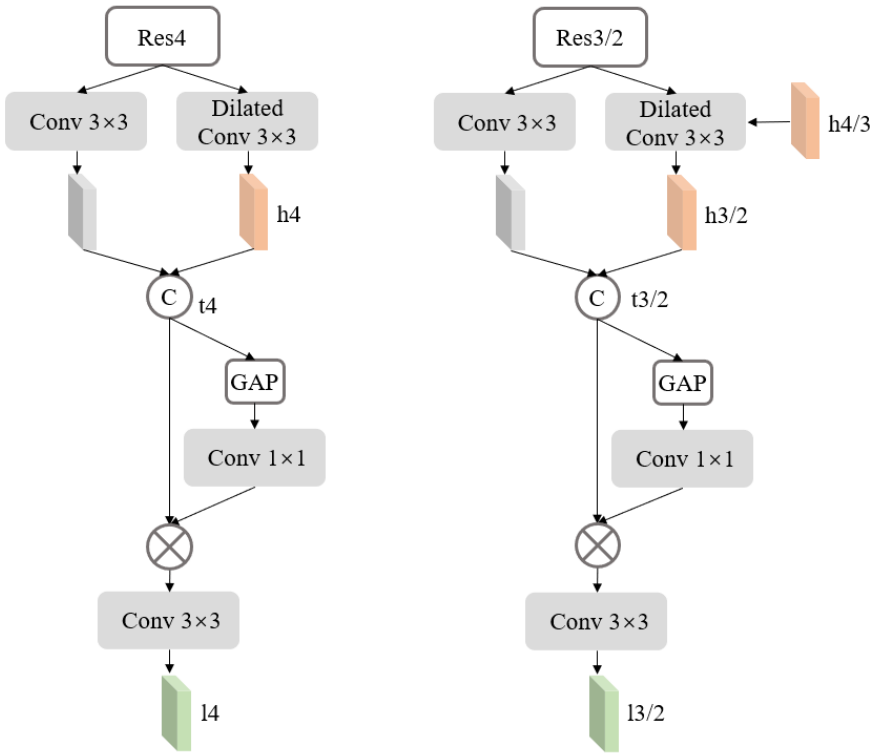

图 5 尺度感知注意模块

图中左边部分是Res_4特征对应的 SEM 模块，右边部分是Res_2和Res_3对应的 SEM 模块。具体来说，给定输入特征 $Res_i \in \mathbb{R}^{C\times H\times W}$，首先采用 3×3 的空洞卷积进行特征提取，如公

式 3-5 所示：

$$h_i = \begin{cases} Dconv(Res_i),\ i = 4 \\ Dconv\big(Res_i + up(h_{i+1})\big), i = 2,3 \end{cases} \tag{3-5}$$

其中$Dconv$表示空洞卷积、批归一化和 ReLU 激活函数的组合。同时Res_i还经过了一个并行的 3×3 普通卷积，将其与空洞卷积得到的特征h_i执行拼接操作，如公式 3-6 所示：

$$t_i = Concat(h_i, Conv(Res_i)),\ i = 2,3,4 \tag{3-6}$$

之后分为两个并行的分支，其中一个分支先经过全局平均池化后，实现了从空间维度对特征进行了压缩，将特征维度从$\mathbb{R}^{C\times H\times W}$压缩为$\mathbb{R}^{C\times 1\times 1}$，得到的向量在某种程度上具有全局性的感受野。然后在卷积部分之前引入了一个可学习的ω权重参数来学习每个通道的权重，最后通过卷积部分进行一个降维，最后其输出是包含通道重要性的特征，再与先前的特征进行一个融合，得到最终的输出。计算过程如公式 3-7 所示：

$$l_i = Conv^{3\times 3}(Conv^{1\times 1}(\omega \cdot GAP(t_i)) \cdot t_i),\ i = 2,3,4 \tag{3-7}$$

最后，将 SAM 模块得到的 l_i 特征首先通过 1×1 的卷积改变其通道数，这里得到的三个特征一方面直接融合作为辅助特征与人体解析分支融合，另一方面，解码并融合成为边缘预测特征$edge$。

