文章编号：0253-374X(2022)10-1385-07

DOI：10.11908/j.issn.0253-374x.22226

基于多尺度融合增强的服装图像解析方法

陈丽芳，余恩婷

（江南大学 人工智能与计算机学院，江苏 无锡 214000）

摘要：基于卷积神经网络中的各个层次特征，提出了一种基于多尺度融合增强的服装图像解析方法。通过融合增强模块，在考虑全局信息的基础上对包含的语义信息和不同尺度特征进行有效融合。结果表明：在Fashion Clothing测试集上的平均F1分数达到60.57%，在LIP(Look Into Person)验证集上的平均交并比(mean intersection over union，MIoU)达到54.93%。该方法可以有效地提升服装图像解析精度。

关键词：服装图像解析；多尺度融合增强网络；卷积神经网络

中图分类号：TP399　　**文献标志码**：A

Clothing Image Parsing Method Based on Multi-scale Fusion Enhancement

CHEN Lifang, *YU Enting*

(School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214000, China)

Abstract: By using the features of each level in convolutional neural network, a clothing image parsing method based on multi-scale fusion enhancement was proposed. Through the fusion enhancement module, the semantic information and the features in different scales were effectively fused with the consideration of global information. The results show that the average F1 score on the Fashion Clothing test set reaches 60.57%, and the mean intersection over union (MIoU) on the Look Into Person (LIP) validation set reaches 54.93%. The method can effectively improve the accuracy of clothing image parsing.

Key words: clothing image parsing; multi-scale fusion enhanced network; convolutional neural network

随着服装和互联网行业的快速发展，服装图像解析作为图像处理的一个重要应用有着巨大的发展前景。服装图像解析的目标是对服装图像各个部分的组成进行像素级别的识别，将服装图像按照若干个类别划分为若干个区域。服装图像解析是计算机视觉中一项特定形式的细粒度分割。因此，服装图像解析研究对服装检索[1]、服装推荐[2]和服装合成[3]等领域的发展有着重要意义。Liu等[4]通过深度卷积网络来学习丰富的语义信息以克服不同的身体部位和服装间的语义模糊，同时采用不进行下采样的网络为小尺度对象保留分辨率和局部细节信息，并设计了一个桥梁模块在2个并行的网络间交换互补信息，从而提升网络解析性能。Luo等[5]利用对抗网络解决低级局部和高级语义的不一致性，该网络采用2个鉴别器，分别作用于低分辨率标签图和高分辨率标签映射的多个像素块，强制实现语义和局部的一致性，而且避免了处理高分辨率图像时对抗网络收敛性差的问题。Wang等[6]将以人体为中心的服装图像解析定义为一个基于人体结构的神经信息融合过程，并建立了结合直接推理、自顶向下推理和自底向上推理的3个层次推理的网络结构，可以明确地捕获人体的组成和分解关系，进而提高服装图像解析精度。Gong等[7]首先通过图内推理在一个数据集内的标签之间学习和传播特征信息，然后通过图间转移在多个数据集之间传输语义信息，分析和编码不同数据集之间全局语义一致性以增强图传递能力，实现多层次的解析任务。现有解析方法没有很好地解决服装类别丰富且尺度差异大等问题，导致服装图像解析效果有待提升。因此，提出一种多尺度融合增强网络，在充分发挥深度卷积网络中各个层次特征优势的基础上，利用通道注意力机制增强特征表达，从而提高服装图像解析效果。

收稿日期：2022-05-10
基金项目：国家自然科学基金(61872166)
第一作者：陈丽芳(1973—)，女，教授，硕士生导师，主要研究方向为数字图像处理、深度学习理论与应用、目标三维重建。E-mail：may7366@163.com

论文拓展介绍

1 相关工作

1.1 服装图像解析

服装图像解析在人工智能等领域具有广阔的应用前景。Chen等[8]将深度卷积网络提取到的多尺度特征输入到注意力模块,分别学习各个尺度特征在每个位置上的权重,输出不同信息重要性差异的权值图,然后将不同尺度的权值图分别乘以原始特征,调整不同像素对于不同类别的重要性。Zhao等[9]在利用金字塔结构获取多尺度特征的基础上,提出语义感知模块和边界感知模块,其中语义感知模块选择与类别相关的有区分性的特征,防止不相关的特征被合并到一起,而边界感知模块将多尺度特征与对象边界有效结合,实现精确的局部定位和部分区域间的准确识别。Luo等[10]设计了金字塔残留池结构以捕获全局和局部的上下文信息,利用一种可信指导多尺度监督方法,有效地整合和监督不同尺度的上下文信息,从而解决了人为误标标签时导致的标签解析混乱问题。

