文章编号：2095-2163(2019)03-0247-09 中图分类号：TP391 文献标志码：A

基于深度学习的手分割算法研究

向 杰[1]，卜 巍[2]，邬向前[1]

（1 哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001；2 哈尔滨工业大学 媒体技术与艺术学院，哈尔滨 150001）

摘 要：第一视角的人手分割在人机交互、虚拟现实方面具有非常重要的应用价值，但是由于图像中人手区域占比较大，精确的人手分割仍然是一个很具有挑战性的问题。本文提出一种基于深度学习的手部分割算法，利用卷积神经网络强大的特征提取能力提取人手图像的特征，模仿人类视觉注意力机制提出 Attention 模块为特征图中的不同区域赋予更具辨别性的权值，同时为了能有效地提取不同尺度物体的特征，设计空洞卷积 DCB 模块在同一尺度大小的特征图上提取不同尺度特征。在 3 个人手数据集上的实验结果表明本文提出的算法能够有效地分割出手部区域并超越了其它的算法，达到了最优的分割效果。

关键词：手部分割；深度学习；注意力机制；空洞卷积

Hand segmentation research based on deep learning

XIANG Jie[1], BU Wei[2], WU Xiangqian[1]

（1 School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China;
2 School of Media Technology and Art, Harbin Institute of Technology, Harbin 150001, China）

【Abstract】 Hand segmentation in egocentric views has very important application value in human-computer interaction and virtual reality. But precise hand segmentation is still a very challenging problem because the appearance of hands can vary greatly in images. This paper proposes a hand segmentation algorithm based on deep learning. The feature extraction ability of Convolution Neural Network is used to extract the features of hand images. Imitating the human visual attention mechanism, the research proposes Attention module to assign more discriminative weights to different regions in the feature map. And to extract the features of objects of different scales effectively, the paper also designs Dilated Convolution Block module to extract different scale features from the feature map of the same size. Experiment results on three hand datasets show that the proposed method can segment hands efficiently and outperform other methods, and achieve the state-of-the-art.

【Key words】 hand segmentation; deep learning; attention mechanism; dilated convolution

0 引 言

Google Class、GoPro 和 Narrative Clip 等可穿戴设备的日益普及，使得计算机视觉中以自我为中心的第一视角研究成为一个快速增长的领域。可穿戴设备产生大量的数据，这使得自动分析其记录的内容（例如，浏览、搜索和可视化）、描述生活记录中的事件、识别日常生活活动等成为一种需要。在以自我为中心的第一视角视频中，大部分的工作都涉及到理解相机佩戴者的活动和行为。在本文中，研究关注的是以自我为中心的第一视角视频中一个非常关键的实体：手。在人们的日常生活中，手是无处不在的。手的姿势和配置告诉人们计划做什么或者人们注意到了什么。因此，手的检测、分割和跟踪是以自我为中心的视觉中的基本问题，在机器人、人机交互、计算机视觉、增强现实等领域有着广泛的应用。在以自我为中心的视频中提取手部区域是理解精细运动的关键一步，例如手-对象操作和手眼协调。

本文着重在现实的日常环境中解决以自我为中心的第一视角的视频中像素级手分割的任务。大量的研究在第三视角或监控视频中解决了这个问题，然而，在第一视角视频中，对这个问题的研究相对较少。本文计划通过设计基于深度学习的语义分割算法对第一视角视频中的手进行分割。

本次研究基于 Bambach 等人[1] 提出的 Egohands 数据集，该数据集对 2 个有交互动作的参与者的手进行了像素级的标注。据分析所知，该数据集是唯一的聚焦于人与人之间交互动作的、第一

基金项目：国家自然科学基金（61472102,61672194）；山东省自然科学基金（ZR2016FM04）。

作者简介：向 杰（1993 - ），男，硕士研究生，主要研究方向：数字图像处理、计算机视觉、深度学习等；卜 巍（1977 - ），女，博士，副教授，主要研究方向：数字媒体技术、数字图像处理、医学图像分析等；邬向前（1973 - ），男，博士，教授，博士生导师，主要研究方向：数字图像处理、模式识别、生物特征识别等。

通讯作者：卜 巍 Email：buwei@ hit.edu.cn

收稿日期：2018-06-21

视角的、并具有像素级标注的人手数据集,故而本文将基于该数据集来验证所提出的语义分割算法。同样,文中也将在 GTEA[2] 数据集及其最新扩展的 EGTEA 数据集上验证了本文提出的算法。

本文的主要贡献总结如下:

(1)提出了一个针对手分割的基于深度学习的语义分割算法,利用卷积神经网络(CNN)强大的自动提取特征的能力来自动提取手部特征,从而能够端到端地训练语义分割网络。

(2)模仿人类视觉机制,提出了 Attention 网络模块,增强对手分割贡献大的特征的权重,减小贡献小的特征的权重,使得网络更具有特征辨别性。

(3)提出空洞卷积 DCB 模块,在同一尺度的特征图上提取不同尺度的特征,对不同大小的图片中的目标、即手的分割更加精确。

(4)提出的针对手分割的语义分割算法在 3 个数据集,即 Egohands、GTEA 和 EGTEA 上取得了超越先前算法的效果,获得了当前最优的分割效果。

1 相关工作

目前已有一些基于以自我为中心的第一视角的手分割研究。Ren 等人[3] 以及 Fathi 等人[4] 提出一种查找手部区域具有不规则光流模式的方法来分割手,研究中假设在日常生活中以自我为中心的第一视角视频中,当人与人或其它对象交互时,背景为静态的,具有规则的光流模式,手作为前景区域具有动态的不规则的光流模式,利用手部区域不规则的光流模式来进行手分割。Li 等人[5] 假设视频中没有人的交互动作存在,认为视频中的所有手都属于以自我为中心的观看者,提出一种利用场景级特征探针为每个环境选择最佳局部颜色特征的光照感知方法来进行手分割。然而这种假设并不能概括生活中所有的人手活动。Lee 等人[6] 提出一种在第一视角的视频中检测分割交互中的手的方法,同时也提出了一种概率图模型,利用空间排列来消除手部类型的歧义,即区分是观察者的手、还是交互者的手。然而,此类方法只考虑了实验室条件下的交互动作,对于具有复杂背景情形下的交互动作却并未纳入研究范畴。