4 实验

4.1 数据集

Look Into Person (LIP)：LIP数据集[12]在2016年LIP挑战赛中用于人工解析任务，其中的图像均裁剪自COCO数据集中的人物实例，覆盖了具有挑战性的视角、姿势，以及严重的遮挡、多样化的外观、不一致的分辨率等真实场景的共50,462张注释图像，其中的30,462张图像用于训练，10,000张图像用于验证，10,000张图像用于测试，数据量大且具有挑战性。这些图像在像素级别上被精细地标记为19个语义人体部位类别(包括6个身体部位和13件衣服)和一个背景类别。

LIP_B：我们将 LIP 数据集中所有真值图片进行了二值转换，将所有前景的像素值赋值为 1，所有背景的像素值赋值为 0，由此将 LIP 变成了一个二类别的数据集，并将其记为 LIP_B。

4.2 评价指标

对于人体解析算法的评价方法可以分为定性评估和定量评估。定性评估通过将解析结果可视化，并与其真值及其他方法的可视化结果进行对比，可以直观地看出算法实现的解析效果。定量评估则以数值的形式来呈现，人体解析任务常用的评价指标包括像素准确率PA、平均像素准确率mPA和平均交并比mIoU，其中以平均交并比作为主要的评价指标，另外两个作为参考指标。

4.3 参数设置

我们采用ResNet-101作为骨干网络，并在256×128、256×256以及384×384三种输入大小下对模型进行训练。其他参数设置如表1所示

表 1 实验参数设置

参数变量	参数值
训练次数（epoch）	150

批大小（batch_size）	8
初始学习率（lr）	0.001
学习衰减率（lr_decay）	0.9
优化函数（ptimizer）	SGD
动量大小（momentum）	0.9
权重衰减系数（weight_decay）	0.0005
随机翻转（flip_prob）	0.5
随机缩放（scale_factor）	0.25
随机旋转（rotation_factor）	30

4.4 对比实验

将提出的双重注意力及边缘约束的人体解析方法在 LIP 数据集上进行训练和测试，并与其他现有的人体解析方法进行对比。表 2 展示了以 384×384 输入大小为例，本文方法与其他方法在 LIP 数据集的各个类别上的 IoU 对比情况。本文的方法在大多数的类别上可以取得较好的解析效果。CE2P 方法采用特征金字塔结构获取上下文信息，并使用边缘分支提供辅助信息，其模型的解析效果较之前的方法有明显的提升。相较于 CE2P，本文的方法在帽子、手套、围巾这些小物体类别上分别实现了 1.73、5.56、8.93 个百分点的性能提升，除此之外，在连衣裙、外套、裤子、连体裤、裙子等类别上分别提升了 5.03、0.87、1.31、2.55、8.53 个百分点的提升。

表 2 与其他方法在 LIP 各个类别上的 IoU 比较

	SS-NAN[27]	**MMAN**[28]	**JJPNet**[29]	**CE2P**[12]	**DAEC(ours)**
Background	88.67	84.75	86.26	87.67	88.16
hat	63.86	57.66	63.55	65.29	67.02
Hair	70.12	65.63	70.20	72.54	72.59
Glove	30.63	30.07	36.16	39.09	44.65
Sunglasses	23.92	20.02	23.48	32.73	33.31
u-clothes	70.27	64.15	68.15	69.46	69.62
dress	33.51	28.39	31.42	32.52	37.55
coat	56.75	51.98	55.65	56.28	57.15
socks	40.18	41.46	44.56	49.67	47.26
pants	72.19	71.03	72.19	74.11	75.42
j-suits	27.68	23.61	28.39	27.23	29.78
scarf	16.98	9.65	18.76	14.19	23.12
skirt	26.41	23.20	25.14	22.51	31.04
face	75.33	69.54	73.36	75.50	75.60
l-arm	55.24	55.30	61.97	65.14	64.74

r-arm	58.93	58.13	63.88	66.59	67.29
l-leg	44.01	51.90	58.21	60.10	57.41
r-leg	41.87	52.17	57.99	58.59	58.17
l-shoe	29.15	38.58	44.02	46.63	46.65
r-shoe	32.64	39.05	44.09	46.12	48.10
mIoU	47.92	46.81	51.37	53.10	54.73