1.2 注意力机制

深度学习中的注意力机制源于人类视觉特性,当人类观察事物时,选择性地获取所观察事物的重要特征,忽略不重要特征。深度学习中的注意力机制借鉴了人类的视觉机制,旨在自适应地聚集有相关性的特征,帮助深度学习模型对输入的信息赋予不同的权重,获取更有用的特征,所以注意力机制被广泛应用于语义分割、目标识别和图像分类等计算机视觉领域。Hu等[11]提出通道注意力模块,从通道的维度学习特征的重要程度,选择性地提升对当前任务有用的特征并抑制对当前任务用处不大的特征。由于在特征提取时卷积操作经常把通道和空间信息混合在一起,因此模型效果仍然不够好。Woo等[12]在Hu等[11]提出的通道维度注意力机制的基础上,又提出了同时考虑通道和空间位置维度的混合注意力机制,进而有效地整合全局的上下文特征表达。与其他注意力机制不同的是,Wang等[13]通过自注意力模块Non-local计算任意2个位置之间的相互作用,直接捕获远距离依赖关系,然后将相关性作为权重表征其他位置和当前待计算位置的相似度。Hu等[14]定义了一个聚集算子,有效地聚合给定空间范围上的特征响应,同时设计一个激发算子调整聚合后的特征大小,并将其作为注意力特征重新分发给原始特征,从而以更轻量级的方式提升网络性能。

2 网络设计

2.1 网络结构

提出的多尺度融合增强网络结构如图1所示。首先,以步长为16的ResNet101[14]作为编码网络,提取多个层次的特征图,同时将Conv-1、Res-2、Res-3、Res-4和Res-5的输出依次表示为L_4、L_3、L_2、L_1和H_1;其次,在解码网络中根据分辨率大小分为4个阶段进行解码,并且每个阶段都使用了如图2所示的融合增强模块(FEM),融合不同层次的语义和不同尺度的特征,从而提升预测结果精度;然后,将4个融合增强模块输出的特征图H_5、H_4、H_3和H_2串联拼接在一起,进一步细化特征对各个类别的感知能力;最后,通过线性插值、1×1卷积和argmax操作得到服装图像各个对象的解析结果。

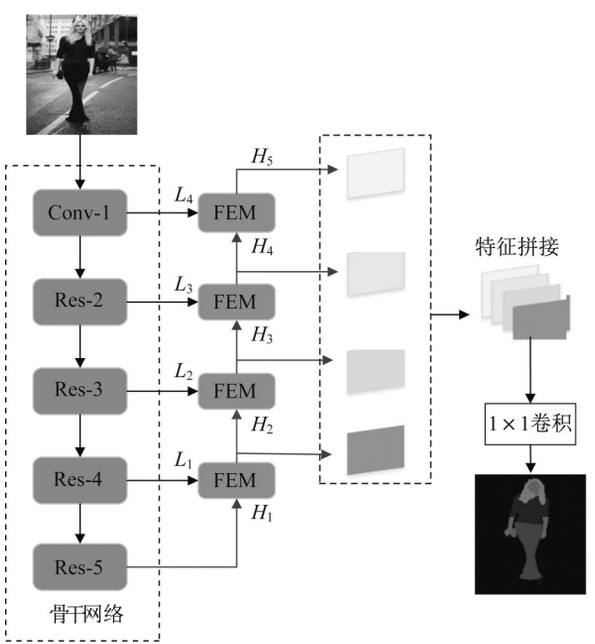

图1 多尺度融合增强网络结构

Fig.1 Structure of multi-scale fusion enhanced network

2.2 融合增强模块

如图2所示,融合增强模块的输入来自编码网络跳跃连接的低层特征L_i和解码网络的深层特征H_i。考虑到显存(2块8 GB显卡)的有限性,用1×1卷积将低层特征L_i和深层特征H_i的通道维度均降低为256。为了融合不同层次的低层特征L_i和深层特征H_i,先将深层特征H_i通过线性插值得到与低层特征L_i相同的分辨率,然后将两者串联在一起。受文献[15]中用不同大小的感受野提取不同尺度特征信息的启发,用3×3、5×5和7×7卷积提取串联后的