更加接近本文工作的研究是 Bambach 等人[1] 提出的,即提出了一种基于肤色检测的方法,该方法首先生成一组可能包含手区域的包围框,然后使用 CNN 检测识别手,最后使用 GrabCut[7] 方法对其进行分割,Aisha 等人[8] 微调当下最好的基于自然图像的语义分割网络 RefineNet[9] 用于手分割,获得了目前最优的结果。

除了基于第一视角的手分割外,基于第三视角的手检测分割也已可见到相应的研究。比如,Mittal 等人[10] 利用可变形部件模型 DPM[11] 和基于肤色的启发式先验进行手的定位检测。Zimmermann 等人[12] 基于单张 RGB 图像进行手的检测和姿势估计。

2 算法设计研究

2.1 网络结构

本文把手部区域分割视为一个语义分割问题,也就是像素级别的分割,是一个密集预测的问题,目标是将属于手部区域的像素和属于背景的像素分离开来,即判定图片中每一个像素是属于手部区域、还是非手部区域。

本文针对手部区域分割设计的语义分割网络如图 1 所示。该网络由 3 部分组成:主干网络(Backbone),空洞卷积模块(Dilated Convolutional Block,DCB),空间通道压缩扩张(Spatial Channel Squeeze-and-Excitation,SCSE)模块。文中,将对此做出如下研究分述。

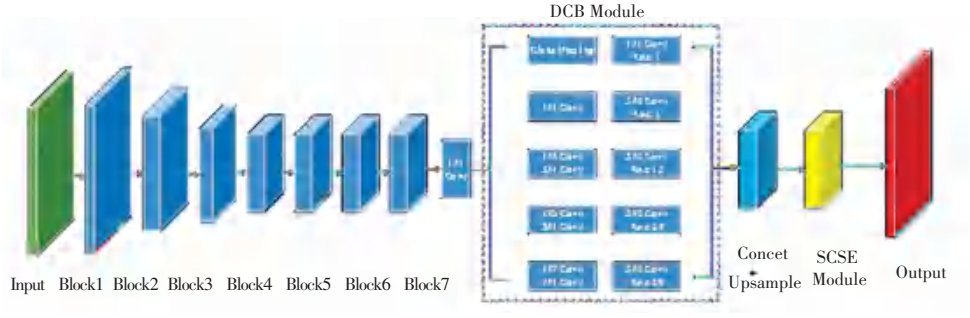


图 1 语义分割网络

Fig. 1 Semantic segmentation networks

2.2　主干网络

研究采用的主干网络是 38 层的 Wide ResNet[13]。Wide ResNet 由 7 个 Block 组成。其中，第 1 个 Block 单元由卷积核为 3×3 的卷积层组成，第 2～5 个 Block 由残差单元（Residual Unit，RU）组成，残差单元由 2 个级联的 3×3 卷积层构成，并按照 ResNet 的残差结构添加跳跃连接（Skip Connection）来学习残差映射。网络的最后 2 个 Block6、Block7 由级联的卷积核，分别为 1×1、3×3、1×1 的卷积层构成，目的是减少网络参数量。Wide ResNet 采用网络加宽，即增加每一层网络的特征通道的方式来提高网络性能，在图像分类上获得了很好的性能，并且参数量也稳定地控制在合理的范围内。除此之外，在语义分割任务上将 38 层的 Wide ResNet 改变为全卷积 CNN 也获得了很好的结果。

2.3　DCB 模块

DCB 模块的作用是多尺度提取特征图的语义特征。深度神经网络学习通过组合低层特征形成更加抽象的高层特征表示全局属性或目标类别，以发现数据的分布式特征表示。网络低层学习到的一般是物体的角点、边缘、局部轮廓等特征，网络高层学习到的一般是物体的抽象的表示，因此，结合网络低层特征和高层特征或者在同一特征图尺度上获取不同尺度的特征对于提高网络对不同大小的图片中的目标、即手的的分割更加精确。

研究提出针对语义分割的空洞卷积模块（Dilated Convolutional Block，DCB），一个在同一尺度的特征图上提取不同尺度特征的多分支卷积模块。DCB 的内部结构可分为 2 个组件：多分支卷积层以及随后的空洞卷积层。其中，多分支卷积层由 5 个子分支组成，分别是 Global Pooling 分支、1×1 卷积分支，1×3 和 3×1 分支、1×5 和 5×1 卷积分支、1×7 和 7×1 卷积分支，除 Global Pooling 分支外，其余 4 个分支其后都级联一个不同比率的空洞卷积层，本文选择的空洞卷积层的比率分别是 1、12、24、36。DCB 模块能够有效地提高特征提取的效率，针对同一尺度的特征图，不同大小的卷积核可以多尺度地提取物体特征，空洞卷积可以有效增大卷积核感受野，这对于语义分割任务十分重要。

