2020 年 10 月 第 10 期
Vol. 41 No. 10 2020

结合编解码网络的人体解析方法

李博涵,许 敏,王 凯,孙 翔,谭守标

(安徽大学 电子信息工程学院,合肥 230000)
E-mail: tsb@ ustc. edu

摘 要: 人体解析是语义分割的一个子任务,只对图片中的人物进行分割而忽略背景信息. 人体解析任务由于其复杂性,导致现有网络分割不够精确. 本文针对该情况提出了一种编解码网络. 在编码器中,对特征提取网络的下采样倍数进行调整以得到合适分辨率的特征图. 在解码器中,通过金字塔池化网络来提取上下文信息,并采取空间加通道的双注意力模块来修正特征图. 本文的网络与经典的编解码网络在公开的人体解析数据集(LIP)上进行了对比,较 Unet 提升了 10.70% MIOU,较 Deeplabv3 + 提高了 1.93% MIOU. 实验结果表明,本文网络的特征提取能力以及解析能力更加适合人体解析任务.

关 键 词: 语义分割; 人体解析; 编解码网络; 金字塔池化; 注意力模块
中图分类号: TP391 **文献标识码:** A **文 章 编 号:** 1000-1220(2020)10-2184-05

Human Parsing Method Combined with Encoder-decoder Network

LI Bo-han, XU Min, WANG Kai, SUN Xiang, TAN Shou-biao
(School of Electronics and Information Engineering, Anhui University, Hefei 230000, China)

Abstract: Human parsing is a sub task of semantic segmentation, which only segments the characters in the image and ignores the background information. Due to the complexity of human parsing, the existing network segmentation is not accurate enough. This paper proposes an encoder-decoder network for this situation. In the encoder, the down sampling multiple of the feature extraction network is adjusted to get the appropriate resolution of the feature map. In the decoder, the pyramid pooling network is used to extract the context information, and the dual attention modular of space plus channel is used to modify the feature map. The network in this paper is compared with the classic encoder-decoder network on the open human parsing dataset (LIP), which is 10.70% more Miou than Unet and 1.93% more Miou than deeplabv3 +. The experimental results show that the feature extraction ability and analytical ability of this network are more suitable for human parsing.

Key words: semantic segmentation; human parsing; encoder-decoder network; pyramidal pooling; attention module

1 引 言

语义分割是目前计算机视觉领域一个非常关键的问题. 从宏观上来讲,语义分割解决的问题是根据某种规则将图片的每一个像素点分配与其相对应的类别标签,所以语义分割可以被看作是像素级别的分类任务[1]. 语义分割能在很多具有挑战性的视觉任务中应用,比如: 遥感测绘、自动驾驶、人机交互、医疗图像分割[2]等等.

在传统语义分割算法中,有阈值分割法、区域分割法、边缘分割法、基因编解分割法、小波变换分割法等等[3]. 传统语义分割算法对特征的描述能力有限,在场景复杂的情况下难以有良好的表现.

自从 Alex 在 2012 年的 ImageNet 比赛上使用卷积神经网络取得了惊人的效果之后,卷积神经网络进入了快速发展的时期,尤其是在计算机视觉领域,卷积神经网络替代了大部分的传统机器学习算法. 由于其强大的特征学习能力,目前卷积神经网络对于图片的分类技术已经非常成熟,但是将其运用到语义分割中却是计算机视觉领域一项非常具有挑战性的任务. 2015 年,Long 等人[4]提出了一种全连接卷积神经网络(Fully Convolutional Network, FCN),该网络将全连接层用对应尺度的卷积层去替代,然后通过上采样的方式把特征图恢复成原始输入图片的大小. 该算法首次以端到端的形式进行图像分割,并且在自然场景下取得了较好的效果. 由于 FCN 算法的提出,类 FCN 算法也被相继提出. Badrinarayanan 等人提出了编解码网络 SegNet,该网络由两部分组成,原始图片由编码网络提取深层语义信息,再由解码网络恢复到输入图片的分辨率,并输出相应的分割结果[5]. SegNet 提出的编码-解码思想是目前语义分割领域中最热门的设计思路之一. Ronneberger 等人提出了一种全新的特征融合方法,即在维度上进行拼接,先对原图进行多次卷积下采样,再将底层特征图上采样与次底层特征图进行维度拼接,重复多次后恢复到原始输入图像大小,由于该网络结构在形状上酷似大写的字母 U,作者命名为 Unet[6]. Zhao H 等人提出的金字塔场景解析网络,利用不同层次的上下文信息,把全局特征与局部特征加