特征，分别得到特征图F_{33}、F_{55}和F_{77}。将特征图F_{33}、F_{55}和F_{77}串联在一起，并应用1×1卷积对其进行融合，进而增强网络对不同尺度服装对象的感知。虽然这种网络结构很好地提取了多尺度上下文信息，但是没有考虑全局信息，因此结合Hu等[11]提出的通道注意力机制，从全局角度对特征图进行优化。首先，对特征图F_{357}进行全局池化和Sigmoid激活，形成大小为256×1×1的全局信息图；然后，通过乘法对全局信息图中3个不同尺度的特征图F_{33}、F_{55}和F_{77}分别加权，从而起到强调重要信息、抑制无用信息的作用；最后，将3个加权后的不同尺度特征图进行串联拼接得到最终的输出特征H_{i+1}。

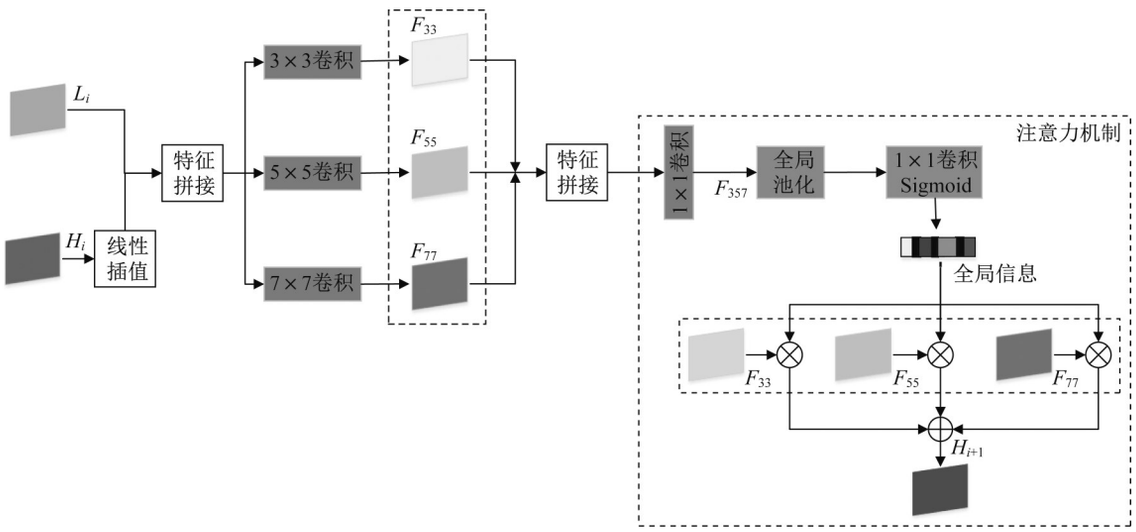

图2　融合增强模块

Fig.2　Fusion enhancement module

3　实验结果与分析

3.1　实验设置

所有实验均是在2个NVIDIA GTX1070 GPU服务器上利用Ubuntu18.04、Python3.6和Pytorch0.4.1搭建的深度学习框架。使用步长为16的预训练好的ResNet101[16]作为骨干网络。在训练和测试时采用的图像大小为320×320。初始学习率设置为0.003，并使用“Poly”学习率策略对学习率进行调整。在训练过程中采用随机梯度下降方法对网络进行训练，动量和权重衰减分别设置为0.9和0.000 5，同时采用随机的图像缩放（从0.5到1.5）、裁剪和左右翻转对数据进行增强。提出的多尺度融合增强网络采用标准的多类的交叉熵损失函数监督网络的学习，使网络预测结果不断接近真实值，在网络参数不断迭代更新过程中实现端到端的学习。