表 3 展示了本文方法与其他方法在 LIP 数据集上的实验结果对比情况。使用 256×128 作为输入大小时，本文的方法取得了 51.33%的 mIoU，与语义分割领域的方法 DeeplabV2 和 Attention 相比效果更好，因为人体解析方法能够关注到更加细粒度的信息。MuLA 和 JPPNet 方法将人体结构信息和人体解析任务结合，证明了人体姿态估计任务可以帮助提升人体解析模型的性能。但本文的方法较两者有明显提升，当采用 384×384 大小时，本文的方法在 mIoU 较 MuLA 和 JPPNet 提高了 5.43 和 3.36 个百分点。即使采用较小的 256×256 作为输入大小时，本文的方法在 mIoU 上较 MuLA 和 JPPNet 仍分别提高了 4.26 和 2.19 个百分点。与 CE2P 相比，在同等输入大小下，本文的方法在 PA、mPA 和 mIoU 上分别提高了 0.25、4.01 以及 1.63 个百分点，进一步证明了本文方法的有效性。

表 3 与其他方法在 LIP 上的比较

Method	**Input_size**	**PA**	**mPA**	**mIoU**
DeeplabV2[30]	-	82.66	51.64	41.46
Attention[2]	-	83.43	54.39	42.92
SS-NAN[27]	321×321	87.59	56.03	47.92
MMAN[28]	256×256	85.24	57.60	46.93
MuLA[31]	256×256	88.50	60.50	49.30
JPPNet[29]	384×384	86.48	62.25	51.37
CE2P[12]	384×384	87.37	63.20	53.10
PGECNet[32]	473×473	87.50	65.66	54.30
ECHP[33]	384×384	87.66	66.09	54.50
DAEC(ours)	256×128	86.41	63,43	51.33
	256×256	87.34	65.36	53.56
	384×384	87.62	67.21	54.73

4.5 消融实验

本文在 LIP 以及 LIP_B 数据集上对所提模块进行了组合训练，以验证其有效性。其中，Baseline 由骨干网络、金字塔模块和高分辨率保持模块组成。DAB 表示双重注意力模块，

ECB 则表示边缘约束分支。

LIP 表 4 展示了以 384×384 输入大小为例，本文方法在 LIP 数据集的不同类别上的 IoU 实验结果。根据表中的结果可以看出，DAB 和 ECB 的加入均能提高模型的分割准确性。具体来说，加入 ECB 的模型在裙子、短裤、外套等类别上实现了更好的分割效果。同时，加入 DAB 的模型在太阳镜、袜子、围巾以及左右腿、左右鞋上等多数类别上的实现了更好的分割效果，说明 DAB 对模型性能的提升作用更大。除此之外，加入 ECB 的模型在背景类别上的 IoU 明显提高，证明了边缘约束分支在区分类间像素上有着明显的优势。

表 4 不同模块在 LIP 各个类别上的 IoU 比较

	Baseline	**+DAB**	**+ECB**	**DAEC(ours)**
Background	87.83	87.97	88.18	88.16
hat	66.40	66.80	67.76	67.02
Hair	71.75	72.05	72.27	72.59
Glove	42.52	43.08	43.18	44.65
Sunglasses	30.25	31.23	29.82	33.31
u-clothes	69.46	69.93	70.20	69.62
dress	37.03	37.58	37.63	37.55
coat	56.03	57.01	57.26	57.15
socks	48.06	48.68	48.66	47.26
pants	75.04	75.40	75.57	75.42
j-suits	32.50	32.44	32.35	29.78
scarf	18.67	20.03	19.30	23.12
skirt	27.70	26.28	28.51	31.04
face	74.66	74.96	75.40	75.60
l-arm	64.86	65.47	65.52	64.74
r-arm	67.47	67.84	68.23	67.29
l-leg	58.11	59.16	58.44	57.41
r-leg	57.80	58.04	57.82	58.17
l-shoe	46.47	46.38	46.21	46.65
r-shoe	47.36	48.07	47.34	48.10
mIoU	54.00	54.42	54.48	54.73

