文章编号：2095-4980(2021)06-1065-05

基于深度学习的三叉神经区域自动检测及 TensorRT 加速

张倩宇[1]，贾 维[2]，彭 博[*1]

(1.西南石油大学 计算机科学学院，四川 成都 610500；2.成都市温江区人民医院 放射科，四川 成都 610000)

摘 要：利用深度学习技术对颅脑核磁共振图像(MRI)中三叉神经区域进行自动检测可为后续三叉神经分割提供可靠的输入图像，有效解决了人工筛选三叉神经对临床医生专业素养要求高、耗时长等弊端。采用 YOLO 网络自动检测颅脑核磁共振图像中三叉神经区域提高推理速度，并系统性地评估 NVIDIA TensorRT 框架在不同计算平台下的推理性能。实验结果表明，通过 YOLO 目标检测网络能够准确检测出三叉神经所在的区域，同时在 NVIDIA TensorRT 框架下，当输入的颅脑 MRI 分辨率为(204×204)时，CPU 平台、嵌入式 GPU 平台、桌面 GPU 平台及专业 GPU 计算卡平台下，YOLOv2 网络检测优化后的三叉神经目标的每秒帧率分别可达到 0.1 FPS,23.4 FPS,112.5FPS 和 793.7 FPS，这为后续开发便携式的三叉神经分割设备提供了可参考的重要依据。

关键词：颅脑核磁共振图像；目标检测网络；三叉神经；TensorRT；加速

中图分类号：TN911.73 **文献标志码**：A **doi**：10.11805/TKYDA2020136

Automatic detection of trigeminal neural region based on deep learning and TensorRT acceleration

ZHANG Qianyu[1]，JIA Wei[2]，PENG Bo[*1]

(1.School of Computer Science, Southwest Petroleum University，Chengdu Sichuan 610500，China；2.Department of Radiology，Wenjiang District People's Hospital，Chengdu Sichuan 610000，China)

Abstract：Manual screening of trigeminal nerves requires high professional quality and is time consuming for clinicians. Using deep learning to automatically detect trigeminal nerve regions in cranial Magnetic Resonance Imaging (MRI) can provide a reliable input image for subsequent trigeminal nerve segmentation. YOLO(You Only Look Once) network is utilized to automatically detect the trigeminal nerve region of the cranial magnetic resonance image to improve the inference speed, and to systematically evaluate the inference performance of the NVIDIA TensorRT framework under different computing platforms. The experimental results show that the YOLO target detection network can accurately detect the area where the trigeminal nerve is located. Simultaneously, under the NVIDIA TensorRT framework, when the input brain MRI resolution is (204×204), the YOLOv2 network detects the optimized trigeminal nerve through the CPU platform, embedded GPU platform, desktop GPU platform and professional GPU computing card platform, the frame rates per second can reach 0.1 FPS, 23.4 FPS, and 793.7 FPS . This provides important reference for the subsequent development of portable trigeminal neural segmentation equipment.

Keywords：craniocerebral Magnetic Resonance Imaging(MRI)；object detection network；trigeminal nerve；TensorRT；accelerate

对颅脑磁共振(MR)图像中三叉神经区域的自动检测，首先是提取图像的特征，然后根据目标检测算法，再对该区域进行标定定位的自动检测过程。三叉神经是混合神经，由于三叉神经的根部位于颅后窝桥小脑三角区内，

收稿日期：2020-04-01；**修回日期**：2020-05-13
基金项目：四川省苗子工程重点项目(2018RZ0093)；四川省人社厅留学回国人员科技活动资助项目
*通信作者：彭 博 email:bopeng@swpu.edu.cn