2.4　SCSE 模块

研究提出的 SCSE 模块，又可以称为空间通道 Attention 模块，如图 2 所示。空间通道压缩扩张模块由空间 Attention 子模块和通道 Attention 子模块构成，分别对应图 2 中的 2 个分支。对于空间 Attention 模块，使用卷积核大小为 1×1，步长为 1 的卷积层与输入的大小为 $H \times W \times C$ 的特征图进行卷积操作，输出大小为 $H \times W \times 1$ 的特征图，将特征图通道方向压缩为一维，再将输出的特征图经过 Sigmoid 层使得特征图的激活值范围为[0，1]，最后将输出的特征图与原输入特征图做点乘得到大小为 $H \times W \times C$ 的特征图，空间 Attention 为特征图中空间位置的不同点重新赋予了不同的权重值，使得目标相关的空间位置点得到更大的权值，减小不相关的空间位置点的权重。对于通道 Attention 模块，将输入的大小为 $H \times W \times C$ 的输入特征图首先通过全局池化（Global Pooling）层获得大小为 $1 \times 1 \times C$ 的特征图，将特征图的空间方向、即长度方向和宽度方向压缩为一维，接着通过第一层全连接（FC）层将特征图变为 $1 \times 1 \times (C/r)$ 大小，其中 r 为缩放参数，本文选择的参数 r 值为 8，然后通过第二层 FC 层将特征图恢复到 $1 \times 1 \times C$ 大小，并经过 Sigmoid 层使得特征图的激活值范围为[0，1]，最后将输出的特征图与原输入特征图做点乘得到大小为 $H \times W \times C$ 的特征图，通道 Attention 通过参数来为每个特征通道生成权重，其中参数被学习用来显式地建模特征通道间的相关性，将经过 Sigmoid 层输出的权重与原输入特征图相乘可以看作是对特征图不同通道的重新赋权，使得目标相关的通道权重得以提升，不相关的通道权重得以抑制。

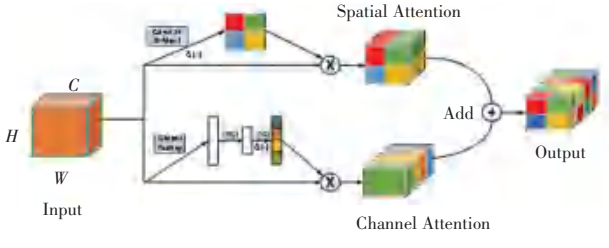


图 2　空间通道压缩扩张模块

Fig. 2　Spatial Channel Squeeze-and-Excitation module

根据空间 Attention 子模块和通道 Attention 子模块的不同组合方式，本文还提出了 2 种形式的 SCSE 模块，依次命名为通道优先空间通道压缩扩张（Channel first Spatial Channel Squeeze - and - Excitation，CSCSE）模块和空间优先空间通道压缩扩张（Spatial first Spatial Channel Squeeze - and - Excitation，SSCSE）模块，分别如图 3 和图 4 所示。

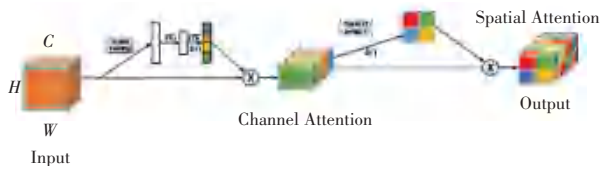


图 3 通道优先空间通道压缩扩张模块

Fig. 3 Channel first Spatial Channel Squeeze – and – Excitation module

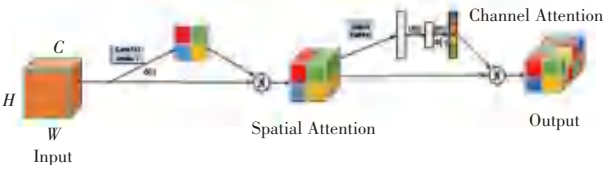


图 4 空间优先空间通道压缩扩张模块

Fig. 4 Spatial first Spatial Channel Squeeze – and – Excitation module

3 网络训练和测试

本次研究的任务是训练一个 CNN 语义分割模型,该模型可以自动地在一张人手图片中分割出人手区域的 mask 图像。

相比于图像识别只需要图像级别的类别标签即可,语义分割任务则需要精细的像素级标注的 mask 图像作为标签,而标记图像的 mask 标签,往往耗时耗力,因此,在语义分割任务中,有标签的标注好的图像数量很少。为了能够较好地训练语义分割网络,数据增强操作必不可少,并且一般采用在 ImageNet[14] 图像识别数据集上预训练的模型参数作为主干网络的初始化参数。ImageNet 是一个大规模的图像分类数据库,包含着数目可观的带有图像类别标签的自然图像,深度学习图像分类网络往往在 ImageNet 上训练测试,因为 ImageNet 数据库数据采集自自然环境,图像覆盖面广且类别宽泛,能够有效地验证分类模型是否性能良好。目前,各个常见的深度神经网络模型都有基于 ImageNet 预训练的模型。

图像分类网络最后的网络层一般是用于分类的全连接 FC 层,全连接层的权重矩阵是固定的,即每一层特征图(feature map)的输入必须是固定大小(即与权重矩阵正好可以相乘的大小),所以网络最开始的输入图片尺寸必须固定,才能保证传送到全连接层的特征图的大小与全连接层的权重矩阵相匹配。全连接层可以看作是卷积核完全覆盖特征图的特殊的卷积层。目前的语义分割网络都是基于全卷积神经网络 FCN,即网络中不存在全连接层,FCN 可以接受不同大小的图片作为输入。

研究将在 ImageNet 上预训练的用于图像分类的 Wide ResNet 作为语义分割网络的主干网络,首先需要将其转换为全卷积神经网络:将全局池化层(Global Pooling)和最后一层用于分类的全连接层去掉。语义分割的目的是要密集预测图片中每一个像素点所属类别,为了尽可能多地捕获特征图中的低层局部信息和高层语义信息,本文将图像的下采样次数设定为 3 次,即经过主干网络输出的特征图大小是原输入图片的 1/8 大小。同时,为了使卷积核能够有效获取更大范围特征,本文采用空洞卷积的方法来扩大卷积核的感受野,其中,主干网络第 5、第 6、第 7 个 Block 分别使用比率为 2、4、8 的空洞卷积来扩大卷积核的感受野。

