文章编号 1004-924X(2019)12-2722-08

基于 DeepLabV3＋与超像素优化的语义分割

任凤雷[1,2],何 昕[1],魏仲慧[1],吕 游[1*],李沐雨[1,2]
(1.中国科学院 长春光学精密机械与物理研究所,吉林 长春 130033;
2.中国科学院大学,北京 100049)

摘要:针对基于深度学习的 DeepLabV3＋语义分割算法在编码特征提取阶段大量细节信息被丢失,导致其在物体边缘部分分割效果不佳的问题,本文提出了基于 DeepLabV3＋与超像素优化的语义分割算法。首先,使用 DeepLabV3＋模型提取图像语义特征并得到粗糙的语义分割结果;然后,使用 SLIC 超像素分割算法将输入图像分割成超像素图像;最后,融合高层抽象的语义特征和超像素的细节信息,得到边缘优化的语义分割结果。在 PASCAL VOC 2012 数据集上的实验表明,相比较 DeepLabV3＋语义分割算法,本文算法在物体边缘等细节部分有着更好的语义分割性能,其 mIoU 值达到 83.8%,性能得到显著提高并达到了目前领先的水平。

关 键 词:深度学习;DeepLabV3＋;超像素;语义分割
中图分类号:TP394.1 **文献标识码**:A **doi**:10.3788/OPE.20192712.2722

Semantic segmentation based on DeepLabV3＋ and superpixel optimization

REN Feng-lei[1,2], HE Xin[1], WEI Zhong-hui[1], LÜ You[1*], LI Mu-yu[1,2]

(1. *Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun* 130033, *China*;
2. *University of Chinese Academy of Sciences, Beijing* 100049, *China*)
* *Corresponding author, E-mail: lvyou*8863@163.*com*

Abstract: To tackle the problem where by DeepLabV3＋ loses considerable detail information during feature extraction, which leads to poor segmentation results in the edges of the objects, this study proposed a semantics segmentation algorithm based on DeepLabV3＋ and optimized by superpixels. First, a DeepLabV3＋ model was chosen to extract semantic features and obtain coarse semantic segmentation results. Then, the simple linear iterative clustering algorithm was used to segment the input image into superpixels. Finally, high-level abstract semantic features and detailed information of the superpixels were fused to obtain edge optimized semantic segmentation results. Experiments conducted on the PASCAL VOC 2012 dataset show that compared to DeepLabV3＋, the proposed algorithm had superior performance in terms of detail parts such as edges of objects, and the value of mIoU reached 83.8%. The proposed algorithm thus outperformed other state-of-the-art algorithms in terms of semantic segmentation.

Key words: deep learning; DeepLabV3＋; superpixel; semantic segmentation

收稿日期:2019-06-24;**修订日期**:2019-08-17.
基金项目:吉林省科技发展计划资助项目(No. 20180201013GX)

1　引　言

图像的语义分割旨在为图像中的每个像素分配类别标签，即实现像素级的分类。近年来，语义分割已经逐渐发展成为计算机视觉领域的基础且重要的研究内容，并被广泛应用于自动驾驶、场景理解及机器人导航等领域[1-2]。

传统的算法依赖于手动提取的特征并结合分类器实现语义分割，如随机森林（Random Forests）[3]，Boosting[4]，支持向量机（Support Vector Machines，SVM）[5]等。这些算法拥有较少的参数和计算复杂度，但是通常受限于所选取的特征无法描述像素的高级语义信息，所以其语义分割结果无法满足实际应用需求。

近年来，深度学习技术在计算机视觉领域的快速发展，深度卷积神经网络（Convolutional Neural Network，CNN）在图像分类[6-8]、图像识别[9-10]等领域都得到了广泛应用，在此基础上，像素级的语义分割技术也获得了巨大进步。基于 CNN 的语义分割模型首先由 Long 等提出并被定义为全卷积网络（Fully Convolution Network，FCN）[11]，该模型取消了传统卷积神经网络中的全连接层，并使用卷积层代替，其性能相比较传统语义分割算法有了突破性地进步，但由于连续池化以及下采样，特征图分辨率逐渐减小，很多细节信息被丢失，导致语义分割结果过于粗糙。U-Net[12]，SegNet[13]等方法采用更加优雅的 Encoder-Decoder 网络结构，用低层的信息帮助高层的语义特征恢复图像细节信息，但此类方法无法解决物体存在多尺度的问题而造成语义分割失败。Chen 等提出的 DeepLab[14]模型采用扩张卷积代替池化层来增大感受野，并使用 ASPP（Atrous Spatial Pyramid Pooling）模块解决物体存在不同尺度的问题。之后，Chen 等又提出了 DeepLabV3[15]，通过优化模型并改善训练策略得到了更好的语义分割效果。在前述基础上，Chen 等提出了 DeepLabV3＋[16]，通过加入解码模块解决在 DeepLabV3 模型中由特征图直接上采样来恢复原始图像分辨率大小导致大量细节信息丢失的问题，并取得了先进的语义分割性能。