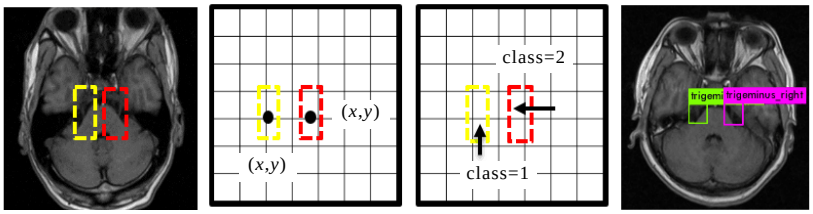
此处血管、神经等组织非常复杂，对临床医生来说，如不能很好地分离三叉神经根和周围血管，会造成颅内神经损伤，并导致一系列复杂的并发症。利用目标检测算法对三叉神经进行分割，可为医生在临床诊断时，分割提取三叉神经附近血管的同时提供可视化的参考依据。传统的三叉神经识别的方法是首先通过人工在一组颅脑 MR 中寻找有三叉神经的颅脑层面并对该位置进行标定，颅脑 MR 图像中各种组织分布密集[1]，对颅脑 MR 图像中神经及其附近血管病变的诊断结果易受到影像科医师个人的理论知识和从业经验的影响。因此，实现颅脑中三叉神经实时性自动检测是三叉神经及附近血管疾病检测所需的关键步骤。近年来随着深度学习在场景目标检测领域取得了极大的成功，如基于区域的卷积特征提取算法(Regions with CNN features，RCNN)[2]及其改进算法 Fast-RCNN[3]，Faster-RCNN[4]，单点多盒探测器(Single Shot Multibox Detector，SSD)[5]，基于回归的目标检测 YOLO[6-8]等，这些基于深度学习的目标检测算法相较于传统的目标检测算法先提取特征再进行模型和目标匹配的方式在识别率上有很大的提升，且在如肿瘤检测、肺结节检测、淋巴结检测等医学检测方向均取得不错的结果。Liu 等人[9]利用肺部 X 线计算机断层摄影(Computer Tomography，CT)图像的三维特性，采用了 3D Faster RCNN 实现了对肺结节的检测，Wang 等人[10]通过 Mask RCNN 从腹部横断位 CT 层切片中获取淋巴结的位置，实现淋巴结的自动检测。Almasni 等人[11]利用 YOLO 实现了乳腺肿块检测方法。

实现三叉神经区域的自动检测需要关注以下问题：1) 如何克服颅脑中其他目标组织的干扰，对三叉神经区域快速准确地定位；2) 如何在保证三叉神经定位精确度的基础上，对 YOLO 网络进行简化，实现推理阶段的加速，满足实时性的要求。为解决上述问题，本文采用 YOLO 网络架构对颅脑三叉神经及血管复合体区域自动定位检测，并在部署阶段采用 TensorRT 框架针对训练好的 YOLO 网络进行优化和系统性的评价。本文的工作致力于通过深度学习方法提供一个快速的 MR 图像中的三叉神经区域自动检测，为实现三叉神经的精准分割提供可靠的输入图像，同时也为开发便携式诊断设备提供参考的计算依据。

1 方法

1.1 基于 YOLO 网络的三叉神经检测定位

以本文的类别检测为例，图 1 为本文所用 YOLO 网络的基本检测原理。图 1(a)是 YOLO 将图像调整为统一大小作为神经网络的输入，通过图 1(b)将图像分成 S × S 的网格，对每个网格进行检测，每个网格输出 B 个标定框，其中包含 2 部分信息：位置信息(*x*,*y*,*w*,*h*)以及输出置信度。图 1(c)是通过训练网络，得到每个预测网格中的 2 个类别概率及 2 个边框坐标左右两侧三叉神经所在区域的标记，图 1(d)是图像的最终预测结果。



(a) input image (b) S × S grid on detection (c) final detection (d) output image

Fig.1 Schematic of YOLO network detection

图 1 YOLO 网络检测原理图

YOLO 网络使用均方和误差作为其损失函数，并由坐标误差、交并比(Intersection-over-Union，IoU)误差和分类误差三部分组成。具体如式(1)：

$$\begin{aligned} Loss=&\lambda_{\text{coord}}\sum_{i=0}^{S\times S}\sum_{j=0}^{2}T_{i,j}^{\text{obj}}\left(x_i-\hat{x}_i\right)^2+\left(y_i-\hat{y}_i\right)^2+\lambda_{\text{coord}}\sum_{i=0}^{S\times S}\sum_{j=0}^{2}T_{i,j}^{\text{obj}}\left[\left(\sqrt{w_i}-\sqrt{\hat{w}_i}\right)^2+\left(\sqrt{h_i}-\sqrt{\hat{h}_i}\right)^2\right]+\sum_{i=0}^{S\times S}\sum_{j=0}^{2}T_{i,j}^{\text{obj}}\left(C_i-\hat{C}_i\right)^2+\\ &\lambda_{\text{noobj}}\sum_{i=0}^{S\times S}\sum_{j=0}^{2}T_{i,j}^{\text{obj}}\left(C_i-\hat{C}_i\right)^2+\sum_{i=0}^{S\times S}T_{i,j}^{\text{obj}}\sum_{C\in\text{classes}}\left[p_i(C)-\hat{p}_i(C)\right]^2 \end{aligned} \tag{1}$$