在测试阶段,给定一个未知测试图片,经过训练好的语义分割网络,分割出图片中的人手区域 mask。

4 实验

研究拟在 3 个数据集,诸如 EgoHands 数据集、Georgia Tech Egocentric Activity(GTEA)数据集和 Extended Georgia Tech Egocentric Activity(EGTEA)数据集上分别进行语义分割网络的训练、验证和测试。

本节首先介绍使用的 3 个数据集,并详细解读了数据集的构成和训练、验证、测试数据集的划分,接着探究了本文使用的评测标准,最后则剖析论述了各个数据集的训练过程和测试结果。对此可做阐释分述如下。

4.1 数据集介绍

(1) EgoHands 数据集。EgoHands 数据集是一个收录人与人之间交互动作的数据集,包含 48 个使用 Google Class 记录的视频片断,每一个视频片段记录 2 个演示者玩拼图(puzzle)、拼卡片(cards)、玩层叠游戏(jenga)或者下国际象棋(chess)的手部交互动作,这些视频数据是在办公室、庭院和卧室三种不同的环境下拍摄。数据集里面汇集了超过 15 000 个人手实例,每一个视频包含 100 张手工精细标注的人手区域 mask 图片,一共有 4 800 张标注的人手 mask 图片。发布该数据集的作者按照 75%、8%、17%的比例将 4 800 张图片划分为训练集、验证集和测试集。本文也遵循这一划分比例。

(2) GTEA 数据集。GTEA 数据集采集了记录日常生活中的 7 种活动的视频,视频采集在同一环境条件下进行,没有记录人与人之间的交互动作,在静态光照条件下采集数据集。分割数据集涉及到人体的手及手臂区域,一共包含 663 张人工精细标注的图片数据。本文根据数据集作者的数据集进行划

分，将数据集中的 367 张图片作为训练集、92 张图片作为验证集、204 张图片作为测试集。

(3) EGTEA 数据集。EGTEA 数据集是 GTEA 数据集的最新扩增版本，包含 28 h 的烹饪视频片段，数据集还提供了相关视频片段的音频、人体动作标注和跟踪信息用于其它视觉任务。数据集精细标注了 13 847 张人手的 mask 图片，共包含 15 176 个人手实例。由于数据集作者未能提供关于人手分割图片数据集的训练、验证、测试数据划分。本文按照约 7:1:2 的比例划分带标签的手部图像数据集为训练集、验证集和测试集，其中，训练集为 7 906 张图片，验证集为 1 844 张图片，测试集为 4 097 张图片。

4.2　评测标准

语义分割中通常使用许多标准来衡量算法的性能。为了便于解释，假设如下共有 $k+1$ 个类别(从 L_0 到 L_k 其中包含一个背景类)，p_{ii} 表示本属于 i 类且预测为 i 类的像素数量，即真正预测正确的像素数量；p_{ij} 表示本属于 i 类但被预测为 j 类的像素数量，即假正；p_{ji} 表示本属于 j 类但被预测为 i 类的像素数量，即假负。二分类分割常用的评测标准可综合表述如下。

(1) 平均交并比(mean Intersection over Union, *mIOU*)：语义分割的标准度量。计算 2 个集合的交集和并集之比，这 2 个集合为真实值(ground truth，标签值)和预测值(predicted segmentation)。在每个类上计算 *IOU*，再取平均值。研究推得数学定义公式如下：

$$mIOU=\frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}},\quad(1)$$

(2) 平均召回率(mean Recall, *mRec*)：预测像素为 i 类且原像素属于 i 类的像素数量与所有原像素为 i 类的像素数量的比值，其中，原像素为 i 类的像素包括预测为 i 类且原像素属于 i 类和本属于 i 类但被预测为 j 类的像素。研究推得数学定义公式如下：

$$mRec=\frac{\sum_{i=0}^{k}p_{ii}}{\sum_{i=0}^{k}p_{ii}+\sum_{j=0}^{k}p_{ij}},\quad(2)$$

(3) 平均精确率(mean Precision, *mPrec*)：预测像素为 i 类且原像素为 i 类的像素数量与所有预测为 i 类的像素数量的比值，其中，原像素为 i 类的像素包括预测为 i 类且原像素属于 i 类和本属于 j 类但被预测为 i 类的像素。研究推得数学定义公式如下：

$$mPrec=\frac{\sum_{i=0}^{k}p_{ii}}{\sum_{i=0}^{k}p_{ii}+\sum_{j=0}^{k}p_{ji}},\quad(3)$$

(4) 像素精度(Pixel Accuracy, *PA*)：标记正确的像素占总像素的比例。研究推得数学定义公式如下：

$$PA=\frac{\sum_{i=0}^{k}p_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k}p_{ij}}.\quad(4)$$