3.2　数据集

实验中使用的数据集是公共数据集Fashion Clothing数据集和LIP（Look Into Person）数据集。Fashion Clothing数据集由Clothing Co-Parsing[17]、Fashionista[18]和Colorful Fashion Parsing Data[19]3个服装数据集组成。由于这些数据集有一些细微的差异，因此现有算法通常把这3个数据集的标签统一为18个类别，最后得到4 371幅像素级别标注的图像。LIP数据集[20]是包含20个类别50 462张图像的大型数据集，其中训练集包含30 462张图像，验证集包含10 000张图像，测试集包含10 000张图像。

在Fashion Clothing数据集中使用像素准确率、前景准确率、平均精确率、平均召回率和平均F1分数5个评价指标对网络性能进行评估。在LIP数据集中使用像素准确率、平均准确率、每个类别的交并比和平均交并比4个评价指标对网络性能进行评估。

3.3　消融实验

表1为各模块消融实验对比结果。√表示在多尺度融合增强网络中加入了此部分，×表示在多尺度融合增强网络中没有加入此部分。F_{357}表示利用3×3、5×5和7×7卷积提取的多尺度特征图，GP表示加入了全局池化结构对特征图进行优化，C表示将4个融合增强模块（FEM）输出的特征图H_2、H_3、H_4和H_5串联在一起。从表1可以看出，在加入3×3、5×5和7×7卷积后卷积感受野变大，从而提取到更丰富的上下文信息，有效地

解决服装图像目标尺度差异较大的问题。多尺度融合增强网络在加入全局池化结构后提高了模型的特征表达能力,使得解析准确率明显提升。最后,将4个融合增强模块输出的特征图H_2、H_3、H_4和H_5串联拼接在一起,进一步增强融合后的特征信息。同时,为了验证融合增强模块个数对网络性能的影响,在解码网络中从左到右使用了不同数量的融合增强模块并在表2中列出了实验结果。由表2可见,随着解码网络中使用的融合增强模块数量的增加,像素准确率、前景准确率、平均精确率、平均召回率和平均F1分数5个评价指标都有明显的提升,进一步验证了融合增强模块在增强模型特征表达和提升网络性能的有效性。

表1 各模块消融实验结果

Tab.1 Results of ablation experiments for each module

方法	F_{357}	GP	C	像素准确率/%	前景准确率/%	平均精确率/%	平均召回率/%	平均F1分数/%
骨干网络	×	×	×	92.03	69.23	53.60	58.38	55.89
多尺度融合增强网络	√	×	×	93.43	74.24	58.11	61.74	59.87
多尺度融合增强网络	√	√	×	93.51	74.33	58.27	62.22	60.18
多尺度融合增强网络	√	√	√	93.13	73.12	58.73	62.54	60.57

表2 不同融合增强模块个数下实验结果

Tab.2 Experimental results under different fusion enhancement modules

融合增强模块个数	像素准确率/%	前景准确率/%	平均精确率/%	平均召回率/%	平均F1分数/%
1	92.47	70.84	54.44	58.67	56.48
2	93.09	72.68	57.16	60.75	58.90
3	93.43	74.02	58.33	61.95	60.08
4	93.51	74.33	58.27	62.22	60.18

3.4 Fashion Clothing数据集上的性能对比

表3给出了多尺度融合增强网络与其他先进方法在Fashion Clothing数据集上的性能对比。从表3可以看出,多尺度融合增强网络与TGPNet(trusted guidance pyramid network)[10]和TPRR(typed part-relation reasoning)[21]相比,像素准确率、前景准确率、平均精确率、平均召回率和平均F1分数都有明显的提升。这主要是由于多尺度融合增强网络充分利用了编码过程中提取到的所有特征,增强了不同层次特征的信息表达,更适合纹理复杂、目标差异大的服装图像,因此提高了服装图像解析各个评价指标的值。为更清晰地展示多尺度融合增强网络在服装图像分割效果上的提升,在图3可视化了不同方法在Fashion Clothing数据集上的解析结果。从图3c可以看出,TGPNet[10]和TPRR[21]等方法都将半身裙错误地分割为连衣裙,只有本方法准确地解析出半身裙的整个轮廓,表明本方法可以更精准地区分易混淆的类别。从图3a和图3b可以看出,其他方法均没有关注到眼镜和腰带这种类别较小的目标,本方法却精确地分辨出小尺度的眼镜和腰带。因此,相比其他方法本方法可以给予尺度差异较大的目标对象均衡的关注,从而提升服装图像解析效果。

