计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200900046

基于边缘计算的图像语义分割应用与研究

王赛男[1]　郑雄风[2]
1 江苏联合职业技术学院南京工程分院　南京 211135
2 南京邮电大学计算机学院、软件学院、网络空间安全学院　南京 210023

摘　要　随着深度学习在医学影像分割、药品检测等医学领域的广泛应用,语义分割技术承载了举足轻重的地位。语义分割融合了目标检测和图像识别两大技术,旨在将图像分割成多组具有特定语义的区域,属于像素级别的密集分类问题。然而为了推动移动视觉识别技术的有效发展,传统深度学习模型在功耗、内存管理、实时性等方面都无法满足移动设备的要求。边缘计算是一种有效将计算、网络、存储、带宽等能力从主机端延伸到移动边缘端的新型架构模式,从而实现在有限计算资源环境下的模型推理运行。因此,文中尝试在基于边缘 TPU 协处理器的开发板上完成 FCN,SegNet,U-Net 等经典图像语义分割模型的转换、部署及推理运行,并在采集的真实药品数据集上验证提出的语义分割模型的正确性及性能。

关键词:深度学习,语义分割,边缘计算,边缘 TPU
中图法分类号　TP181

Application and Research of Image Semantic Segmentation Based on Edge Computing

WANG Sai-nan[1] and ZHENG Xiong-feng[2]
1 Nanjing Engineering Vocational College, Jiangsu Union Technical Institute, Nanjing 211135, China
2 School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China

Abstract　With the extensive application of deep learning in medical imaging segmentation, drug detection and other medical fields, semantic segmentation technology plays a pivotal role. Semantic segmentation combines two techniques of target detection and image recognition. It aims to segment the image into multiple groups of regions with specific semantics, which is a dense classification problem at the pixel level. However, in order to promote the effective development of mobile visual recognition technology, the traditional deep learning model cannot meet the requirements of mobile devices in terms of power consumption, memory management, and real-time performance. Edge computing is a new architecture mode that effectively extends the computing, network, storage, and bandwidth capabilities from the host to the mobile edge to implement model inference operations in a limited computing resource environment. Therefore, this paper attempts to complete the transformation, deployment and inference operation of the classic image semantic segmentation model, such as FCN, SegNet, U-Net, etc, on the development board based on the edge TPU coprocessor, and verifies the correctness and performance of the proposed semantic segmentation model on the collected real drug dataset.

Keywords　Deep learning, Semantic Segmentation, Edge Computing, Edge TPU

1　引言

众所周知,图像语义分割[1](Image Semantic Segmentation,ISS)是计算机视觉和模式识别任务中不可或缺的关键技术,它融合了传统的目标检测和图像分类两大核心理念,旨在将原始数据(例如平面图像)分割成多组具有特定语义类别的区域,即感兴趣区域。语义分割属于像素级别的密集分类问题,在处理图像时,会将图像中每个像素分配到某个对象类别,既要求识别出目标位置,也要求标出每个对象的边界。例如在药品图像分割领域,如图 1(b)所示,语义分割模型不仅要识别出不同药片所在的位置,还要标出每个药片的边界;在医学影像分割领域,如图 1(d)所示,语义分割模型不仅要识别出该人脑是否患有肿瘤,还要明确标出病灶区域范围。因此,与传统分类或检测任务相比,语义分割最大的区别在于其像素级的密集预测能力,因而具有广阔的应用前景。

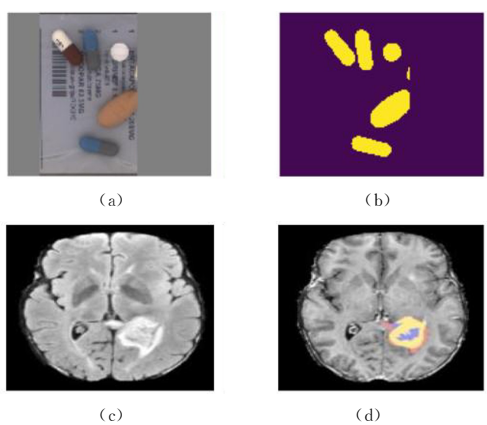

图 1　图像语义分割实例

Fig. 1　Examples of image semantic segmentation

近年来,随着深度学习技术的发展,基于深度学习的语义分割算法层出不穷,如基于编码器-解码器(encoder-decoder)

