基于 Mask-RCNN 的服装识别与分割

张泽堃[1]，张海波[2,3,*]

（1.北京服装学院 信息中心，北京 100029；
2. 北京服装学院 服装材料研究开发与评价北京市重点实验室 北京市纺织纳米纤维工程技术研究中心，北京 100029；
3.北京服装学院 图书馆，北京 100029）

摘　要：本文提出了一种基于 Mask-RCNN 和数据集 DeepFashion2 的服装识别与分割的方法。基于 Mask-RCNN 的服装识别与分割是基于卷积神经网络的思想，在深度学习框架下通过多线程迭代训练，在 ResNet 网络中得到目标特征后，再通过 RPN 和 RoI Align 将特征输入不同的全连接分支，最后得到具有优化权重的目标检测模型。在不同场景的服装图像中，该模型可以更快更准确的识别出服装并将其分割。

关键词：服装识别；服装分割；Mask-RCNN；ResNet；DeepFashion2；TensorFlow

中图分类号：TS941.2　　**文献标识码**：A　　**文章编号**：1673－0356(2020)06－0020－05

根据中国服装协会发布的《2018－2019 年度中国服装行业发展报告》[1]显示，2018 年全年限额以上服装品类零售额实现 7 742.8 亿，累计增长 9.1%，增速较 2017 年提高了 1.1%，网上销售增长了 22%，大型零售增长 0.99%。2018 年全年服装行业规模以上企业主营业务收入 17 106.57 亿。服装行业发展迅速，计算机行业进军服装领域，如何理解、区分和识别不同的服装图像，以及如何处理海量的服装图像，并从中提取出有用的信息成为当前的研究热点。除了电商的以图搜物外，服装的智能搭配和服装定制都用到了服装的图像识别与分割。目前网购服装的检索主要以文字检索为主，但随着服装数量剧增，服装款式变化多样，传统的图像处理已无法满足当前快速、智能准确的要求，而这也使得计算机在服装图像的处理上遇到了瓶颈。

从最初的人工标注到使用卷积神经网络来训练模型，期间出现了很多图像识别和分割的算法。早期的识别方法有静态图像中的分割、边缘提取、运动检测等，如局部模板方法、光流检测法等，这些方法速度较慢，识别率较低，误报率也高。随着深度学习[2]的不断发展，深层卷积神经网络在图像的处理上更加针对服装图像的特征，具有独特的优势。Kim[3]等为了识别摄像机拍摄的灰度图像中的任务，使用边缘提取的方法提取服装图像的纹理特征，计算欧几里得距离来判定特征向量之间的相似性。只使用服装边缘的直方图特征向量，并不适用于图像缩放、旋转、扭曲等形变。Hidayati[4]等提出了一种基于视觉差异化风格元素的服装风格自动分类方法，而不用低层次特征或模糊关键词来识别服装风格，基于服装设计理论，确定了一组对识别服装风格的特定视觉风格至关重要的风格元素，将服装风格元素归一化为上身的特征向量和下身的特征向量，通过一个判别函数来判断输入特征向量到类别标签的映射关系，利用判别函数来确定服装的类型。Liu[5]等设计了用于服装图像分类检测的 FashionNet，提出了 DeepFashion 数据集，这是一个具有全面注释的大型服装数据集。Luo[6]等研究了频场景中识别服装的相关技术，在 ImageNet 数据集上预训练 GoogleNet 模型，并根据服装本身的特点对网络进行微调，以完成服装的检索任务。

根据现有研究，目前使用较多的检测算法有 R-CNN、Fast R-CNN、Faster R-CNN 和 Mask-RCNN，检测结果的准确性和速度不断提高，但是缺点是需要大量的训练数据。Mask-RCNN[7]是目前最新的实例分割架构，引入了 RoI Align，增加了一个分支用于分割任务，在时间上有一定的优化。而且 DeepFashion2 数据集在 DeepFashion 数据集的基础上进行了优化。

本文基于 Mask-RCNN 和 DeepFashion2 的服装识别与分割，对传统深度学习目标检测算法的训练数据和特征提取器做了调整，使得模型更加适用于服装识别和服装分割，得到的结果更加准确。