式中：引入惩罚项 λ_{coord}=5 来修正坐标误差；$T_{i,j}^{\text{obj}}$ 为预测网格中是否含有目标，用 0 和 1 表示；*W* 和 *h* 为标记框的宽和高；*C* 为存在目标对象的概率；*p* 为类别概率；用惩罚项 λ_{noobj}=0.5 来修正 IoU 误差。

在测试推理阶段，每个网格的条件类别概率将会与该网格中的每个目标边框里的置信度相乘，见式(2)。

$$Pr(Class_i\,|\,Object)\times Pr(object)\times IOU_{\text{pred}}^{\text{truth}}=Pr(class_i)\times IOU_{\text{pred}}^{\text{truth}} \tag{2}$$

式中：*Pr* 是指是否含有目标的概率；*IOU* 代表真实目标与预测标定框的交集面积。

通过式(2)得到每个网格中具体类别的概率信息，以及目标边框是否含有物体的概率和它对应坐标的信息。通过极大抑制来筛选出最后需要的目标边框，最后将本文对应的 2 个类别和目标位置显示在图片上。

1.2 TensorRT 加速 YOLO 目标检测网络

NVIDIA TensorRT 是一种高性能神经网络推理引擎。通过优化 TensorRT 可以获得更小、更快、更高效的计算流图，优化后拥有更少层网络结构以及更少的内核运行次数。本文在 Caffe 框架进行模型推理加速，使用 NvCaffeParser 导入模型，导入时输入网络结构 YOLOv1,v2,v3 的 prototxt 文件及 caffemodel 文件，进行推理过程。具体操作流程见图 2。

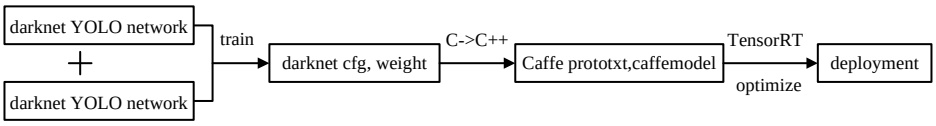


Fig.2 Optimization flow chart
图 2 优化流程图

YOLOv1-tiny 网络采用 Leaky Relu 作为激活函数，根据 IPlugin 接口类成员函数和 Leaky Relu 层的原理，设计 LeakyReluPlugin 类，然后将 LeakyReluPlugin 类插入到 YOLOv1-tiny 网络层中。YOLOv2 网络中添加了 Route 层和 Reorg 层，Leaky-Relu 层采用 YOLOv1 的 Plugin 方式进行添加，Route 层在 Caffe 框架中与全连接层的意义一致，采用 Caffe 框架和 TensorRT 框架下支持的 Concat 层完全替代。Reorg 层，使用 Plugin 的方式进行实现。YOLOv3 网络层中存在 Resample 层、Yolo 层、Route 层、Upsample 层等。其中 Resample 层执行的是 Shaortcut 操作和 Route，用 eltwise 层、concat 层完全替代，Upsample 层使用 Plugin 的方式进行实现。

1.3 基于深度学习的三叉神经区域自动检测模型整体结构

图 3 为本文提出的基于 YOLO 网络的三叉神经目标检测流程图。从左到右依次为输出 MR 图像、通过 YOLO 网络训练出目标检测模型、通过 TensorRT 框架优化 YOLO 网络、在推理阶段得到定位目标区域结果的基础上加速其推理过程，从而完成对三叉神经区域的自动检测。