通信作者:王赛男(wangsn@njevc.edu.cn)

结构的完全卷积网络（Fully Convolutional Networks，FCN）[2]、SegNet[3]、U-Net 网络[4]，以及谷歌推出的基于空洞卷积（Dilated Convolutions）的 DeepLab[5]等。

传统基于深度学习的语义分割模型的训练及推理过程都发生在主机端或云端，主要依赖丰富的硬件设备，即强大的算力资源。然而，随着移动领域的快速发展，传统的语义分割模型从大小、运行效率、功耗等方面都无法满足有限计算资源下的移动设备。在嵌入式设备领域，既要求模型大小越小越好并且模型精度没有显著下降，还要求模型具备低功耗、实时性等特性。

针对以上问题，边缘计算[6]概念应运而生，推动了移动视觉识别技术的有效发展。边缘计算是指在网络边缘执行计算的一种新型计算框架，边缘是指从数据源到云计算中心路径之间的任意资源和网络资源，边缘计算更多地聚集在边缘设备本身，将计算任务迁移到产生源数据的边缘设备上。例如在深度学习领域，依赖强大的云计算中心在海量数据集上进行模型训练，然后将训练完成的模型文件经过量化后迁移到边缘端（例如手机端、树莓派、谷歌 Coral 开发板[7]等嵌入式设备）进行推理，从而实现相应的应用。

2 相关工作

2.1 语义分割

图像语义分割主要分为传统图像语义分割方法和基于深度学习的图像语义分割方法。在深度学习算法普及之前，研究者一般使用纹理基元森林（TextonForest）或者随机森林（Random Forest）方法来构建用于语义分割的分类器，但传统语义分割方法十分依赖特征的选择，特征选择的质量直接影响了算法性能。随着卷积神经网络（Convolutional Neural Network，CNN）在图像识别领域的优异性能逐渐凸显，基于 CNN 改进的图像语义分割算法也相继出炉，如 FCN，SegNet，U-Net 等。

完全卷积网络（Fully Convolutional Networks，FCN）[2]是基于深度学习的语义分割的开山之作，作为首个基于编码器-解码器（encoder-decoder）的深度图像语义分割算法，其首次将 CNN 最后几层全连接都换成卷积，由此获得多维的 feature map，在其后接 softmax 以获得每个像素点的分类信息，从而实现像素化的密集分类，其体系结构如图 2 所示。

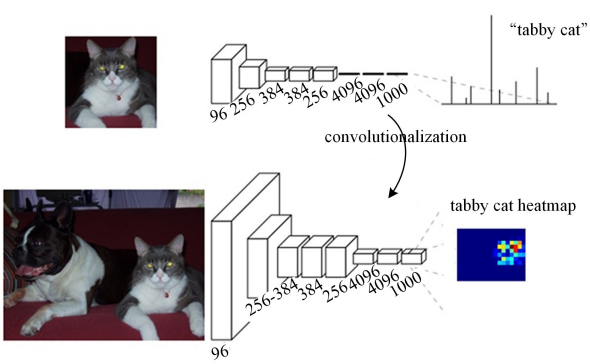

图 2 FCN 体系结构图

Fig. 2 Architecture of FCN

然而，FCN 的缺点也很明显，首先是训练比较繁琐，需要训练 3 次才能够得到 FCN-8s，而且得到的结果并不精细，对图像的细节不够敏感；其次是对各个像素进行分类时没有考虑到像素之间的关系，忽略了基于像素分类的分割方法中使用的空间规整步骤，缺乏空间一致性。