1　DeepFashion2 数据集

收稿日期：2019-10-23；**修回日期**：2019-10-30

作者简介：张泽堃(1994-)，男，硕士在读，研究方向为服装智能搭配。
* **通信作者**：张海波(1970-)，男，副研究员，博士，研究方向为服装情感学、图书馆信息化，E-mail：hbdmzhb@126.com。

1.1　DeepFashion2 数据集概况

近年来，时尚产业愈发火爆，时尚服装图像分析成为了热点。现有最大的时尚数据集 DeepFashion 存在标记较为稀疏，没有定义服装姿态，没有对每个像素进行掩膜标注的缺点。为了解决上述缺陷，提出了 DeepFashion2[8]，它是一个大规模的基准集，能全面的进行服装分类和服装图像的标注，包含 49.1 万张时装图像，图片可分为 13 种流行的服饰类别。DeepFashion2 定义了相对全面的任务，包括服装的检测和识别，关键点的标记和服装姿态估计，服装分割，服装的验证和检索。所有的服装图像都有丰富的标注。它有 81 万个服装项目，每个项目都有丰富的注释，其中每件都标有不同的比例、不同大小的遮挡、不同的缩放大小、不一样的视角、精准的边界框、密集的标注和每个像素的掩膜。表 1 为 DeepFashion2 与其他数据集的比较。

表 1　DeepFashion2 与其他数据集的比较

	WTBI[9]	DARA[10]	DeepFashion	ModaNet[11]	FashionAI	DeepFashion2
时　间/年	2015	2015	2016	2018	2018	2019
图片大小	425k	182k	800k	55k	357k	491k
类　别	11	20	50	14	41	13
B-box	39k	7k	X	X	X	801k
Landmarks	X	X	120k	X	100k	801k
Masks	X	X	X	119k	X	801k
Pairs	39k	91k	251k	X	X	837k

DeepFashion2 的贡献主要有三个：(1)构建了拥有丰富标注的大规模数据集，推动了时尚图像分析的发展。拥有丰富的任务定义和最大数量的服装标签。它的标注至少是 DeepFashion 的 3.5 倍，是 ModaNet 的 6.7 倍，是 FashionAI 的 8 倍。(2)在数据集上定义了全部任务，包括服装检测、服装姿态、服装分割与检索。(3)使用 Mask-RCNN 对数据集进行识别和分割。

1.2　DeepFashion2 数据集和基准

1.2.1　数据标签

(1)类别和边界框。对服装图片进行人工标注，并为每个服装项目指定一个类别标签。通过重新对 DeepFashion 的类别进行分组，得到了 13 个没有歧义的服装类别。

(2)服装标签、轮廓和骨架。由于不同类别的衣服(如上下半身服装)有不同的形变和外观变化，通过捕捉服装的形状和结构将特征点连接。将每类服装进行人工标注，每个特征点都被将指定为“可见”或“遮挡”。然后，将特征点通过一定顺序连接后生成轮廓和骨架，还将标注区分为两种类型，即轮廓点和连接点。以上过程控制了标签的质量，生成的骨架有助于人工重新检查这些标记是否具有较高的识别效率。只有当轮廓覆盖整个项目时，标记的结果才合格，否则将重新确定关键点。

(3)掩膜。使用两个阶段的半自动方式为每个项目标记像素掩膜。第一阶段自动从轮廓生成掩膜。在第二阶段，要求人工重新定义掩膜，因为当呈现复杂的人体姿势时，生成的掩膜可能不准确。例如，当从人腿交叉侧视图拍摄图像时，标记会不准确，这时掩膜需要人工调整，如图 1 所示。

图 1　掩膜出现识别错误时进行人工调整

1.2.2　基　准

使用 DeepFashion2 的图像和标签构建了四个基准。对于每个基准测试，训练集图像 39.1 万，验证集图像 3.4 万，测试集图像 6.7 万。

(1)服装检测。通过识别边界框和类别标签来检测图像中的衣服。根据 COCO 数据集，评价标准为 AP_{box}、$AP_{box}^{IoU=0.50}$ 和 $AP_{box}^{IoU=0.75}$。