在二分类分割问题中，上述 4 种评测标准都能有效地评测算法的性能。本文的手分割是一个二分类分割任务，因此，研究即将以上述 4 种评测标准用于算法性能的研究考证。

4.3　实验与结果分析

针对前文探讨论述的 4 个数据集，本节将分别给出其实验结果及结果分析，详情参见如下。

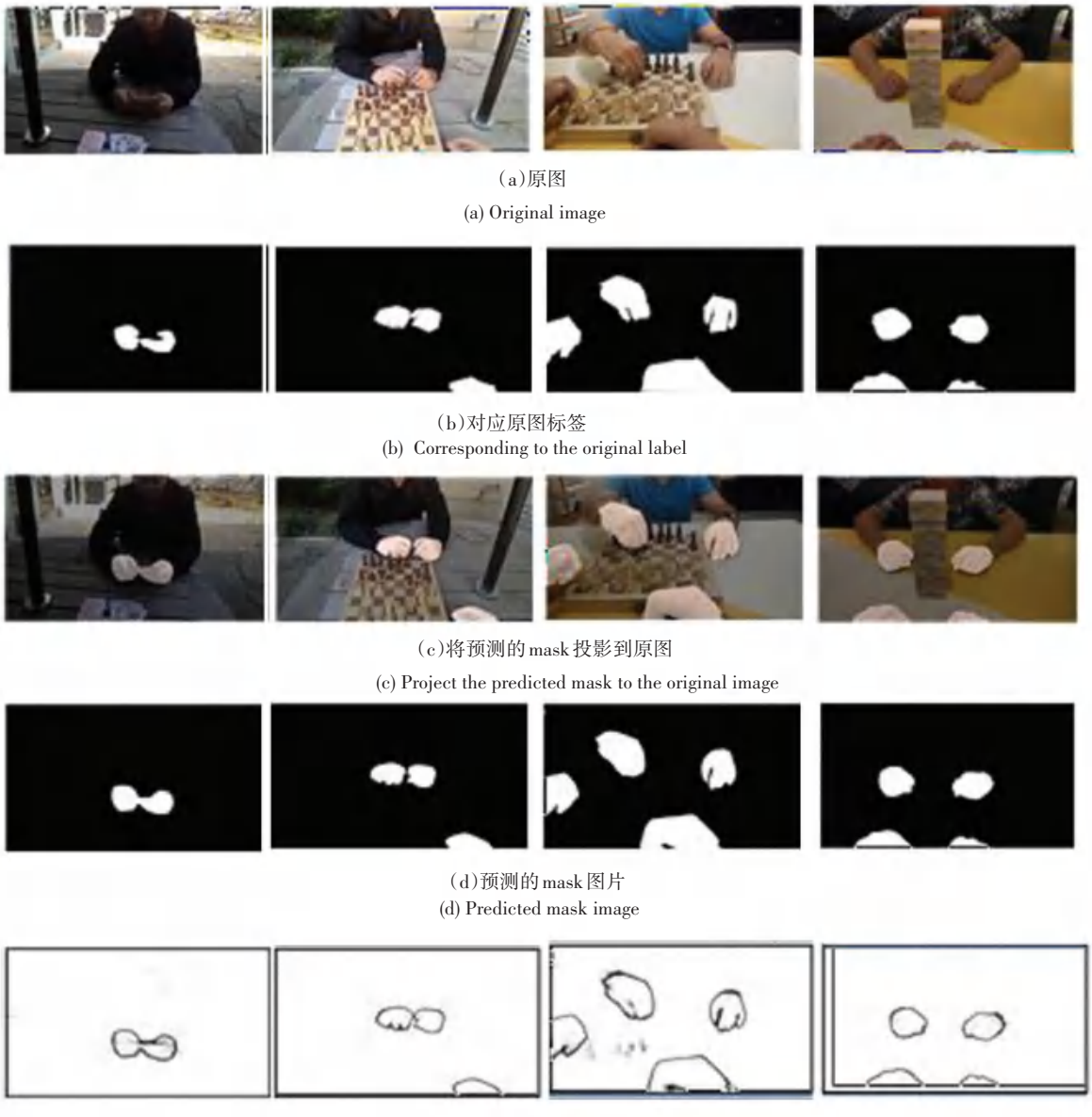
(1) EgoHands 数据集。研究中根据 EgoHands 数据集作者的数据划分来训练验证模型，并在测试集上测试模型。为了验证设计的 3 种 Attention 结构，本节分别训练不包含 Attention(noAttention)结构和包含 3 种不同 Attention(CSCSE、SSCSE、SCSE)结构的模型，并分别测试其模型效果，给出各评测指标的定量评测结果。同时，与前人在 EgoHands 数据集上的分割结果在各个评测指标上进行了对比，最终对比结果见表 1。

表 1　Egohands 数据集实验结果对比

Tab. 1　Comparison of experiment results of Egohands dataset

算法(algorithm)	平均交并比(*mIOU*)	平均召回率(*mRec*)	平均精确率(*mPrec*)	像素精度(*PA*)
Bambach et al[1]	0.556	N/A	N/A	N/A
Aisha et al[8]	0.814	0.919	0.879	N/A
The proposed(noAttention)	0.856	0.924	0.869	0.969
The proposed(CSCSE)	0.865	0.933	0.889	0.969
The proposed(SSCSE)	0.867	0.923	0.863	0.970
The proposed(SCSE)	**0.869**	**0.937**	**0.897**	**0.972**

从表 1 可以看出,分割模型中有 Attention 结构比没有 Attention 结构好,其中,具有 SCSE 结构的 Attention 模块在各个评测指标上都获得了最好的性能。因此,在下文的其它数据集的实验中,模型都使用具有 SCSE 结构的 Attention 模块。部分 EgoHands 数据集测试集可视化结果如图 5 所示。