表3 不同方法在Fashion Clothing数据集上的性能对比

Tab.3 Comparison of performance between different methods on Fashion Clothing dataset

方法	像素准确率/%	前景准确率/%	平均精确率/%	平均召回率/%	平均F1分数/%
Yamaguchi等[18]提出的方法	81.32	32.24	23.74	23.68	22.67
Paper doll parsing[22]	87.17	50.59	45.80	34.20	35.13
DeepLabV2[23]	87.68	56.08	35.35	39.00	37.09
Attention[8]	90.58	64.47	47.11	50.35	48.68
TGPNet[10]	91.25	66.37	50.71	53.18	51.92
Compositional neural information fusion[6]	92.20	68.59	56.84	59.47	58.12
TPRR[21]	93.12	70.57	58.73	61.72	60.19
多尺度融合增强网络	93.13	73.12	58.73	62.54	60.57

3.5 LIP数据集上的性能对比

为了进一步验证本方法的有效性和泛化性,表4给出了多尺度融合增强网络与其他方法在LIP数据集上的解析结果。从表4可以看出,多尺度融合增强网络与PGECNet相比,像素准确率、平均准确率和平均交并比3个评价指标分别提升了0.15%、

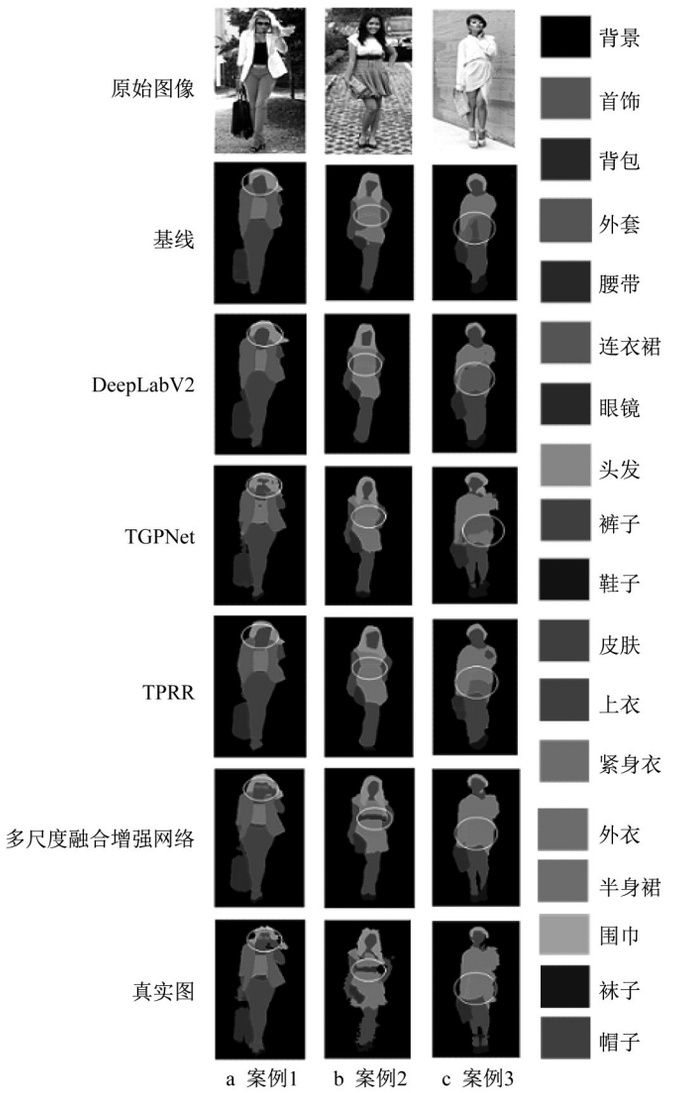

图3　不同方法在Fashion Clothing数据集上的解析结果对比

Fig.3　Comparison of parsing results between different methods on Fashion Clothing dataset

1.14%和0.63%。表5给出了不同方法在LIP数据集上每个类别的交并比。从表5可以看出,多尺度融合增强网络对大部分服装类别的解析都得到了较高的精度,也表明本方法对于不同尺度的目标类别都是有效的。本方法的解析效果不仅在连衣裙、外套和连衣裤等较大的服装类别上有明显的提升,还在帽子、手套和眼镜等较小的服装类别上有明显的改善。因此,验证了多尺度融合增强网络在融合低层特征、深层特征以及增强特征表达方面的有效性。