(2)特征点估计。预测每个图像中检测到的每个服装项目的标志。采用 COCO 用于人体姿态估计的评估指标，通过计算关键点 AP_{pt}、$AP_{pt}^{OKS=0.50}$ 和 $AP_{pt}^{OKS=0.75}$ 的平均精度，其中 OKS 表示目标特征点相

似性。

(3)图像分割。将类别标签(包括背景标签)分配给项目中的每个像素。评估指标是平均精度,包括在掩膜上计算的 AP_{mask}、$AP_{mask}^{IoU=0.50}$ 和 $AP_{mask}^{IoU=0.75}$。

2 Mask-RCNN 服装识别

2.1 检测框架设计思路

在检测模型训练阶段,对具有初始参数的卷积神经网络进行迭代训练,并通过 Tensorboard 来查看训练过程,从而不断修改和优化训练模型的参数,最终得到目标检测模型。在模型测试阶段,将待检测样本输入之前得到的目标检测模型并得到检测结果。主要有检测模型的训练和模型测试两个阶段,如图 2 所示。

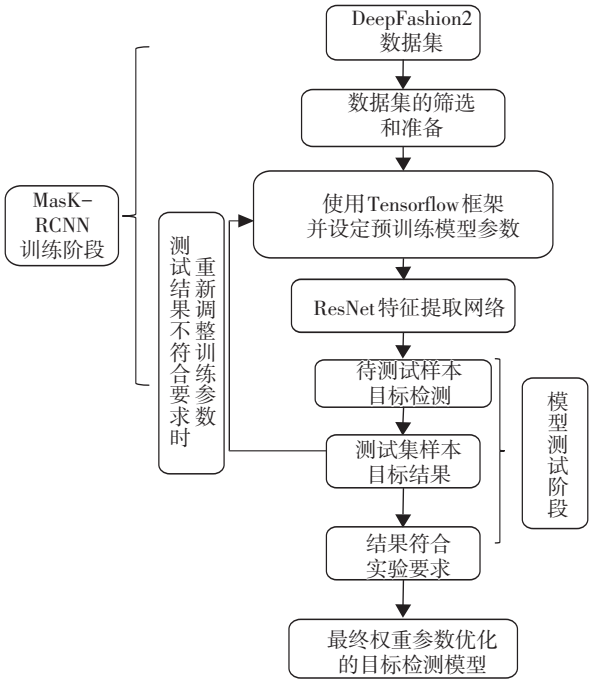

图 2 服装识别检测模型框架

目标检测方法的训练包括 6 个步骤:(1)获取 DeepFashion2 数据集源文件;(2)对训练样本进行预处理,将 DeepFashion2 中的 json 文件转换成 dataset 备用;(3)将 dataset 输入到 ResNet 中,得到对应训练服图像的特征图,对特征图中的每一点设定预定的 RoI,得到多个候选 RoI;(4)将候选的 RoI 送入 RPN 进行二分类(输出为前景或背景)和 BB(bounding box)回归,过滤掉一部分候选的 RoI,对剩下的 RoI 进行 RoIAlign 操作;(5)将得到的 RoI 进行 N 类别分类、BB 回归和 MASK 生成;(6)重复步骤 4 和步骤 5,训练完所有的服装样本后得到最终的检测模型。

服装检测的方法包括两个步骤:(1)利用测试的服装样本对检测模型进行测试,最终得到新样本的检测结果。(2)测试结果不符合要求时重新进行模型的调整与参数训练,并且重新训练模型,若测试结果符合要求,则得到最终的目标检测模型。

2.2 Mask-RCNN 简介与原理

Mask-RCNN 是何凯明等人在 Faster R-CNN[12] 基础上提出的目标实例分割模型。该模型能够有效地检测图像中的目标并为每个实例生成高质量的分割掩码。如图 3 所示,该模型通过在 Faster R-CNN 已存在的 B-box 识别分支旁并行地添加了一个用于预测目标掩码的分支。掩码分支是一个应用到每个 RoI 上的小型 FCN(全卷积网络),能够预测 RoI 中每个像素所属的类别,从而实现准确的实例分割。