收稿日期: 2020-03-30 收修改稿日期: 2020-04-21 基金项目: 国家自然科学基金项目(61772032)资助. 作者简介: 李博涵,男,1997 年生,硕士研究生,研究方向为计算机视觉; 许 敏,女,1994 年生,硕士研究生,研究方向为计算机视觉; 王 凯,男,1995 年生,硕士研究生,研究方向为计算机视觉; 孙 翔,男,1991 年生,硕士研究生,研究方向为计算机视觉; 谭守标(通讯作者),男,1976 年生,博士,教授,研究方向为图像、视频分析.

入到预测中,增强了网络的表现力,在多个数据集上验证了算法的有效性[7].

人体解析作为语义分割的一个子问题,其忽略背景信息,只对目标人体进行精细分割,如图 1 上半部分所示,而语义分割则会对图片中的所有目标进行解析,如图 1 下半部分所示[8].

图 1　人体解析与语义分割

Fig. 1　Human parsing and semantic segmentation

人体解析对于给定一张包含人的图片,将人按照身体部位以及服饰衣着分割开来,包含的类别一般有人的脸、头发、外套、裤子、背包、帽子等等[9]. 人体解析在诸多领域中都有应用,如智能安防、行人重识别、虚拟穿搭等等. 因此,人体解析由于其潜在的广泛应用,受到了越来越多的关注[10]. Gong 等人提出了一种自监督结构的敏感学习方法用于人体解析,在不需要额外监督的前提下,将人体姿势结构加入到了解析结果中[11]. Liang 等人提出了一种局部到全局再到局部的框架,从底层提取语义信息,将交叉层的语义信息、全局语义信息、局部超像素语义信息进行融合,取得了 state-of-the-art 的效果[12]. Ruan T 等人提出了一种多任务学习来进行人体解析,通过增加边缘检测网络辅助优化模型表现[13].

2　结合编解码网络的人体解析方法

人体解析数据集的特点是输入分辨率较低以及类别多且错综复杂. 本文提出的结合编解码网络的人体解析模型如图 2 所示. 输入图片(a) 首先会经过编码网络即卷积神经网络(b) 得到初步的特征图,针对输入分辨率低的问题,本文在卷积神经网络下采样阶段选择了合适的下采样倍数以确保在上采样时能够保留足够的语义信息. 初步特征图尺度单一且无法体现各像素、各通道间的权重性. 于是本文将初步特征图经过上下文提取网络(c) 后得到多尺度的特征图,再将该特征图经过双注意力模块(d) 给像素和通道赋予权重,最终上采样输出得到预测结果.

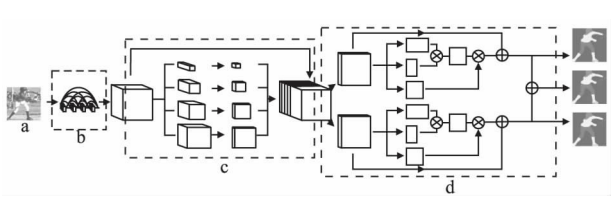

图 2　人体解析模型网络结构

Fig. 2　Network structure of human parsing model

2.1　卷积神经网络提特征

在深度学习领域,尤其是在计算机视觉领域,卷积神经网络有着无可替代的重要地位. 卷积神经网络是一种端到端的前馈神经网络,由卷积层、池化层以及全连接层构成. 近年来,更多的学者采取用 1* 1 的卷积层去代替全连接层,这样会带来两个好处,一个是不会改变图像的空间结构,第二点是由于全连接层的输入尺寸是固定的,导致全连接层的参数个数由图像的大小而定,这显然违背了端到端训练的思想,而卷积层的输出尺寸是任意的,是因为卷积核的参数个数与输入图像的信息无关.