表4　不同方法在LIP数据集上的性能对比

Tab.4　Comparison of performance between different methods on LIP dataset

方法	像素准确率/%	平均准确率/%	平均交并比/%
DeepLabV2[23]	82.66	51.64	41.64
Attention[8]	83.43	54.39	42.92
ASN(adversarial network)[24]			45.41
MMAN(macro-micro adversarial network)[5]	85.24	57.60	46.93
JPPNet(joint body parsing & pose estimation network)[25]	86.39	62.32	51.37
CE2P(context encedding with edge perceiving)[26]	87.37	63.20	53.10
PGECNet(pyramidical gather-excite context network)[27]	87.50	65.66	54.30
多尺度融合增强网络	87.65	66.80	54.93

表5 不同方法在LIP数据集上每个类别的交并比

Tab.5 Comparison of per-class IoU between different methods on LIP dataset

类别	各方法交并比/%							
	DeepLabV2[23]	Attention[8]	ASN[24]	MMAN[5]	JPPNet[25]	CE2P[26]	PGECNet[27]	多尺度融合增强网络
帽子	56.48	58.87	56.92	57.66	63.55	65.29	66.36	67.16
头发	65.33	66.78	64.34	65.63	70.20	72.54	72.83	72.86
手套	29.98	23.32	28.07	30.07	36.16	39.09	40.76	42.64
眼镜	19.67	19.48	17.78	20.02	23.48	32.73	32.85	35.19
上衣	62.44	63.20	64.90	64.15	68.15	69.46	69.93	69.67
连衣裙	30.33	29.63	30.85	28.39	31.42	32.52	33.78	36.70
外套	51.03	49.70	51.90	51.98	55.65	56.28	56.48	56.82
袜子	40.51	35.23	39.75	41.46	44.56	49.67	48.86	48.18
裤子	69.00	66.04	71.78	71.03	72.19	74.11	74.51	74.95
连衣裤	22.38	24.73	25.57	23.61	28.39	27.23	28.20	33.13
围巾	11.29	12.84	7.97	9.65	18.76	14.19	25.16	21.42
半身裙	20.56	20.41	17.63	23.20	25.14	22.51	26.52	26.51
脸部	70.11	70.58	70.77	69.54	73.36	75.50	75.34	75.79
左胳膊	49.25	50.17	53.53	55.30	61.97	65.14	65.69	66.22
右胳膊	52.88	54.30	56.70	58.13	63.88	66.59	67.33	68.82
左腿	42.37	38.35	49.58	51.90	58.21	60.10	59.36	60.30
右腿	35.78	37.70	48.21	52.17	57.99	58.59	58.82	59.48
左鞋	33.81	26.20	34.57	38.58	44.02	46.63	47.77	46.95
右鞋	32.89	27.09	33.31	39.05	44.09	46.12	47.78	47.78
背景	84.53	84.00	84.01	84.75	86.26	87.67	87.74	88.05
平均交并比/%	41.64	42.92	45.41	46.81	51.37	53.10	54.30	54.93

4 结语

提出了一种基于多尺度融合增强的服装图像解析方法。通过融合增强模块设计,在提取不同尺度特征的基础上,利用通道注意力机制优先考虑全局特征,增强多尺度特征信息,达到获取更多细节特征的目的。实验结果表明,本方法不仅可以提升较大目标的解析效果,还对帽子、腰带和眼镜等小物体的解析效果有明显改善。虽然本方法对较小对象的解析结果有所改善,但是与其他类别相比,小目标对象的解析精度仍然较低。在未来的研究中,将考虑利用目标检测技术定位小目标对象,从而提升小目标对象的解析精度。

作者贡献声明:

陈丽芳:模型网络结构构思、设计、分析,论文修改与校对。

余恩婷:模型网络结构程序与实验设计,论文撰写与修改。

参考文献:

[1] LIU S, SONG Z, LIU G, *et al*. Street-to-shop: cross-scenario clothing retrieval via parts alignment and auxiliary set [C]// 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence: IEEE, 2012: 3330-3337.