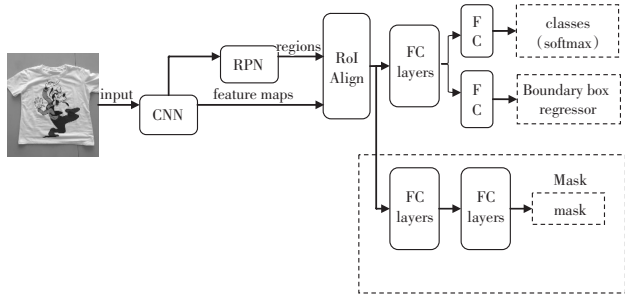

图 3 Mask-RCNN 结构图

Mask-RCNN 的技术要点主要有三个:(1)使用 ResNet+FPN 来提取图像特征。(2)使用 RoIAlign 替代 RoIPooling,引入了一个插值过程,先通过双线性插值到 14×14,再 pooling 到 7×7,解决了仅通过 Pooling 直接采样带来的 Misalignment 对齐问题。(3)每个 RoIAlign 对应 $k \times m^2$ 维度的输出。k 对应类别个数,即输出 k 个掩膜,m 对应池化分辨率 7×7。

2.3 RoI Align 操作

RoI Align 是取消量化操作和整数化操作,并保留小数,使用双线性内插的方法获得坐标为浮点数的图像数值,将整个特征聚集过程转化为一个具有连续性的操作。RoI Align 不是简单的补充出候选区域边界上的坐标点,然后进行池化,而是重新设计。图 4 中虚线框表示的是 5×5 的特征图。虚线部分表示的是 feature map,实线表示 RoI,如图 4 所示将 RoI 切分成 4 个 2×2 的单元格,之后在每个实线的方形区域中选择 4 个采样点,除了这 4 个点还选取离该采样点最近的 4 个特征点,如图 4 中黑色小方格的 4 个顶点,并且通过双线性插值的方法得到每个采样点的像素值;最

后计算每个小区域的像素值，并生成 2×2 的特征图。

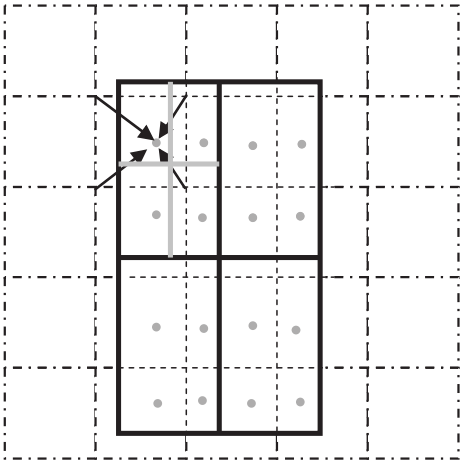

图 4　RoI Align 原理

2.4　FPN(特征金字塔网络)

FPN 的提出是为了实现更好的 feature maps 融合，FPN 采用了自上而下的侧向连接来融合不同尺度的特征，使用 3×3 的卷积来消除混叠现象，来预测不同尺度的特征，不断重复以上的过程，最终得到最佳分辨率。FPN 的优点在于，它可以在不增加计算量的情况下提高多个尺度上小物体的准确性和快速检测能力。图 5 为特征融合原理图。左边的底层特征层通过 1×1 的卷积得到与上一层特征层相同的通道数；上层的特征层通过上采样得到与下一层特征层一样的长和宽再进行相加，从而得到了一个融合好的新的特征层。

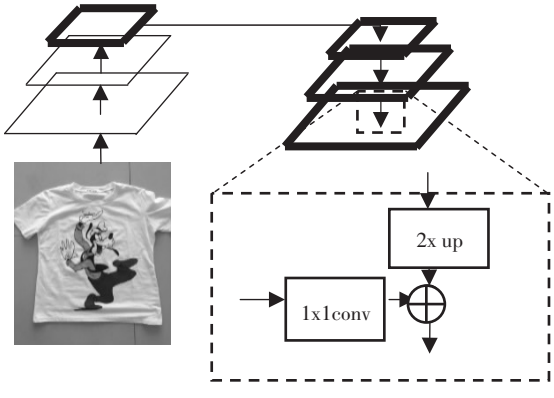

图 5　特征融合原理图

2.5　Loss Function

Mask-RCNN 的损失函数为：

$$L = L_{cls} + L_{box} + L_{mask} \quad (1)$$

式中 L_{cls} 和 L_{box} 与 Fast RCNN 中定义的分类和回归损失相一致，mask 分支对于每一个 RoI 都有 $k \times m^2$ 维度的输出，k 个分辨率为 $m \times m$ 的二值 mask。L_{mask} 为平均二值交叉熵损失。对于一个属于第 k 个类别的 RoI，L_{mask} 仅仅考虑第 k 个 mask。