1998 年,LeCun 首次提出卷积神经网络,即 LeNet-5 网络. LeNet-5 的设计主要用于手写数字识别,且大大促进了手写数字识别的商业化使用. 然而在这几十年间里,卷积神经网络由于技术限制等因素,发展较为缓慢. 2014 年,牛津大学视觉几何组提出了 VggNet,其创新性的采用多个小卷积核来代替大卷积核,如一个 5* 5 的卷积核,可以被两个 3* 3 的卷积核所替代,一个 7* 7 的卷积核可以被三个 3* 3 的卷积核替代,成功构建了 16 和 19 层的神经网络. 之所以采取小卷积核,是因为在减少参数量的同时,还增加了层与层之间的非线性变换,使得神经网络对特征的学习更强. 同年,Google 提出 Inception-v1 网络,即 GoogLeNet 网络,其主要解决了网络的无限堆叠导致的过拟合以及计算量增加的问题. 但是这些操作从根本上,都没有解决网络加深导致梯度弥散的情况,直到 2015 年何恺明提出的残差网络 Resnet,其主要思想就是在网络中增加了直连通道,Resnet 在 ILSVRC2015 比赛上夺得冠军,且参数量远小于 VggNet.

此前的网络结构是将输入做一个非线性变换,而直连通道的思想则是将当前网络块输出的结果和之前网络块的结果直接进行数值上相加的一个操作,允许保留之前网络块的一定比例的输出[14],如式(1) 所示:

$$X_i = F_i(X_{i-1}) + X_{i-1} \tag{1}$$

其中,X_i 表示第 i 层的输出,F 表示非线性变换. 残差结构的思想解决了传统卷积神经网络在信息传递的时候导致的信息丢失、梯度消失和梯度爆炸等情况,以至于很深的网络无法训练.

Gao[15] 等人提出的 Densenet 的思路和 Resnet 一致,都是如何复用特征以达到梯度不弥散,其主要区别在于 Densenet 建立的连接不只与前一层,而是前面所有层和后面层的连接以及 Resnet 是将前面层与后面层进行逐像素数值上的相加,而 Densenet 采取的是通道的合并,如式(2) 所示:

$$X_i = F_i([X_0, X_1, \cdots, X_{i-1}]) \tag{2}$$

其中,第 i 层输入 X_i 为之前所有层的特征拼接,即 $[X_0, X_1, \cdots, X_{i-1}]$,$F_i$ 定义为 3 个连续的函数,即 Batchnormalization、Relu 以及 3* 3 的卷积. 如此巧妙的设计使得 Densenet 在计算量更少的前提下比 Resnet 拥有更好的效果.

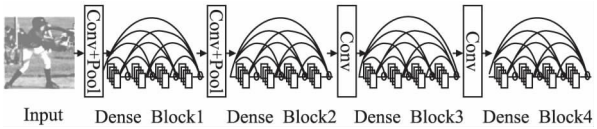

图 3　卷积神经网络提取特征

Fig. 3　Feature extraction by convolution neural network

本文的特征提取网络采用的是修改后的 Densenet-121,结构如图 3 所示. 由于人体数据集,分辨率相对来说较小,多

次下采样会导致特征图更加小,损失过多的空间细节,于是本文取消了第 3 个和第 4 个 Dense Block 后的池化层,因此,特征图的分辨率为输入图片的 1/8. 为了达到端到端的效果,本文取消了网络的全连接部分.

2.2 上下文提取网络

由于卷积神经网络提取的特征是单一尺度,无法很好的表达上下文信息,而人体数据集各类别信息紧密连接,所以针对该情况,特征图需要同时包含全局特征和局部特征才能够在上采样恢复时,在边缘处更加平滑,更接近真实值. 本文借鉴金字塔池化模块[7]来提取上下文信息,并且针对人体解析数据的特殊性,选择合适的池化方式,使其更符合情形,达到特征增强的目的,如图 4 所示.