虽然现有的基于深度学习的语义分割算法不断优化了分割效果并获得了巨大的成功，但仍然存在如下问题：基于深度学习的语义分割算法在特征提取阶段通过连续堆叠池化层或降采样来增大感受野，因此很多的物体边缘等细节信息在卷积过程中被丢失，即使最近的语义分割算法提出采用扩张卷积代替池化层来产生密集的预测的策略，以 DeepLabV3＋为例，其编码输出的特征图相较于输入图像分辨率减小了 16 倍，仍然有很多的细节信息被丢失，导致现有算法在物体边缘部分分割效果不佳。

近年来，使用超像素的思想已经被广泛应用于不同领域。如 Ren 等提出利用道路图像超像素信息构建道路模型从而实现了对道路的检测[17]。Zhao 等提出利用超像素和条件随机场来优化 FCN 深度学习模型的语义分割效果[18]。

本文提出了基于 DeepLabV3＋与超像素优化的语义分割算法。首先使用 DeepLabV3＋模型提取图像语义特征并得到粗糙的语义分割结果；为了恢复在池化或下采样过程中丢失的物体边缘等图像细节信息，使用 SLIC 超像素分割算法将输入图像分割成超像素图像，并将高层的抽象语义特征和超像素的物体边缘细节信息进行融合，得到最终的语义分割结果。

2　本文算法

本文提出的基于 DeepLabV3＋与超像素优化的语义分割算法如图 1 所示（彩图见期刊电子版），其中黄色框表示 DeepLabV3＋的编码部分；红色框表示 DeepLabV3＋的解码部分，本文利用此编码-解码结构来提取图像语义特征并得到粗糙的语义分割结果；图中灰色框表示超像素优化部分，其作为后处理部分通过融合高层抽象语义特征和物体边缘细节信息优化语义分割结果。总的来说，高层卷积特征获得精确的语义信息，超像素边缘特征获得精确的空间信息。

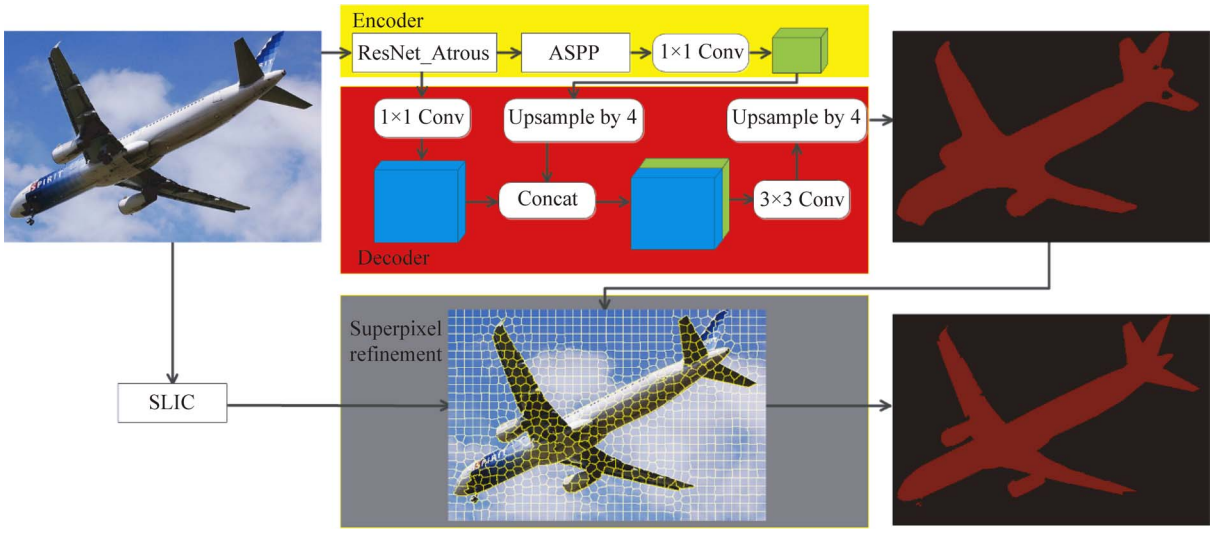

图 1 本文算法原理图

Fig. 1 Diagram of proposed algorithm

2.1 DeepLabv3+

DeepLabV3+是目前最为先进的语义分割算法之一，其在 DeepLabV3 模型的基础上，利用其作为编码模块输出特征图，并添加解码模块实现语义分割。DeepLabV3 使用深度残差网络(ResNet_101)[19]提取语义信息，采用扩张卷积(Atrous Convolution)来控制输出特征图的分辨率并扩大卷积核的感受野。以二维特征图为例，假设卷积核为 w，当扩张卷积作用于输入特征图 x，对于输出特征图 y 中的每个位置 i，有：

$$y[i]=\sum_k x[i+r\times k]w[k], \tag{1}$$

其中：r 表示扩张率。扩张卷积的示意图如图 2 所示。图 2(a)～图 2(c)分别对应卷积核 3×3 扩张率为 1，2，4 的扩张卷积。

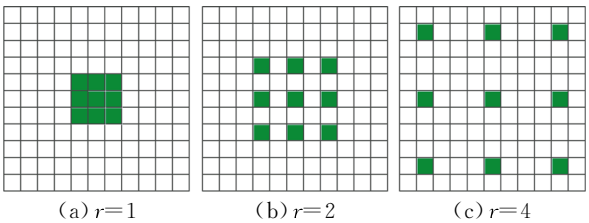

图 2 扩张卷积示意图

Fig. 2 Outline of atrous convolution

此外，DeepLabV3 使用扩张空间金字塔池化(Atrous Spatial Pyramid Pooling，ASPP)模型，应用带有不同扩张率的扩张卷积和图像级特征来得到不同尺度的丰富的语义信息。ASPP 模型如图 3 所示。