3　试验部分

3.1　深度学习框架与预训练模型的选取

目前深度学习的学习框架有很多，深度学习的模型需要大量时间和海量训练样本进行训练，在考虑到硬件水平的前提下，试验使用 COCO 2014 数据集作为预训练模型。COCO 2014 数据集拥有 9 000 多张图片，包含了自然图片和生活中常见的物品图片，也有较多的服装图像，因此试验采用迁移学习方法将 COCO 2014 数据集训练得到的权重模型作为服装检测算法模型的预训练模型，在此预训练模型的基础上使用 DeepFashion2 作为训练集再进行样本训练，通过迁移学习的方式不但可提升训练效率，而且能有效地提升检测模型的整体检测精度和模型性能。

3.2　数据集的处理

由于 DeepFashion2 已经得到了各图片的 json 文件，故不需要再使用 LebelMe 进行手动标注，只需将 json 文件转换为 dataset 即可。试验在 Ubuntu18.04、CUDA9.0 环境下进行。试验参数设置如下：初始学习率 0.000 01，每迭代 2 000 次缩小 10 倍。为了使训练效果和模型性能更好，选取服装图片有不同遮挡、不同缩放和不同姿态。

3.3　试验结果与分析

得到训练模型后，使用测试代码进行测试，得到的测试效果如图 6 所示。

图 6　测试效果

试验选取了不同服装姿态的图片，其中矩形框表

示检测服装的位置，矩形框上的数字表示的是属于不同服装类别的概率大小。为了提高检测的准确率，将模型中的矩形框概率的阈值设置为 0.7。一方面减少了网络中确定服装图像边框的计算量，提升计算速度。另一方面防止发生过拟合。经过测试，把 NMS(non maximum suppression)在 RPN 网络的预测阶段在 proposal layer 的阈值设定为 0.7 时，试验结果较好，被标注的服装识别概率均高于 0.764。

试验中也存在识别失败的例子，如图 7 所示。原因可能是服装占比较大或较小、形变较明显、放大不正规和有较大的遮挡等。

由于服装本身的易形变和拍照的角度问题，总体上 Mask-RCNN 识别率较高，达到了预期的效果。为了提高识别准确率，可以将训练集的数据进行筛选。也可以增加训练集数量来提高训练模型的性能。

图 7 服装识别失败例子

4 结语

基于 Mask-RCNN 和 DeepFashion2 的服装检测模型可以更好的识别和分割服装，更好地促进服装识别算法的发展，更好地理解时尚图像。通过使用 Mask-RCNN 对 DeepFashion2 数据集进行测试，得到较为精确的结果，可以通过分割得到的服装进行智能搭配等。这为探索服装形象的多领域学习提供了基础，也为以后进一步优化服装提取算法提供了借鉴。同时，在 DeepFashion2 中引入更多的评估指标，例如深度模型的大小、运行时间和内存消耗，可解释现实场景中的时尚图像。

参考文献：

[1] 中国服装协会.2018－2019 年度中国服装行业发展报告[R].2019.

[2] LECUN Y, KAVUKCUOGLU K, FARABET C. Convolutional networks and applications in vision[C]//Proceedings of 2010 IEEE International Symposium on Circuits and Systems. IEEE, 2010: 253－256.

[3] KIM K J, PARK S M, CHOI Y J. Clothing identification based on edge information[C]//2008 IEEE Asia-Pacific Services Computing Conference. IEEE, 2008: 876－880.

[4] HIDAYATI S C, YOU C W, CHENG W H, *et al*. Learning and recognition of clothing genres from full-body images[J]. IEEE Transactions on Cybernetics, 2017, 48(5): 1 647－1 659.

[5] LIU Z, LUO P, QIU S, *et al*. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 1 096－1 104.

[6] XIAO L, YICHAO X. Exact clothing retrieval approach based on deep neural network[C]//2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference. IEEE, 2016: 396－400.