为了改进 FCN，基于编码器-解码器的 SegNet[3]应运而生，其中，编码器使用池化层逐渐缩减输入数据的空间维度，而解码器通过反卷积层等网络层逐步恢复目标的细节和相应的空间维度。从编码器到解码器之间，通常存在直接的信息连接，以帮助解码器更好地恢复目标细节。SegNet 体系结构如图 3 所示，其核心在于解码器对较低分辨率的输入特征图进行上采样。具体地，解码器使用了在相应编码器的最大池化步骤中计算的池化索引来执行非线性上采样。这种方法不必学习上采样的。经上采样后的特征图是稀疏的，因此其随后使用可训练的卷积核进行卷积操作，生成密集的特征图。

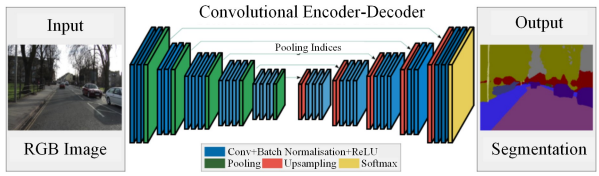

图 3 SegNet 体系结构

Fig. 3 Architecture of SegNet

U-Net[4]是根据 FCN 改进的经典网络，其网络结构如图 4 所示。该网络首先进行 Conv＋Pooling 下采样；然后 Deconv 反卷积进行上采样，将 crop 之前的低层 feature map 进行融合；然后再次上采样。重复这个过程，直到获得输出 388×388×2 的 feature map，最后经过 softmax 获得 output segment map。总体来说 U-Net 与 FCN 思路非常类似。与 FCN 逐点相加不同，U-Net 将特征在 channel 维度拼接在一起，形成更"厚"的特征。

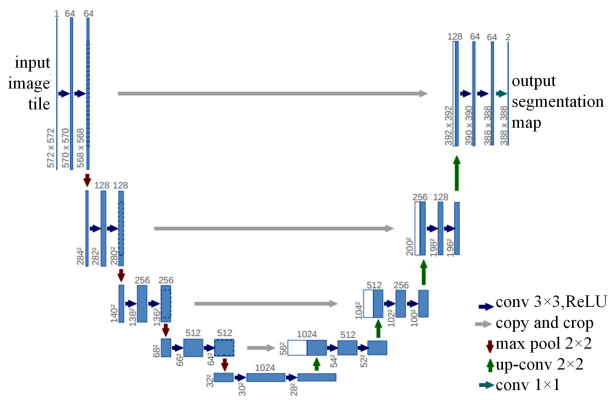

图 4 U-Net 体系结构图

Fig. 4 Architecture of U-Net

2.2 边缘计算

边缘计算[6]是指利用靠近数据源的边缘地带来完成运算程序，其主要计算节点以及应用部署在靠近终端的数据中心，这使得其在服务的响应性能和可靠性方面都高于传统中心化的云计算概念。边缘计算拥有以下特征。

（1）低延时：边缘计算聚焦实时、短周期数据的分析，能够更好地支撑本地业务的实时智能化处理与执行。

（2）高效率：由于边缘计算距离用户更近，在边缘节点处实现了对数据的过滤和分析，因此效率更高。

（3）更节能：云计算和边缘计相结合，其成本只有单独使用云计算的 39%。

（4）环境流量压力小：在进行云端传输时通过边缘节点进行一部分简单数据处理，能够缩短设备响应时间，减

少从设备到云端的数据流量。

3 基于 Edge TPU 的药品图像语义分割

本节主要描述了基于边缘计算的图像语义分割算法在药品图像识别中的应用。首先明确我们的语义分割模型，并在云端或主机端对其进行训练；其次详细描述了训练完成的模型如何部署到边缘设备，包括量化、转换、编译、部署四大步骤；最后在边缘端进行模型推理。整个实现流程如图 5 所示。

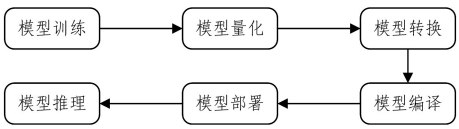

图 5 边缘推理的实现流程

Fig. 5 Flow chart of edge inference

(1)模型训练：随着计算机算力的不断增强，海量数据的增长，深度学习算法的提出使得用更大量数据训练更深的网络成为可能，并且在特定的场景下，一些图像识别算法的准确率已经超越了人类。因此，基于深度学习的语义分割模型取得了不错的进展，本文尝试搭建 FCN，SegNet，U-Net 三大主流深度语义分割模型体系结构并在 Google 云服务器端进行迭代训练以得到较优的分割性能。