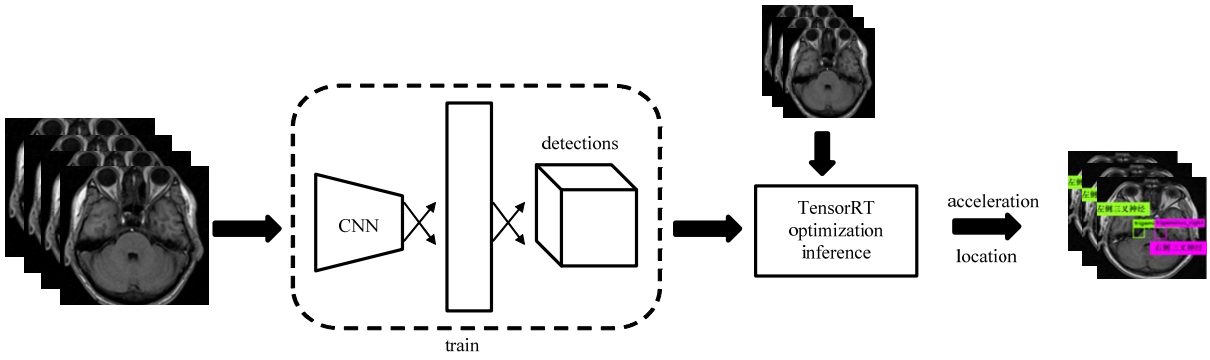


Fig.3 Flow chart of trigeminal nerve detection model
图 3 三叉神经检测模型流程图

1.4 训练数据描述及训练数据过程

本文使用的实验数据集以 PASCAL VOC(Visual Object Class)[12]格式为例，具体的数据来自成都市温江区人民医院提供的颅脑 MR 图像组，为增加实验数据的鲁棒性，通过图像的剪裁、旋转等一系列数据增强处理将筛选出的 156 张图像增加到 445 张，作为最终的训练素材，并标注 2 个类别，分别为左侧三叉神经根和右侧三叉神经根，设置训练集和测试集的比例为 8:2。

在训练过程首先修改网络的配置文件，设置 YOLOv1,v2,v3 参数，在训练时，YOLOv1,v2,v3 的网络层数分别为 16,32,106，训练配置参数的 Batch Size 大小为 32，动量为 0.99，Decay 为 0.000 5，学习率为 0.000 1。

1.5 实验环境说明

为了较系统和全面地测试 TensorRT 加速基于 YOLO 网络的三叉神经自动识别的性能，本文搭建了 CPU 平台、嵌入式 GPU 平台、桌面 GPU 平台及专业 GPU 计算卡平台，用于性能测试。其中 GPU 型号分别为 GeForce GTX TITAN X,NVIDIA Quadro P6000,Telsa K10,NVIDIA Jetson TX2。以上计算机平台均安装有 NVIDIA CUDA 8.0、Python2.7，并配置相应的库如 Darknet,Caffe,Numpy,OpenCV3 等用于数据处理。

2 结果

2.1 三叉神经区域检测精确度结果

在测试集中分别随机选取(384×512),(467×467),(204×204)大小颅脑 MR 图像，设置对照实验，将训练得到的 YOLOv1,v2,v3 的权重模型分别在 CPU 平台、嵌入式 GPU 平台、桌面 GPU 平台及专业 GPU 计算卡平台下进行推理。利用训练得到的 YOLOv1,YOLOv2,YOLOv3 权重模型对颅脑 MR 测试集进行三叉神经目标区域的检测，采用平均精确度(mean Accuracy Precision，mAP)[13]作为验证 YOLOv1,YOLOv2,YOLOv3 目标检测网络的评价指标，得到的 mAP 值分别为 58%,73%,93%。通过对比以上 3 个网络的平均准确率可以看出 YOLOv3 网络的准确度是最高的，由此可以推断网络层数的加深以及经过网络层的优化与改进可以提高目标检测的准确率。

2.2 三叉神经区域检测时间结果分析

不同分辨率时精确度为 Float32 类型的图像分别在 YOLOv3,YOLOv2,YOLOv1-tiny 网络下推理的时间对比见表 1~表 3。

表 1 不同分辨率时在 YOLOv3 网络下推理的时间对比(单位：ms)

Table1 Time comparison of reasoning under YOLOv3 network at different resolutions(unit: ms)

resolution/pixel	precision type	CPU	GPU			TensorRT			Jetson TX2 GPU	Jetson TX2 TensorRT
			GeForceGTX TITAN X	Quadro P6000	Tesla K10	GeForceGTX TITANX	Quadro P6000	Tesla K10		
384×512	Float32	20 685	21.00	19.50	150.12	6.677	6.129	41.83	435.25	85.21
467×467	Float32	20 787	20.00	20.00	149.67	6.534	6.012	42.56	434.08	87.50
204×204	Float32	20 609	18.83	19.21	149.90	5.898	[5.353]	39.65	432.87	86.44