(a)原图

(a) Original image

(b)对应原图标签

(b) Corresponding to the original label

(c)将预测的 mask 投影到原图

(c) Project the predicted mask to the original image

(d)预测的 mask 图片

(d) Predicted mask image

(e)预测的 mask 的边缘图

(e) Predicted mask edge map

图 5 EgoHands 数据集测试结果

Fig. 5 Test results of EgoHands dataset

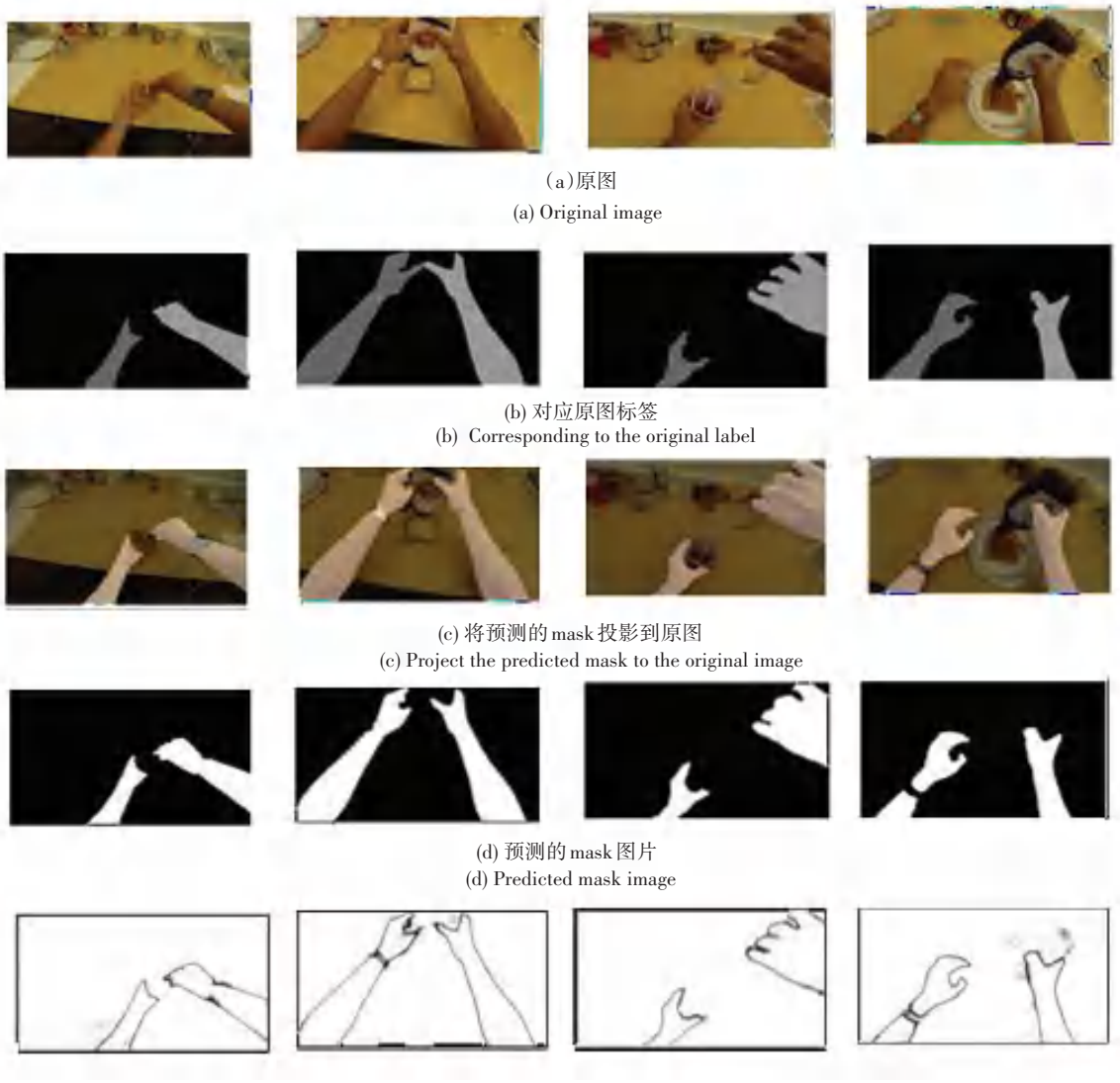
(2) GTEA 数据集。研究中按照数据集作者的数据集划分方式划分训练集、验证集和测试集合,在训练集上训练模型,而每训练达到一个 *epoch* 后则在验证集上验证模型,最后,用最终训练完成的模型在测试集上测试模型。与 Aisha 等人在 GTEA 数据集上的算法性能进行比较,实验结果见表 2。

表 2 GTEA 数据集实验结果对比

Tab. 2 Comparison of experimental results of GTEA dataset

算法(algorithm)	平均交并比(*mIOU*)	平均召回率(*mRec*)	平均精确率(*mPrec*)	像素精度(*PA*)
Aisha et al[8]	0.821	0.869	0.928	N/A
The proposed	**0.855**	**0.877**	**0.957**	**0.976**