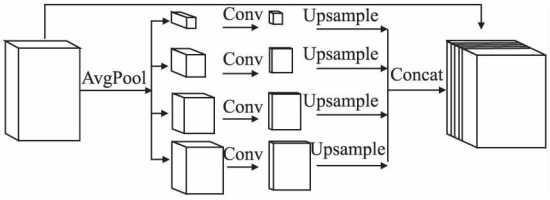

图 4 上下文提取网络

Fig. 4 Context extraction network

对于卷积神经网络提取的特征图,经过平均池化层后得到 B* C* 1* 1,B* C* 2* 2,B* C* 3* 3,B* C* 6* 6 这样四种尺度的特征图(C 代表通道数,与输入特征图的通道数一致). 这四种尺度的特征图会经过 1* 1 的卷积,将其通道数降至原来的 1/4,再上采样至池化之前的宽高,最后与原特征图进行通道合并.

2.3 双注意力模块

注意力模块可以理解为用一个参数很小的网络来修正我们的特征图. 本文在上采样阶段引入双注意力模块[16],来整合上下文网络提取的高维特征图. 其中该注意力模块分为空间注意力模块和通道注意力模块,如图 5 所示. 由于人体数据

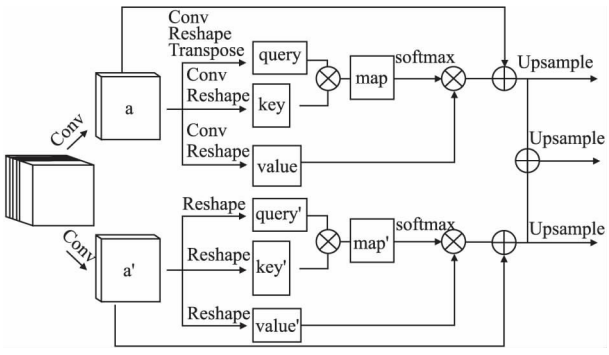

图 5 双注意力模块

Fig. 5 Dual-attention model

集的特殊性,图像中各个类别之间有着依赖性,而空间注意力模型能很好的处理多层次的依赖[17]. 空间注意力模型如图 5 的上半部分所示. 对于特征图 $a \in R^{C^* H^* W}$,将特征图上的每个点看成是长度为通道数 C 的特征向量,特征图 a 通过 1* 1 的卷积、压缩、转置提取长度为 $C/8$ 的向量 $query \in R^{N^* C/8}$,其中 $N = H^* W$. 同样,a 通过 1* 1 的卷积、压缩提取长度为 $C/8$ 的向量 $key \in R^{\frac{C}{8}^* N}$. 利用矩阵的乘法,相当于特征空间变化,可以利用位置信息,于是 $query$ 向量与 key 向量的积 $map \in R^{N^* N}$ 便可以表示原特征图上每个点的相似度,这一点与全连接层非常相似,但是全连接层会破坏原始数据,于是对于该相似度,经过 $softmax$ 生成总和为 1 的权重向量后,权重向量与 $value$ 图相乘得到了在空间上的注意力权值,最后将该权值与原特征图相加. 便得到了空间注意力模型的输出.

对于特征提取网络输出的高维度特征图,经过金字塔上下文提取模块后,维度进一步提升. 而卷积操作只是在一个局部的空间进行操作,无法有效提取通道间的关系[18]. 因此本文将一个通道的特征编码成一个向量,将每个通道赋予一个权重,信息丰富的通道赋予高权重增强信息表达,信息匮乏的通道赋予低权重. 通道注意力模型如图 5 下半部分所示. 对于特征图 $a' \in R^{C^* H^* W}$ 经过通道压缩得到 $query' \in R^{C^* N}$,其中 $N = H^* W$. 同样,a' 经过通道压缩并转置得到 $key' \in R^{N^* C}$. $query'$ 与 key' 的矩阵相乘便得到通道与通道之间的相似度 $map' \in R^{C^* C}$. 该相似度经过 $softmax$ 生成权重与 $value$ 图相乘得到权值,与原特征图相加便得到了通道注意力模型的输出. 最后注意力模块分别输出空间注意力的结果、通道注意力的结果以及两个注意力进行元素求和后的结果. 三个分支的输出分别进行卷积、上采样后与真实标签图进行损失函数的计算,其公式如式(3)所示:

$$loss = L(y_{sum}, \hat{y}) + L(y_{patt}, \hat{y}) + L(y_{catt}, \hat{y}) \tag{3}$$

其中,L 表示交叉熵损失函数,y_{sum} 表示双注意力特征融合分支上采样的预测图,y_{patt} 表示空间注意力分支上采样的预测图,y_{catt} 表示通道注意力分支上采样的预测图,$\hat{y}$ 表示真实标签.