(2)模型量化：由于深度学习模型网络体系结构庞大，参数众多，在云端训练所得的模型往往无法直接部署到边缘移动设备上进行推理。移动端都是轻量型设备，内存空间有限，为了解决该问题，首先需要对原始语义分割模型进行量化。深度神经网络的量化技术主要可以分为两类：训练完成后量化和训练过程中进行量化[8]。与改变密度方法不同，量化的方法属于改变网络多样性的方法，容易造成精度损失。一般训练完神经网络后再执行量化的技术主要针对已完整训练的网络，移植性较好。本文即采用了该量化思想，将原始 32 位精度浮点数 FP32 下训练的语义分割模型量化到 8 位精度整数 INT8 下的模型，其实现细节如图 6 所列。

```
bazel build tensorflow/tools/graph_transforms:transform_gaph
bazel-bin/tensorflow/tools/graph_transforms/transform_graph \
  --in_graph=model.pb \
  --out_graph=quantized_model.pb \
  --inputs=input \
  --outputs=output\
```

图 6 Tensorflow 量化工具

Fig. 6 Tensorflow Quantilization Tool

(3)模型转换：为了能够在边缘 TPU 开发板上运行，首先对量化之后的模型进行转换，将其转成特定设备支持的文件格式类型，例如 Google TPU 开发板支持的.tflite 文件、天数智芯 EPU 支持的.tflex 文件等。本文采用 google 开源的 Tensorflow Lite 工具提供的转换接口进行实现，实现细节如图 7 所列。

```
tflite_convert --output_file=model.tflite \
               --graph_def_file=quantized_model.pb \
               --input_arrays=input \
               --output_arrays=output
```

图 7 Tensorflow Lite 转换工具

Fig. 7 Tensorflow Lite conversion tool

(4)模型编译、部署、推理：经转换之后的.tflite 文件需要进行在线编译，然后将编译成功的.tflite 文件发布到边缘端进行模型推理。

4 实验

为了验证上述语义分割模型的训练、量化、转换、部署及推理的正确性，本文在收集的真实药品数据集上进行了语义分割模型的训练，并部署到边缘开发板验证其推理性能。该实验需要做如下准备：1)收集训练中使用的真实药品数据集；2)选择语义分割模型并进行训练；3)完成模型的量化、转换、编译及部署；4)实现边缘推理。

4.1 数据集采集

实验使用真实带标签的药品数据集来训练不同语义分割模型。如图 8 所示，本文采集了 472 张包含不同形状的药品图片作为训练集并对其进行标注，103 张无标签药品图片作为测试集，以验证训练所得语义分割模型在边缘端进行推理的正确性及其分割性能。

(a)药品数据集

(b)对应标签

图 8 药品数据集及对应标签

Fig. 8 Drug data sets and corresponding labels

语义分割模型较传统分类模型最大的区别在于语义分割模型的输出是一张标记了分割结果的图片，而不是简单的一个数字类别。因此对于训练数据的标签，本文采用了 Mask 掩膜思想[7]提取出原始图片的感兴趣区域，从而得到具有语义分割结果的黑白图片，如图 8(b)所示。Mask 实现机制如图 9 所示，假设存在一个 3×3 维的原始图像，通过与自定义的 Mask 矩阵进行逻辑与操作，即原图中的每个像素和 Mask 矩阵中的每个对应像素进行与运算，从而得到包含特定语义区域的效果图。