从表 2 可以看出，本文提出的分割算法在各个性能的对照上都较 Aisha 等人的算法好，部分 GTEA 数据集测试集可视化结果如图 6 所示。



(a)原图
(a) Original image

(b) 对应原图标签
(b) Corresponding to the original label

(c) 将预测的 mask 投影到原图
(c) Project the predicted mask to the original image

(d) 预测的 mask 图片
(d) Predicted mask image

(e) 预测的 mask 的边缘图
(e) Predicted mask edge map

图 6　GTEA 数据集测试结果

Fig. 6　Test results of GTEA dataset

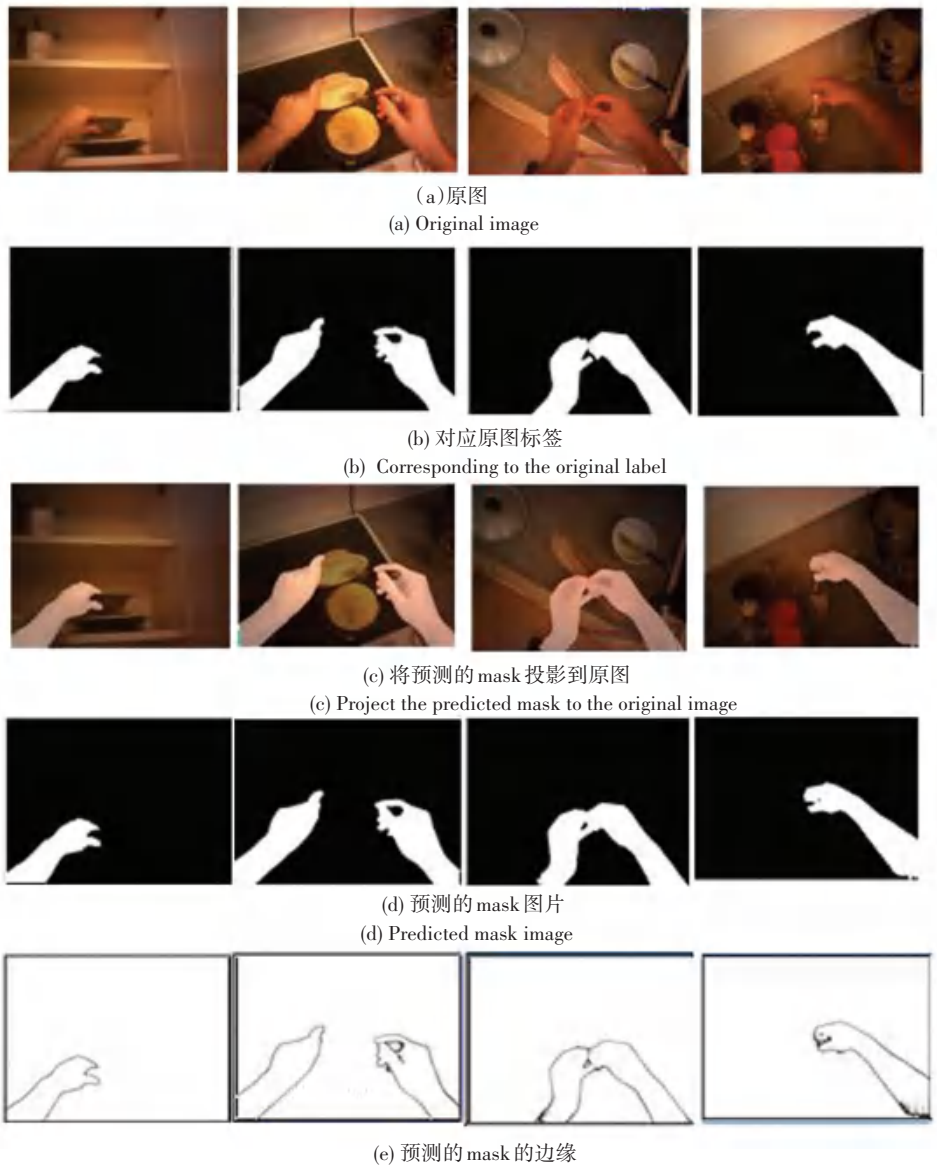
(3) EGTEA 数据集。考虑到该数据集是 GTEA 数据集的最新扩增版本，而做此研究时仍尚未有基于该数据集的研究结果可供对比，因此研究中按照上文所述 EGTEA 数据集的数据划分方法划分训练集、验证集和测试集，并列出本文算法在该数据集上各个评测指标的结果，具体见表 3。

表 3　EGTEA 数据集评测结果

Tab. 3　Evaluation results of EGTEA dataset

算法(algorithm)	平均交并比(*mIOU*)	平均召回率(*mRec*)	平均精确率(*mPrec*)	像素精度(*PA*)
The proposed	**0.932**	**0.906**	**0.963**	**0.990**

从表 3 可以看出，本文提出的分割算法在各个评测指标上都获得了较好的结果。部分 EGTEA 数据集测试集可视化结果如图 7 所示。



(a)原图
(a) Original image

(b) 对应原图标签
(b) Corresponding to the original label

(c) 将预测的 mask 投影到原图
(c) Project the predicted mask to the original image

(d) 预测的 mask 图片
(d) Predicted mask image

(e) 预测的 mask 的边缘
(e) Predicted mask edge map

图 7 EGTEA 数据集测试结果

Fig. 7 Test results of EGTEA dataset

5 结束语

本文将对以自我为中心的第一视角视频中的手分割视为一个语义分割任务,设计了一个基于深度学习的针对手分割的语义分割网络。在该网络中,研究提出 DCB 模块,该模块能够在相当程度上提升特征提取的效率,针对同一尺度的特征图,不同大小的卷积核可以多尺度地提取物体特征,同时空洞卷积可以有效增大卷积核感受野,能够较为成功地提取到图像中不同大小的目标、即手的特征。另外,研究模仿人类视觉注意力机制提出 Attention 模块,在特征图空间和通道方向上分别计算各激活值的概率分布,并与原特征图相乘,为特征图的激活值重新赋权,使得目标相关的特征权重得以提升,不相关的特征权重得以抑制。进一步地,研究还在 EgoHands、GTEA 和 EGTEA 这 3 个相关数据集上分别进行训练测试,获得了当前最优的结果,从各个数据集的测试集结果可以看出,本文提出的语义分割算法可以很好地实现手分割。

参考文献

[1] BAMBACH S, LEE S, CRANDALL D J, et al. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions[C]// IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015: 1949-1957.

[2] LI Yin, YE Zhefan, REHG J M, et al. Delving into egocentric actions[C]// IEEE International Conference on Computer Vision and Pattern Recognition. Portland, OR, USA:IEEE,2015: 287-295.