表 5 展示了分别使用 256×128、256×256 和 384×384 作为输入大小，本文方法在 LIP 数据集上的实验效果。在三种输入大小下，Baseline 分别达到了 45.06%、52.84%以及 54.00%的 mIoU，即随着输入大小的增大，Baseline 的效果在不断提高。在 Baseline 的基础上引入双重注意力模块后，mIoU 分别提高了 3.98、0.12 以及 0.42 个百分点；在 Baseline 的基础上引入边缘约束分支后，mIoU 分别提高了 3.82、0.29 以及 0.48 个百分点。将两者均添加到基础模型上后，mIoU 分别实现了 6.27、0.72 以及 0.73 个百分点的提升。

表 5 不同模块在 LIP 上的比较

Input_size	**method**	**PA**	**mPA**	**mIoU**
256×128	Baseline	84.32	55.75	45.06

	+DAB	85.63	60.81	49.04
	+ECB	85.84	60.22	48.88
	DAEC(ours)	86.41	63,43	51.33
256×256	Baseline	87.28	64.40	52.84
	+DAB	87.23	64.61	52.96
	+ECB	87.25	64.98	53.13
	DAEC(ours)	87.34	65.36	53.56
384×384	Baseline	87.43	65.85	54.00
	+DAB	87.65	66.49	54.42
	+ECB	87.77	65.81	54.48
	DAEC(ours)	87.62	67.21	54.73

LIP_B 表 6 展示了以 256×128 输入大小为例，本章方法在 LIP_B 数据集的不同类别上的 IoU 实验结果。对于前景和背景的区分上，加入 ECB 的模型较加入 DAB 的模型有明显优势，大大提高了类间的可区分性。综合来说，本章的方法能够在二分类的分割任务上实现正确率接近 85%的分割效果。

表 7 展示了采用 256×128 的输入大小，在 LIP_B 数据集上验证模型各个模块的有效性实验。实验结果显示，本章的方法较 Baseline 在 mIoU 上提高了 0.76 个百分点，其中 DAB 和 ECB 分别提高了 0.35 和 0.56 个百分点。

表 6 不同模块在 LIP_B 各个类别上的 IoU 比较

	Baseline	**+DAB**	**+ECB**	**DAEC(ours)**
Background	86.17	86.50	91.83	86.77
Froeground	81.92	82.31	91.75	82.70
mIoU	84.05	84.40	84.61	84.74

表 7 不同模块在 LIP_B 上的比较

Input_size	**method**	**PA**	**mPA**	**mIoU**
256×128	Baseline	91.50	91.45	84.05
	+DAB	91.71	91.65	84.40
	+ECB	91.82	91.75	84.61
	DAEC(ours)	91.94	91.89	84.81

4.6 可视化结果

图 6 和图 7 分别展示了基于 LIP 数据集以及基于 LIP_B 数据集的可视化结果示例，其中 LIP 数据集共包含 20 个类别，LIP_B 数据集共包含 2 个类别。图中从左到右，第一列为原图，第二列为解析真值图，第三列表示粗解析预测 p1 的示例图，第四列为最终的解析预测结果 p2 的示例图，第五列为基于解析真值图生成的边缘真值图，最后一列为边缘预测结果 edge 的示例图。

观察粗解析预测图和解析预测图可知，加入边缘约束分支后，解析预测图能够实现更加精细、准确的分割。例如图 3.7 中第一行人物手臂部分、第二行人物耳朵部分、第三行人物脖子以及手的部分等，以及图 3.8 中第二行人物两个手臂之间的空隙部分、第三行人物的下巴部分等，解析预测图中人物的耳朵部分分割的更加准确。