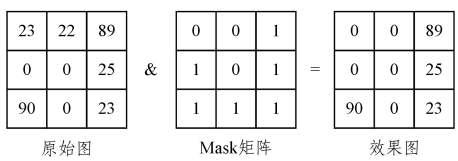

图 9 Mask 实现机制

Fig. 9 Mask Implementation Mechanism

4.2 性能评价指标

近年来，在图像语义分割领域，常用的性能评价指标主要包括像素准确率、平均准确率和平均交并比等[2]。其中，像素准确率(Pixel Accuracy，PA)表示正确分割的像素数量占图像像素总量的百分比，其计算方法如式(1)所示：

$$PA=(\sum_{i=1}^{N}X_{ii})/(\sum_{i=1}^{N}T_i) \tag{1}$$

平均准确率(Mean Accuracy，MA)表示所有类别物体像素准确率的平均值，其计算方法如式(2)所示：

$$MA=(\sum_{i=1}^{N}\frac{X_{ii}}{T_i})/N \tag{2}$$

平均交互比(Mean Intersection over Union，mIoU)表示

分割结果与原始图像真值的重合程度，其计算方法如式(3)所示：

$$mIoU=\left(\sum_{i=1}^{N}\frac{X_{ii}}{T_i+\sum_{j=1}^{N}(X_{ji}-X_{ii})}\right)/N \tag{3}$$

式(1)－式(3)中，N 表示图像像素的类别数量，T_i 表示第 i 类的像素总数，X_{ii} 表示真实类别为 i 且预测类别为 i 的像素总数，X_{ji} 表示真实类别为 i 且预测类别为 j 的像素总数。

从图 8(b)可以看出，本文采集的现实数据包含两个类别(即 $N=2$)，一个类别属于背景，一个类别属于药片所在位置。后期可以针对不同药片进一步细分类。

4.3 软硬件环境

深度学习模型在训练和测试阶段往往对硬件设备要求较高，尤其是在训练时会进行大量的计算并消耗巨大内存和显存，但受限于硬件高昂的价格，本文采用 google 开源的深度学习开发云平台 colab[9] 并使用 tensorflow[10] 深度学习框架进行语义分割模型的训练并保存为 .pb 文件，再对其进行量化、转换、编译，将得到的轻量型模型 .tflite 文件部署到边缘 TPU 开发板(如图 10 所示)进行推理。边缘 TPU 开发板硬件参数如表 1 所列。

表 1 边缘开发板硬件参数

Table 1 Hardware parameters of edge dev board

CPU	NXP i.MX 8M SoC (quad Cortex-A53, Cortex-M4F)
GPU	Integrated GC7000 Lite Graphics
ML accelerator	Google Edge TPU coprocessor
RAM	1 GB LPDDR4
Flash memory	8 GB eMMC
Wireless	Wi-Fi 2×2 MIMO (802.11b/g/n/ac 2.4/5 GHz) and Bluetooth 4.2
Dimensions	48 mm×40 mm×5 mm

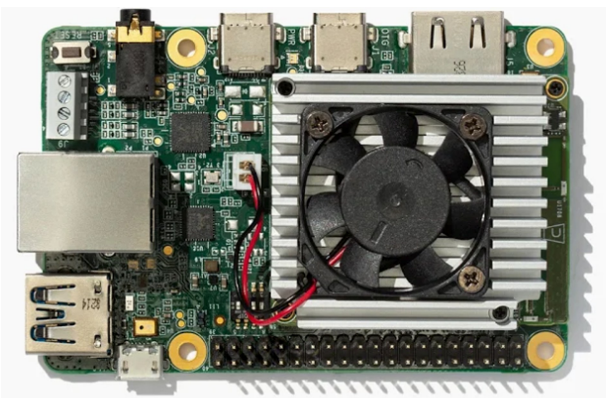

图 10 边缘 TPU 开发板

Fig. 10 Dev Board with Edge TPU

4.4 实验结果分析与对比

为了评估经编译之后的语义分割模型在边缘 TPU 开发板上的分割性能，针对 3 个经典的图像语义分割模型 FCN，SegNet，U-Net 进行了对比实验，其性能表现如表 2 所列。

表 2 图像语义分割模型性能

Table 2 Performance of image semantic segmentation models

模型	PA/%	MA/%	$mIoU$/%	平均耗时/s
FCN	85.3	80.5	72.0	0.500
SegNet	92.1	87.2	83.0	0.062
U-Net	93.5	88.0	85.5	0.065