[2] 徐慧, 白美丽, 万韬阮, 等. 基于深度学习的服装图像语义分析与检索推荐[J]. 纺织高校基础科学学报, 2020,33(3):64.
XU Hui, BAI Meili, WAN Taoruan, *et al*. Semantic analysis and retrieval recommendation of clothing images based on deep learning [J]. Journal of Basic Science of Textile Colleges, 2020,33(3):64.

[3] ZHU S, URTASUN R, FIDLER S, *et al*. Be your own Prada: fashion synthesis with structural coherence [C]// Proceedings of the IEEE International Conference on Computer Vision. Venice :IEEE,2017:1680-1688.

[4] LIU X, ZHANG M, LIU W, *et al*. BraidNet: braiding semantics and details for accurate human parsing [C]// Proceedings of the 27th ACM International Conference on Multimedia. New York: Association for Computing Machinery, 2019: 338-346.

[5] 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). Berlin: Springer, 2018: 418-434.

[6] WANG W, ZHANG Z, QI S, *et al*. Learning compositional neural information fusion for human parsing [C]//Proceedings of the IEEE International Conference on Computer Vision. Seoul:IEEE, 2019: 5703-5713.

[7] GONG K, GAO Y, LIANG X, *et al*. Graphonomy: universal human parsing via graph transfer learning [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern

Recognition. Long Beach: IEEE, 2019: 7450-7459.

[8] CHEN L C, YANG Y, WANG J, *et al*. Attention to scale: scale-aware semantic image segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE,2016: 3640-3649.

[9] ZHAO Y, LI J, ZHANG Y, *et al*. Multi-class part parsing with joint boundary-semantic awareness[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Long Beach: IEEE, 2019: 9177-9186.

[10] LUO X, SU Z, GUO J, *et al*. Trusted guidance pyramid network for human parsing[C]//Proceedings of the 26th ACM International Conference on Multimedia. New York: Association for Computing Machinery, 2018: 654-662.

[11] HU J, SHEN L, SUN G. Squeeze-and-excitation networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 7132-7141.

[12] WOO S, PARK J, LEE J Y, *et al*. CBAM: convolutional block attention module [C]//Proceedings of the European Conference on Computer Vision (ECCV). Berlin: Springer, 2018:3-19.

[13] WANG X, GIRSHICK R, GUPTA A, *et al*. Non-local neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 7794-7803.

[14] HU J, SHEN L, ALBANIE S, *et al*. Gather-excite: exploiting feature context in convolutional neural networks [C]//NIPS' 18: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Red Hook:Curran Associates Inc., 2018: 9423-9433.

[15] LI H, XIONG P, An J, *et al*. Pyramid attention network for semantic segmentation [J/OL]. [2021-05-06]. https://arxiv.org/abs/1805.10180.

[16] 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. Las Vegas: IEEE, 2016: 770-778.

[17] YANG W, LUO P, LIN L. Clothing co-parsing by joint image segmentation and labeling [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014: 3182-3189.

[18] 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. Providence: IEEE, 2012: 3570-3577.

[19] LIU S, FENG J, DOMOKOS C, *et al*. Fashion parsing with weak color-category labels [J]. IEEE Transactions on Multimedia, 2013, 16(1): 253.

[20] 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. Honolulu: IEEE,2017: 932-940.

[21] 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. Seattle: IEEE, 2020: 8929-8939.

[22] YAMAGUCHI K, HADI K M, BERG T L. Paper doll parsing: retrieving similar styles to parse clothing items[C]// Proceedings of the IEEE International Conference on Computer Vision. Sydney: IEEE,2013: 3519-3526.

[23] 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.

[24] LUC P, COUPRIE C, CHINTALA S, *et al*. Semantic segmentation using adversarial networks[C]// Workshop on Adversarial Training, NIPS 2016. Barcelona: IEEE, 2016: 1-9.

[25] 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, 2018, 41(4): 871.

[26] RUAN T, LIU T, HUANG Z, *et al*. Devil in the details: towards accurate single and multiple human parsing[C]//The Thirty-Third AAAI Conference on Artificial Intelligence. Menlo Park: Association for the Advancement of Artificial Intelligence, 2019:4814-4821.

[27] 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, 2020, 31(3): 1016.