[3] REN Xiaofeng, MALIK J. Tracking as repeated figure/ground segmentation[C]// IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR ′07. Minneapolis, Minnesota, USA:IEEE, 2007:1-8.

[4] FATHI A, REN Xiaofeng, REHG J M. Learning to recognize objects in egocentric activities [C]// IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI:IEEE Computer Society, 2011:3281-3288.

[5] LI Cheng, KITANI K M. Pixel-level hand detection in egocentric videos[C]//2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, Oregon, USA: IEEE, 2013:3570-3577.

[6] LEE S, BAMBACH S, CRANDALL D J, et al. This hand is my hand: A probabilistic approach to hand disambiguation in egocentric video[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Columbus, OH, USA:IEEE, 2014:557-564.

[7] TANG M, GORELICK L, VEKSLER O, et al. GrabCut in one cut[C]// IEEE International Conference on Computer Vision. Washington, DC, USA:IEEE, 2013:1769-1776.

[8] AISHA U, BORJI A. Analysis of hand segmentation in the wild [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018). Salt Lake City, UT: IEEE, 2018:1-10.

[9] LIN Guosheng, MILAN A, SHEN Chunhua, et al. Refinenet: Multi-path refinement networks with identity mappings for high-resolution semantic segmentation[J]. arXiv preprint arXiv:1611.06612, 2016.

[10] MITTAL A, ZISSERMAN A, TORR P. Hand detection using multiple proposals[C]// British Machine Vision Conference. Dundee:University of Dundee, 2011:75.1-75.11.

[11] FELZENSZWALB P F, GIRSHICK R B, MCALLESTER D, et al. Object detection with discriminatively trained part - based models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 32(9):1627-1645.

[12] ZIMMERMANN C, BROX T. Learning to estimate 3D hand pose from single RGB images[J]. arXiv preprint arXiv:1705.01389v3, 2017.

[13] WU Zifeng, SHEN Chunhua, HENGEL A V D. Wider or deeper: Revisiting the ResNet model for visual recognition[J]. arXiv preprint arXiv:1611.10080,2016.

[14] RUSSAKOVSKY O, DENG Jia, SU Hao, et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3):211-252.

(上接第 246 页)

[4] GAUTRON P, BOUATOUCH K, PATTANAIK S. Temporal radiance caching[J]. IEEE Transactions on Visualization and Computer Graphics, 2007,13(5):891-901.

[5]孙鑫,周昆,石教英. 可变材质的实时全局光照明绘制[J]. 软件学报, 2008,19(4):1004-1015.

[6] PAN Minghao, WANG Rui, LIU Xinguo, et al. Precomputed radiance transfer field for rendering inter reflections in dynamic scenes[J]. Computer graphics Forum,2007,26(3):485-493.

[7] WANG Rui, WANG Kun, PAN Minghao, et al. An efficient GPU-based approach for interactive global illumination[C]//ACM Transactions on Graphics, 2009,28(3):91(1-8).

[8] VIITANEN T, KOSKELA M, IMMONEN K, et al. Sparse sampling for real-time ray tracing [C]// Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 1: GRAPP. Funchal, Madeira, Portugal:dblp, 2018:295-302.

[9] SHIVARAJU S,PUDUR G. Splay thread cooperation on ray tracing as a load balancing technique in speculative parallelism and GPGPU[J]. The International Arab Journal of Information Technology, 2018, 15(1):167-176.

[10] LEE W J, HWANG S J, SHIN Y, et al. Fast stereoscopic rendering on mobile ray tracing GPU for virtual reality applications[C]//IEEE International Conference on Consumer Electronics. Las Vegas, NV, USA :IEEE,2017:355-357.

[11] PÉRARD-GAYOT A, KALOJANOV J, SLUSALLEK P. GPU ray tracing using irregular grids[J]. Computer Graphics Forum. 2017,36(2):477-486.

[12]郑越. 大规模场景渲染下的分布式光线跟踪算法研究[D]. 长沙:湖南大学, 2014.

[13] CALAZAN R M, RODRÍGUEZ O C, NEDJAH N. Parallel ray tracing for underwater acoustic predictions[M]// GERVASI O. Computational Science and Its Applications-ICCSA 2017. ICCSA 2017. Lecture Notes in Computer Science. Cham :Springer,2017, 10404:43-55.

[14] MONIL M A H. Stingray - HPC: A scalable parallel seismic raytracing system [C]//The 26th Euromicro International Conference on Parallel, Distributed and Network - Based Processing. Cambridge, UK:[s.n.],2018:1-11.

[15] RITSCHEL T, ENGELHARDT T, GROSCH T, et al. Micro-rendering for scalable, parallel final gathering [J]. Acm Transactions on Graphics, 2009, 28(5):132.

[16] GUZEK K, NAPIERALSKI P. Efficient rendering of caustics with streamed photon mapping[J]. Bulletin of the Polish Academy of Sciences, Technical Sciences, 2017, 65(3):361-368.

[17]陈纯毅, 杨华民, 李文辉,等. 基于帧间虚拟点光源重用的动态场景间接光照近似求解算法[J]. 吉林大学学报(工学版), 2013, 43(5):1352-1358.

[18]袁璐. 基于立即辐射度的实时全局光照算法[J]. 现代计算机(专业版), 2018(2):63-66.

[19] KAPLANYAN A, DACHSBACHER C. Cascaded light propagation volumes for real-time indirect illumination[C]// Proceedings of the 2010 ACM SIGGRAPH symposium on Interactive 3D Graphics and Games. Washington, D.C: ACM, 2010:99-107.

[20] CRASSIN C, NEYRET F, SAINZ M, et al. Efficient rendering of highly detailed volumetric scenes with GigaVoxels[M]// Crassin2010 GPU Pro.USA: AK Peters, 2010:643-676.