从表 2 可以看出，基于深度学习的语义分割模型在药品数据集上都取得的较优的分割结果，相对而言，SegNet 与 U-Net 在像素准确率 PA，平均准确率 MA 以及平均交互比 $mIoU$ 评价指标上更优于 FCN。除此之外，我们发现 FCN 分割较为耗时，实时性较差，不适用于在边缘端进行实时推理，而另外两种模型耗时较短，实时性更好，在未来移动端领域中具有更大优势。以上测试结果客观地反映了深度语义分割模型具有较好的分割性能。为了更直观地查看各模型的分割结果，我们给出了 FCN，SegNet 及 U-Net 三大模型对同一张药品图片的分割效果图，如图 11 所示。其中前两列表示原始图像及对应的真实标签，第三列表示各图像语义分割模型对应的输出。从中可以更直观地看出，3 种模型都取得了不错的分割效果，而 SegNet 与 U-Net 对图像的细节更为敏感，分割出的边界要比 FCN 更为清晰。

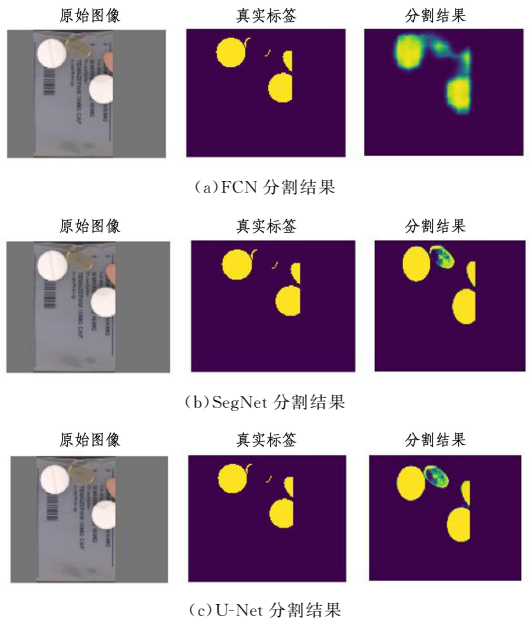

图 11 3 种模型的分割结果

Fig. 11 U-Net segmentation results of three models

为了对比不同语义分割模型在具有更多药品的图片上的分割性能，我们收集了更多类型的药品来测试模型的性能，实验结果如图 12 所示。可以明显看出，当药品数量激增之后(例如第一行收集了 5 种不同药片)，3 个模型的分割性能都有所下降，而 FCN 下降得最为明显，对图像的边界细节不敏感，进而导致分割结果不清晰。

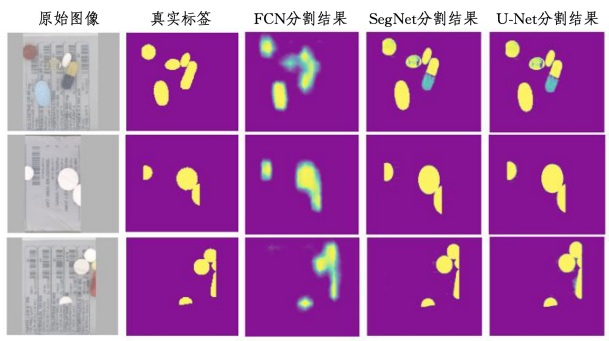

图 12 不同模型的分割结果

Fig. 12 Segmentation result of various models

总而言之，当图像语义分割类别增加之后，基于深度学习的语义分割模型在边缘端的推理性能将有所下降，需要针对该问题做进一步的改进。实验中也发现，云端训练的深度语

义分割模型经过量化之后，准确率会有部分损失，其局限于边缘端的内存限制，实时性也有所下降。若现实应用具有较高的实时性要求，比如在无人驾驶、人脸识别、商品货物目标检测等领域，用户可以尝试使用类似 Tengine[11]，MNN[12]，MACE[13]等开源的边缘端加速推理机制提高模型的推理速度，解决实时性限制。