表 2 不同分辨率时在 YOLOv2 网络下推理的时间对比(单位：ms)

Table2 Time comparison of reasoning under YOLOv2 network at different resolutions(unit:ms)

resolution/pixel	precision type	CPU	GPU			TensorRT			Jetson TX2 GPU	Jetson TX2 TensorRT
			GeForceGTX TITAN X	Quadro P6000	Tesla K10	GeForceGTX TITANX	Quadro P6000	Tesla K10		
384×512	Float32	9 982	8.98	8.92	64.87	1.854	1.641	23.82	217.25	45.25
467×467	Float32	9 930	8.90	8.84	63.59	1.649	1.280	23.01	214.50	44.89
204×204	Float32	9 960	8.91	8.89	62.80	1.543	[1.260]	22.17	213.80	42.72

表 3 不同分辨率时在 YOLOv1-tiny 网络下推理的时间对比(单位：ms)

Table3 Time comparison of reasoning under YOLOv1-tiny network at different resolutions(unit:ms)

resolution/pixel	precision type	CPU	GPU			TensorRT			Jetson TX2 GPU	Jetson TX2 TensorRT
			GeForceGTX TITAN X	Quadro P6000	Tesla K10	GeForceGTX TITANX	Quadro P6000	Tesla K10		
384×512	Float32	12 830	130	130	229.56	4.183	3.92	37.69	498.05	25.39
467×467	Float32	12 780	127	125	223.04	4.256	4.12	37.12	497.41	24.32
204×204	Float32	12 740	128	125	200.83	3.965	[3.87]	36.85	490.12	23.23

由表 1~表 3 可以看出，在精确度一定的情况下，Quadro P6000 显卡的性能最佳，推理时间最短，并且通过 TensorRT 优化后的网络比在 GPU 上部署后直接进行推理的速度有大幅提高。对比发现，Tesla K10 显卡相比其他显卡的性能最差，经过 TensorRT 加速过之后在 Jetson TX2 嵌入式设备上推理耗时都大幅减少。

对比 YOLOv1,v2 和 v3 网络进行推理的时间，YOLOv2 在 CPU,GPU,Jetson TX 2 上的耗时明显少于 YOLOv1-tiny 和 YOLOv3。YOLOv2 相对于 YOLOv1 的优势体现在准确度和速度方面，主要在于 YOLOv2 使用分类网络代替特征提取，并对特征进行了压缩。YOLOv2 相对于 YOLOv3 网络在推理时间上的优势是由于 YOLOv3 的网络深度是 YOLOv2 网络深度的 3 倍多，导致 YOLOv2 在推理时间上明显少于 YOLOv3。

图 4 为输入分辨率(467×467)的 MR 图像后，YOLOv1,v2 和 v3 网络在不同平台下的推理时间对比结果。图 4 中横坐标代表不同平台下推理时间的变化。从图 4 中可以发现通过 TensorRT 加速过的网络推理时间明显比未经过加速的网络推理时间少。在嵌入式平台 Jetson TX2 上也呈现这样的规律。