3 实验结果与分析

3.1 实验数据集

本文实验数据集用的是 Look Into Person(LIP)公开数据集. LIP 数据集是目前为止,数量最大的单人人体解析数据集,一共包含 50462 张图片,20 个类别(包括背景类). 其中训练集共有 30462 张图片,验证集和测试集分别为 10000 张图片. 由于 LIP 数据集规模庞大,且人体姿势幅度大,衣饰、背景复杂,所以 LIP 数据集是人体解析任务中最有挑战性的数据集.

3.2 评价指标

本文将从三个指标去评价本文模型的表现,即像素精度(PA)、均像素精度(MPA)、均交并比(MIOU). 以下公式中 k 表示类别,p 表示像素.

1) 像素精度(PA):计算识别正确的像素占所有像素的比例,如式(4)所示:

$$PA = \frac{\sum_{i=0}^{k} p_{ij}}{\sum_{i=0}^{k} \sum_{j=0}^{k} p_{ij}} \tag{4}$$

2) 均像素精度(MPA):对于每一个类,计算其中识别正确的像素占该类所有像素的比例,最后求所有类的平均,如式(5)所示:

$$MPA = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ij}}{\sum_{j=0}^{k} p_{ij}} \tag{5}$$

3) 均交并比(MIOU):对于每一个类,计算每个类的交并比(IOU),最后求每个类的平均. 均交并比能够预测像素集和

真实像素集之间的比例. 均交并比是语义分割领域使用最广泛的一个评价指标[8],如式(6)所示:

$$MIOU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ij}}{\sum_{j=0}^{k}p_{ij} + \sum_{j=0}^{k}p_{ji} - p_{ii}} \tag{6}$$

3.3　实验细节

本实验的软件平台为 Ubuntu16.04,硬件平台为 NVIDIA RTX2080Ti 显卡、i7-7700k 处理器以及 16G 内存. 网络训练采用的是 Pytorch1.3 框架,本文的特征提取网络使用 ImageNet 的预训练模型进行参数初始化,在训练和测试的时候,网络的输入缩放到 256* 256 的宽高,Batchsize 设定为 14,共训练 70 个 epoch,其中每隔 10 个 epoch 学习率衰减至原来的十分之一,图片经过卷积神经网络后的初步特征图大小为 1024* 32 * 32. 由于人体数据集背景类占比最多,数据较稀疏,所以本文选择 Adam 优化器,其中初始学习率采用默认值 0.001,为了加速收敛,本文的损失函数采用交叉熵损失.

3.4　实验结果与分析

本文在 LIP 数据集上与 Segnet、Unet、Deeplabv3 + 网络进行对比,为了保证计算结果的可对比性,这三种网络在训练时候的数据预处理、损失函数、优化器、学习率都与本文方法保持一致. 测试阶段分别计算其 PA,MPA,MIOU 三种评价指标,实验结果如表 1 所示.

表 1　不同方法在 LIP 数据集的结果

Table 1　Results of different methods in LIP dataset

Methods	PA(%)	MPA(%)	MIOU(%)
Segnet	77.61	40.12	32.28
Unet	78.93	45.21	34.73
Deeplabv3 +	83.37	54.83	43.50
Our	83.91	56.81	45.43

从表 1 中可以看到,本文的方法与其他三种方法有着显著的提升. 像素精度、均像素精度与均交并比都是最优的. 其中,Deeplabv3 + 是近年来编解码网络中精度最高的网络之一,本文方法与 Deeplabv3 + (Resnet101) 相比,在像素精度上提升了 0.54%,均像素精度上提升了 1.98%,均交并比上提升了 1.93%. 因为均交并比最能够反映预测图与真实图的对比情况,所以均交并比是语义分割领域最为常用和关键的指标. 本文较 Deeplabv3 + 在 MIOU 上有 1.93% 的提升,说明了本文方法在人体解析问题上的有效性. 为了更直观的看到各个网络的对比,本文将网络的输出进行可视化,如图 6 所示.