此外，与粗解析预测图相比，解析预测图的边缘通常更加平滑。例如图 3.7 中第二行的人物的背部线条以及手臂线条、第三行的人物的手臂线条、第四行的人物的腿部线条等。

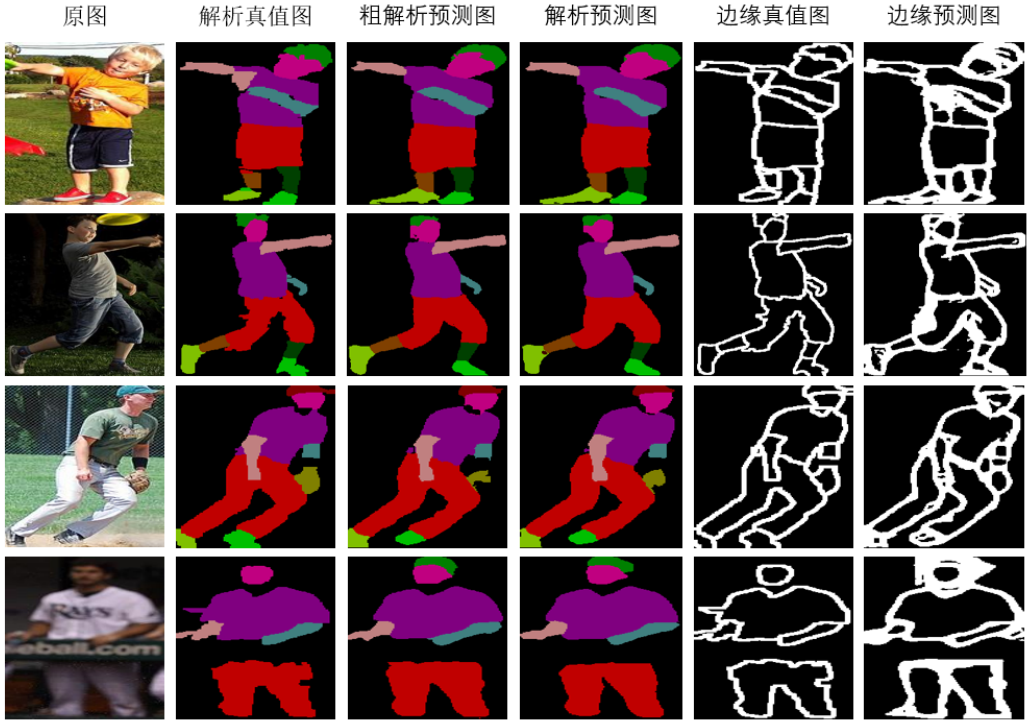

图 6 基于 LIP 数据集的可视化结果示例

图 7 基于 LIP_B 数据集的可视化结果示例

5 总结

通过对两种主流的人体解析方法进行分析，提出了基于双重注意力及边缘约束的人体解析方法。双重注意力模块可以整合特征图中的局部特征及远程依赖关系，弥补卷积操作导致的局部感受野的缺陷。除此之外，为了增强模型区分相邻部分的能力，提高部分边界的准确度，本文采用多任务处理的思想，通过构建双分支结构模型来引入边缘约束，为人体解析分支提供辅助信息，强化对类别信息的区分。最后在 LIP 数据集上的实验结果表明，本文提出的模型在三种不同大小的输入下均实现了分割性能的提升，且随着输入大小的增大，模型的性能在逐步提高。除此之外，在基于各类别部分的消融实验的结果显示，边缘约束分支的加入使得模型在背景类别上的 IoU 明显提高，而双重注意力模块的加入则使得模型在众多类别上实现了更好的分割效果。可视化结果表明，本文的方法能够实现更加精细、准确、流畅的分割效果。

参考文献

[1] Badrinarayanan V, Kendall A, Cipolla R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation[J]. Ieee Transactions On Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495.

[2] Chen L C, Yang Y, Wang J, et al. Attention to Scale: Scale-Aware Semantic Image Segmentation[C]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[3] Chen L C, Zhu Y, Papandreou G, et al. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation[C]. Proceedings of the European conference on computer vision (ECCV), 2018.

[4] Fu J, Liu J, Wang Y, et al. Adaptive Context Network for Scene Parsing[C]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[5] Hou Q, Zhang L, Cheng M M, et al. Strip Pooling: Rethinking Spatial Pooling for Scene Parsing[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[6] Huang Z, Wang C, Wang X, et al. Semantic Image Segmentation by Scale-Adaptive Networks[J]. Ieee Transactions On Image Processing, 2020, 29: 2066-2077.