[7] HE K, GKIOXARI G, DOLLAR P, *et al*. Mask r-cnn[C]//Proceedings of the IEEE International Conference on Computer Vision, 2017: 2 961－2 969.

[8] GE Y, ZHANG R, WANG X, *et al*. DeepFashion2: A versatile benchmark for detection, pose estimation, segmentation and re-Identification of clothing images[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 5 337－5 345.

[9] HADI KIAPOUR M, HAN X, LAZEBNIK S, *et al*. Where to buy it: Matching street clothing photos in online shops[C]//Proceedings of the IEEE International Conference on Computer Vision, 2015: 3 343－3 351.

[10] HUANG J, FERIS R S, CHEN Q, *et al*. Cross-domain image retrieval with a dual attribute-aware ranking network[C]//Proceedings of the IEEE International Conference on Computer Vision, 2015: 1 062－1 070.

[11] ZHENG S, YANG F, KIAPOUR M H, *et al*. Modanet: A large-scale street fashion dataset with polygon annotations[C]//2018 ACM Multimedia Conference on Multimedia Conference, 2018: 1 670－1 678.

[12] REN S, HE K, GIRSHICK R, *et al*. Faster r-cnn: Towards real-time object detection with region proposal networks[C]//Advances in Neural Information Processing Systems, 2015: 91－99.

(下转第 32 页)

Conference Proceedings, 2016.
[3] PRADHAN A K, DAS D, CHATTOPADHYAY R, *et al*. An approach of optimal designing of nonwoven air filter media: Effect offibre fineness[J]. Journal of Industrial Textiles, 2014, 45(6): 1 308－1 321.
[4] JIN G, ZHU C. Artificial neural network modeling for predicting pore size and its distribution for melt blown nonwoven[J]. Sen′i Gakkaishi, 2015, 71(11): 317－322.
[5] 金关秀,祝成炎.孔隙形状对熔喷非织造布过滤品质的影响[J].上海纺织科技,2018, 46(11):15－18.

Modeling of 3D Structure of Mask Filter Material

ZHU Guo-qing[1], HOU Shuang[2], DONG Han-rui[2], MO Xiang-jie[3]
(1. Suzhou Institute of Fiber Inspection, Suzhou 215128, China;
2. Xi′an Polytechnic University, Xi ′ an 710048, China;
3. Guangxi Road Construction Engineering Group Co., Ltd., Laibin 546111, China)

Abstract: Nonwoven materials contained a large number of micro three-dimensional curved channels, which was widely used in the filter layer of masks. Structure determined the performance of the material. In order to establish the mesoscopic three-dimensional structure of the nonwoven fabric, the three-dimensional point clouds of the nonwoven fabric were obtained through sequential 50-layer images collected by the automatic microscope and the software developed by research group. The three-dimensional model of the fabric was visualized in Geomagic Studio software. Three-dimensional reconstructed images of nonwovens showed that the 3D modeling method could simulate the three-dimensional packing structure of fibers in nonwoven fabrics.

Key words: mask; nonwoven web; 3D image reconstruction

(上接第 24 页)

Clothing Recognition and Segmentation Based on Mask-RCNN

ZHANG Ze-kun[1], ZHANG Hai-bo[2,3,*]
(1. Information Center, Beijing Institute of Fashion Technology, Beijing 100029, China;
2. Beijing Engineering Research Center of Textile Nanofiber, Beijing Key Laboratory of Clothing Materials R &D and Assessment, Beijing Institute of Fashion Technology, Beijing 100029, China;
3.Library, Beijing Institute of Fashion Technology, Beijing 100029, China)

Abstract: A method for clothing recognition and segmentation based on Mask-RCNN and data set DeepFashion2 was proposed. Mask-RCNN-based clothing recognition and segmentation was based on the idea of convolutional neural networks. Through the multi-threaded iterative training in the deep learning framework, after obtaining the target features in the ResNet network, the features were input differently through RPN and RoI Align. The branches were connected. Finally the target detection model with optimized weights was obtained. In clothing image of different scenes, the model could identify and segment the garment more quickly and accurately.

Key words: clothing recognition; clothing segmentation; Mask-RCNN; ResNet; Deepfashion2; TensorFlow