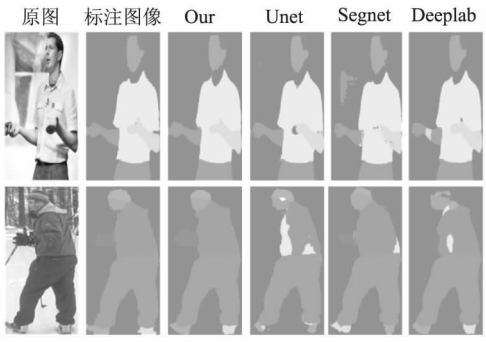

图 6　不同方法的可视化效果

Fig. 6　Visual effects of different methods

从第一组图中可以看到,Unet 在左手处的分割不连续,Segnet 则是出现了大范围的误识别且左手处分割混乱,而 Deeplabv3 + 则是将部分右手区域分割成了左手. 从第二组图中可以看到,Unet 网络易将部分外套类别识别成上衣且左右鞋子分割混乱,Segnet 网络虽然没有大面积的出现外套识别成上衣的情况,但是在头部的边缘分割处不够细腻且弄反了左右鞋子,而 Deeplabv3 + 则是在整个头部区域都分割地十分粗糙. 可以看出,本文方法在正确分类的同时,在边缘分割处也更为细腻.

表 2　本文模块在 LIP 数据集的对比结果

Table 2　Comparison results of this module in LIP dataset

Methods	PA(%)	MPA(%)	MIOU(%)
B-32s	80.40	47.62	37.63
B-16s	82.86	54.08	42.81
B-8s	83.36	55.56	43.79
B-8s + P	83.55	56.12	44.65
B-8s + P + A	83.91	56.81	45.43

为了探究改变下采样倍数、上下文提取高维特征网络以及注意力解码高维特征网络的有效性,本文在 LIP 数据集上进行对比实验,结果如表 2 所示,其中 B 代表本文的基础特征提取网络 Densenet-121,其中 32s 代表从原图中下采样 32 倍再恢复至原图,16s 代表从原图中下采样 16 倍再恢复至原图,8s 代表从原图中下采样 8 倍再恢复至原图,P 代表使用金字塔上下文提取网络提取高维度特征,A 代表用双注意力模块去解码高维度特征. 从表 2 中,可以看出,对于人体解析问题,采用 256* 256 的输入分辨率,8、16、32 倍下采样的效果是逐步递减的. 本文在 8 倍下采样的基础上增 加了金字塔池化和双注意力网络这样的解码网络后,在像素精度上能够提升 0.55%,均像素精度能够提升 1.25%,均交并比能够提升 1.64%.

表 3　各方法的 IOU 值

Table 3　IOU value of each method

Methods	Segnet	Unet	Deeplabv3 +	B-32s	B-16s	B-8s	B-8s + P	B-8s + P + A
bkg	78.57	80.90	84.31	80.12	83.31	84.18	84.28	84.56
hat	43.19	48.47	55.12	47.76	54.49	56.94	57.28	57.49
hair	55.73	59.76	63.39	56.23	61.92	64.73	64.86	65.45
gloves	13.68	14.90	22.95	16.95	24.3	26.90	29.12	30.52
sunglasses	12.31	20.40	17.20	6.38	13.9	22.66	19.59	20.87
upperclothes	48.08	53.03	61.82	57.93	61.54	62.46	61.96	62.84
dress	7.96	7.84	28.15	28.19	27.95	28.20	28.54	27.11
coat	36.98	37.64	49.79	47.41	50.41	50.38	50.01	51.15
socks	25.07	32.29	37.72	27.61	36.6	38.51	39.2	39.41
pants	57.80	60.63	69.15	65.02	68.77	69.31	69.46	69.86
jumpsuits	8.21	9.02	24.19	21.97	26.14	24.58	29.61	27.50
scarf	2.20	3.20	6.00	5.33	10.97	12.98	10.26	13.16
skirt	8.33	10.00	20.71	20.69	21.3	22.00	23.37	21.31
face	63.33	65.05	67.38	60.77	65.02	67.84	67.98	68.34
leftarm	36.54	37.54	49.43	38.12	46.55	47.94	49.19	50.54
rightarm	40.43	41.43	52.85	40.37	49.4	50.85	52.43	53.59
leftleg	33.09	31.89	45.41	39.35	44.37	41.31	44.83	46.55
rightleg	29.61	32.11	45.31	38.72	44.07	41.34	44.43	46.90
leftshoe	23.17	24.14	34.05	26.93	32.38	31.04	33.28	35.12
rightshoe	21.41	24.34	35.23	26.88	32.84	31.62	33.44	36.43
avg	32.28	34.73	43.50	37.63	42.81	43.79	44.65	45.43