[7] Deep learning for human parsing: a survey[J]. Arxiv Preprint Arxiv:2301.12416, 2023.

[8] Chen L C, Barron J T, Papandreou G, et al. Semantic Image Segmentation with Task-Specific Edge Detection Using CNNs and a Discriminatively Trained Domain Transform[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.

[9] Zhao H, Shi J, Qi X, et al. Pyramid Scene Parsing Network[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017.

[10] Wang W, Zhu H, Dai J, et al. Hierarchical Human Parsing with Typed Part-Relation Reasoning[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020.

[11] Liu S, Liang X, Liu L, et al. Fashion Parsing With Video Context[C]. Proceedings of the 22nd ACM international conference on Multimedia, 2014.

[12] Ruan T, Liu T, Huang Z, et al. Devil in the Details: Towards Accurate Single and Multiple Human Parsing[C]. Proceedings of the AAAI conference on artificial intelligence, 2019.

[13] Gong K, Liang X, Zhang D, et al. Look into Person: Self-Supervised Structure-Sensitive Learning and a New Benchmark for Human Parsing[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017.

[14] Long J, Shelhamer E, T D. Fully convolutional networks for semantic segmentation[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015.

[15] Chen L C, Papandreou G, Schroff F. Rethinking atrous convolution for semantic image segmentation[J]. Arxiv Preprint Arxiv:1706.05587, 2017.

[16] Fu J, Liu J, Tian H, et al. Dual Attention Network for Scene Segmentation[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[17] Huang Z, Wang X, Huang L, et al. CCNet: Criss-Cross Attention for Semantic Segmentation[C]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[18] Jin Z, Liu B, Chu Q, et al. ISNet: Integrate Image-Level and Semantic-Level Context for Semantic Segmentation[C]. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

[19] Yu C, Wang J, Gao C, et al. Context Prior for Scene Segmentation[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[20] Dong J, Wang W, Zhu L. Object-Contextual Representations for PointNet[C]. 2022 34th Chinese Control and Decision Conference (CCDC), 2022.

[21] 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.

[22] Xia F, Wang P, Chen L C, et al. Zoom Better to See Clearer: Human Part Segmentation with Auto Zoom Net[J]. Proc. Of the European Conference On Computer Vision (Eccv), 2015, 1.

[23] Zhang X, Chen Y, Zhu B, et al. Part-Aware Context Network for Human Parsing[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020.

[24] Liu S, Feng J S, Domokos C, et al. Fashion Parsing With Weak Color-Category Labels[J]. Ieee Transactions On Multimedia, 2013, 16(1): 253-265.

[25] Zhang Z, Su C, Zheng L, et al. Correlating Edge, Pose with Parsing[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020.

[26] Liu K, Choi O, Wang J, et al. CDGNet: Class Distribution Guided Network for Human

Parsing[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022.

[27] Zhao J, Li J, Nie X, et al. Self-Supervised Neural Aggregation Networks for Human Parsing[C]. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017.

[28] Luo Y, Zheng Z, Zheng L, et al. Macro-Micro Adversarial Network for Human Parsing[C]. Proceedings of the European conference on computer vision (ECCV), 2018.

[29] Liang X, Gong K, Shen X, et al. Look into Person: Joint Body Parsing & Pose Estimation Network and a New Benchmark[J]. Ieee Transactions On Pattern Analysis and Machine Intelligence, 2019, 41(4): 871-885.

[30] 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, 2017, 40(4): 834-848.

[31] Nie X, Feng J, Yan S. Mutual Learning to Adapt for Joint Human Parsing and Pose Estimation[C]. Proceedings of the European Conference on Computer Vision (ECCV), 2018.

[32] Zhang S, Qi G J, Cao X, et al. Human Parsing With Pyramidical Gather-Excite Context[J]. Ieee Transactions On Circuits and Systems for Video Technology, 2021, 31(3): 1016-1030.

[33] Song J, Shi Q, Li Y, et al. Enhanced Context Learning with Transformer for Human Parsing[J]. Applied Sciences, 2022, 12(15): 7821.