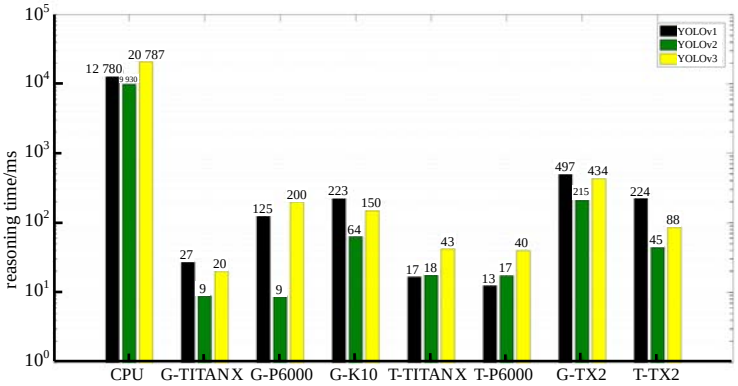


Fig.4 Comparison of reasoning time under different computing platforms

图 4 不同平台下推理时间对比

综合来看，YOLOv2网络在任何一种平台上表现都较优，更适用于颅脑三叉神经目标区域的检测部署。

3　结论

本文以实时检测三叉神经为研究目的，提出了基于YOLO网络的三叉神经区域自动检测方案，该自动检测方案可以输出包含MRI图像中颅脑中三叉神经根及周边区域的完整图像，能够为后续研究三叉神经分割奠定基础。从实验结果来看，YOLO网络实现了从颅脑MR图像中快速、准确地自动定位出三叉神经目标区域，能够为医生的临床观察判断提供良好的参考依据。同时基于NVIDIA TensorRT引擎的性能加速，经过优化后的三叉神经目标检测YOLO网络在CPU平台、嵌入式GPU平台、桌面GPU平台及专业GPU计算卡平台下的推理计算时间为0.1FPS,23.4FPS,112.5FPS和793.7FPS。结果表明通过NVIDIA TensorRT可以在保证检测精确度的情况下实现推理阶段的加速，性能良好的GPU在部署阶段也表现出一定程度的优越性，为后续开发便携式的三叉神经分割设备提供了可供参考的重要依据。

参考文献：

[1] 朱瑞,贾红丽. 医学影像技术在医学影像诊断中的临床应用分析[J]. 影像研究与医学应用, 2017,26(5):146-147. (ZHU Rui,JIA Hongli. Analysis of clinical application of medical imaging technology in medical imaging diagnosis[J]. Imaging Research and Medical Applications, 2017,26(5):146-147.)

[2] GIRSHICK R,DONAHUE J,DARRELL T,et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]// 2014 IEEE Conference on Computer Vision & Pattern Recognition. Columbus,OH:IEEE, 2014:580-587.

[3] GIRSHICK R. Fast R-CNN[C]// 2015 IEEE International Conference on Computer Vision. Santiago:IEEE, 2015:1440-1448.

[4] REN S,HE K,GIRSHICK R,et al. Faster R-CNN:towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017,39(6):1137-1149.

[5] LIU W,ANGUELOV D,ERHAN D,et al. SSD:Single Shot MultiBox Detector[J]. INSPEC, 2016,9905 LNCS:21-37.

[6] REDMON J,DIVVALA S,GIRSHICK R,et al. You only look once:unified, real-time object detection[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). IEEE, 2016:779-788.

[7] REDMON J,FARHADI A. YOLO9000:better,faster,stronger[C]// IEEE Conference on Computer Vision & Pattern Recognition. Honolulu, HI, USA:IEEE, 2017: 6517-6525.

[8] REDMON J,FARHADI A. YOLOv3: an incremental improvement[EB/OL]. (2018)[2019-11-11]. https://pjreddie.com/media/files/papers/YOLOv3.pdf.

[9] 刘迪,王艳娇,徐慧. 基于深度学习的医学图像肺结节检测[J]. 微电子学与计算机, 2019,36(5):5-9. (LIU Di,WANG Yanjiao,XU Hui. Detection of lung nodules in medical images based on deep learning [J]. Microelectronics and Computer, 2019, 36(5):5-9.)

[10] 王嘉奇. 基于深度学习的淋巴结自动检测算法研究[D]. 杭州:浙江大学, 2019. (WANG Jiaqi. Research on automatic detection of lymph nodes based on deep learning[D]. Hangzhou,China:Zhejiang University, 2019.)

[11] ALMASNI M A,ALANTARI M A,PARK J M,et al. Detection and classification of the breast abnormalities in digital mammograms via regional Convolutional Neural Network[C]// 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society(EMBC). Seogwipo,South Korea:IEEE, 2017:1230-1233.

[12] EVERINGHAM M,ESLAMI S M A,GOOL L V,et al. The pascal visual object classes challenge:a retrospective[J]. International Journal of Computer Vision, 2015,111(1):98-136.

[13] EVERINGHAM M,GOOL L V,WILLIAMS C K I,et al. The pascal Visual Object Classes(VOC) challenge[J]. International Journal of Computer Vision, 2010,88(2):303-338.

作者简介：

张倩宇(1995-)，女，新疆维吾尔自治区伊犁哈萨克自治州人，在读博士研究生，主要从事人工智能图像处理的研究.email:1211932526@qq.com.

彭　博(1980-)，男，四川省南部县人，博士，副教授，主要从事图像与信号处理的研究.email:bopeng@swpu.edu.cn.

贾　维(1980-)，女，四川省南充市人，副主任医师，主要研究方向为神经影像.