为了更加仔细地比较各个网络在每个类别上的表现,本

文以 MIOU 指标为例,展示各个网络在每个类别的 IOU,如表 3 所示. 从表中可以看出,Segnet、Unet 对于连衣裙、围巾类别识别效果较差, Deeplabv3 + 整体表现较好,但是在围巾类识别同样不理想. 本文在基准网络上增加了解码网络,最直观的提升在于左右手臂、左右脚以及左右鞋子类别上. 实验表明,本文方法能够很好的将各类别之间的相关信息进行提取,高效地融合全局以及局部的特征.

4 结 论

本文提出了一种结合编解码网络的人体解析方法,图片经过编码网络,即修改后的 Densenet-121 网络后输出特征图,然后用上下文金字塔加上双注意力模块的组合进行解码输出预测图. 对比实验验证了取消编码网络的两个下采样模块和解码网络的有效性,并且在目前最大的人体解析数据集上进行了测试,均优于目前主流的编解码网络.

References:

[1] Pathak D, Shelhamer E, Long J, et al. Fully convolutional multi-class multiple instance learning [C] //International Conference on Learning Representations, 2015.

[2] Garcia-Garcia A, Orts-Escolano S, Oprea S, et al. A survey on deep learning techniques for image and video semantic segmentation [J]. Applied Soft Computing, 2018, 70: 41-65, https://doi.org/10.1016/j.asoc.2018.05.018.

[3] Wang Xiao-chuan. Analysis of 43 cases of MATLAB neural network [M]. Beijing: Beihang University Press, 2013.

[4] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 3431-3440.

[5] 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.

[6] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation [C] //International Conference on Medical Image Computing and Computer-Assisted Intervention, 2015: 234-241.

[7] 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: 2881-2890.

[8] Shao Jie, Huang Qian, Cao Kun-tao. A survey of human body analysis based on deep learning [J]. Journal of University of Electronic Science and Technology of China, 2019, 48(5): 644-654.

[9] Liang X, Liu S, Shen X, et al. Deep human parsing with active template regression [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(12): 2402-2414.

[10] Yamaguchi K, Kiapour M H, Ortiz L E, et al. Parsing clothing in fashion photographs [C] //2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012: 3570-3577.

[11] 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: 932-940.

[12] Liang X, Xu C, Shen X, et al. Human parsing with contextualized convolutional neural network [C] //Proceedings of the IEEE International Conference on Computer Vision, 2015: 1386-1394.

[13] 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, 33: 4814-4821.

[14] 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.

[15] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 4700-4708.

[16] Fu J, Liu J, Tian H, et al. Dual attention network for scene segmentation [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 3146-3154.

[17] Roy A G, Navab N, Wachinger C. Concurrent spatial and channel´ squeeze & excitationín fully convolutional networks [C] //International Conference on Medical Image Computing and Computer-Assisted Intervention, 2018: 421-429.

[18] Hu J, Shen L, Sun G. Squeeze-and-excitation networks [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7132-7141.

附中文参考文献:

[3] 王小川. MATLAB 神经网络 43 个案例分析 [M]. 北京: 北京航空航天大学出版社, 2013.

[8] 邵 杰, 黄 茜, 曹坤涛. 基于深度学习的人体解析研究综述 [J]. 电子科技大学学报, 2019, 48(5): 644-654.