结束语 本文在真实世界采集的药品数据集上对 FCN，SegNet，U-Net 等常用的语义分割模型进行训练、量化、转换、编译等一系列操作，然后将其成功部署到基于边缘 TPU 协处理器的开发板上进行边缘推理，并在功耗、内存管理、实时性等方面的限制条件下验证模型的分割性能。实验结果表明，基于深度学习的语义分割模型在边缘端取得了不错的分割效果，实时性也较好。但对于实时性要求非常高的实际应用，例如无人驾驶，基于边缘端的模型还有待改进。

随着 5G 的普及，深度学习模型边缘化必将成为主流趋势，如何加速边缘端的运行时间也将成为未来研究的重点。

参 考 文 献

[1] GARCIA-GARCIA A, ORTS-ESCOLANO S, OPREA S O, et al. A Review on Deep Learning Techniques Applied to Semantic Segmentation[J]. arXiv:1704.06857v1, 2018.

[2] JONATHAN L, EVAN S, TREVOR D. Fully Convolutional Networks for Semantic Segmentation[C]// The 2015 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). IEEE, 2015:3431-3440.

[3] VIJAY B, ALEX K, ROBERTO C, et al. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation [C]// The 2017 IEEE Transactions on Pattern Analysis and Machine Intelligence. IEEE, 2017:2481-2495.

[4] OLAF R, PHILIPP F, THOMAS B. U-Net: Convolutional Networks for Biomedical Image Segmentation[C]// The 2015 Medical Image Computing and Computer-Assisted Intervention(MICCAI). Springer, 2015:234-241.

[5] CHEN L C, GEORGE P, IASONAS K. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs[C]// The 2018 IEEE Transactions on Pattern Analysis and Machine Intelligence. IEEE, 2018:834-848.

[6] SHI W S, CAO J, LI Y H, et al. Edge Computing: Vision and Challenges[C]// The 2015 IEEE Internet of Things Journal. IEEE, 2015:637-646.

[7] Google. Learn how to build AI products with Coral devices[EB/OL]. https://coral.withgoogle.com/docs/dev-board/get-started/.

[8] JACOB B, KLIGYS S, CHEN B, et al. Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference[C]// The 2018 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). IEEE, 2018:2704-2713.

[9] Google. Welcome to Colaboratory [EB/OL]. https://coral.withgoo-gle.com/docs/dev-board/get-started/.

[10] ABADI M, AGARWAL A, BARHAM P, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems[J]. arxiv:1603.04467, 2016.

[11] OpenAI. Tengine[EB/OL]. https://github.com/OAID/Tengine.

[12] Alibaba. MNN[EB/OL]. https://github.com/alibaba/MNN.

[13] Xiaomi. MACE[EB/OL]. https://github.com/Xiaomi/mace.

WANG Sai-nan, born in 1980, master, lecturer. Her main research interests include machine learning and pattern recognition.

(上接第 270 页)

[14] ROCK S J, WONZY M J. Generating topological information from a bucket of facets [C]// Proceedings of Solid Freeform Fabrication Symposium Proceedings. 1992:251-258.

[15] ZHANG Z, JOSHI S. An improved slicing algorithm with efficient contour construction using STL files[J]. The International Journal of Advanced Manufacturing Technology, 2015, 80(5/6/7/8):1347-1362.

[16] XIAO H B, ZHOU Y Q, LIU M J, et al. Approach to Optimize STL Model for 3D Laser Machining[C]// Proceedings of 2017 2nd International Conference on Computational Modeling, Simulation and Applied Mathematics (CMSAM 2017). 2017: 192-196.

[17] ZHU J, GUO G, YAN Y N. Research on the fast layering algorithm based on model continuity in rapid prototyping manufacturing [J]. China Mechanical Engineering, 2000(5):77-82, 86.

[18] HU H W, ZHANG L C, ZHANG J F, et al. Fast slicing algorithm of surface color AMF model [J]. Journal of Computer Aided Design and Graphics, 2017, 29(11):2108-2116.

XING Jing-pu, born in 1995, postgraduate. His main research interests include 3D printing and slicing algorithm.

LI Feng-qi, born in 1974, Ph.D, professor, is a member of China Computer Federation. His main research interests include 3D printing, intelligent software systems and blockchain.