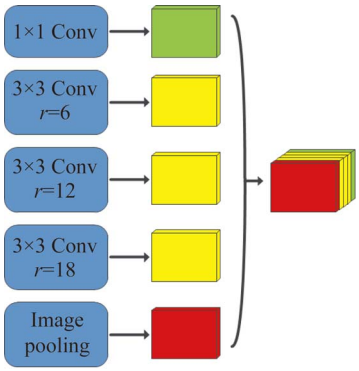

图 3 ASPP 模型示意图

Fig. 3 Outline of ASPP model

定义输入图像空间分辨率与最终输出特征图分辨率的比值为输出步长(Output Stride)。在 DeepLabV3 中输出步长为 16，故本文对最后的特征图直接采用因子为 16 的双线性插值进行上采样处理，处理方式过于粗糙。针对此问题，本文在 DeepLabV3+中加入解码模块，首先将编码特征进行因子为 4 的双线性插值上采样；然后连接低层网络输出的具有相同空间分辨率的特征层；连接后，采用 3×3 的卷积核和因子为 4 的双线性插值上采样来将输出特征图恢复为输入图像的空间分辨率大小，最终完成语义分割。

2.2 Superpixel

尽管 DeepLabV3＋通过添加解码模块恢复了部分细节信息，但仍然采用两次双线性插值上采样来增加特征分辨率，在物体边缘处的语义分割效果不够理想。考虑到超像素具有可以保护物体边缘的特性，本文通过融合高层语义特征和超像素物体边缘信息来优化语义分割结果。

通常，超像素可以被认为是一组位置、颜色、纹理等相似的像素集合。本文使用 Achanta 等提出的 SLIC(Simple Linear Iterative Cluster)超像素分割算法[20]将输入图像分割成超像素图像。首先，它将彩色图像转化为 CIE-Lab 颜色空间，对应每个像素的(L，a，b)颜色值和(x，y)坐标组成一个 5 维向量，然后对此构造距离度量标准。最后通过迭代的方式对图像像素进行局部聚类。其中，迭代计算聚类中心是本算法的关键，而迭代的核心就是计算距离。其计算公式如下：

$$d_c=\sqrt{(l_j-l_i)^2+(a_j-a_i)^2+(b_j-b_i)^2},\quad(2)$$

$$d_s=\sqrt{(x_j-x_i)^2+(y_j-y_i)^2},\quad(3)$$

$$D'=\sqrt{\left(\frac{d_c}{N_c}\right)^2+\left(\frac{d_s}{N_S}\right)^2},\quad(4)$$

其中：d_c 代表颜色距离，d_s 代表空间距离，N_S 表示类内最大空间距离，定义为：

$$N_S=S=\sqrt{\frac{N}{K}},\quad(5)$$

其中：K 表示超像素总数，N_c 表示最大的颜色距离，随图片及聚类不同而不同，取一个固定常数 m 代替。最终的度量公式为：

$$D'=\sqrt{\left(\frac{d_c}{m}\right)^2+\left(\frac{d_s}{S}\right)^2}.\quad(6)$$

图 4 由左到右分别表示 K 值取 100，500，1 000时的超像素分割结果。为提高超像素在物体边缘处分割的精确性，本文设置 K 值为 1 000。

图 4　超像素图像

Fig. 4　Superpixels of input image

2.3 超像素优化

本节通过融合高层语义特征和超像素物体边缘信息来优化语义分割结果。首先得到由 DeepLabV3＋输出的粗糙语义分割结果，然后统计每个超像素内各语义类别所占像素总数，最后选择像素总数最多的语义类别并将其赋给该超像素。具体算法如下。

Algorithm: Superpixel refinement.

1. Input image I and coarse semantic segmentation result L by DeepLabV3＋.
2. Let $C=\{C_1, C_2, \cdots, C_N\}$ refers to all the N kinds of semantic classifications which need to be assigned to each pixel.
3. Segment the input image I using SLIC algorith m intonsuperpixels, thus we have $S=\{S_1, S_2, \cdots, S_n\}$, and there are m_i pixels in the superpixel S_i.
4. In L:
   for $i=1:n$
      for $j=1:m_i$
         Find p_j belongs to which kind of semantic classification in$C$, where p_j means the j_{th} pixel in m_i, e. g., if p_j refers to C_k, let $M_k=M_k+1$, i. e., make M_k plus 1, where M_k means the total number of pixels referring to C_k.
      end
      Find the semantic classification C_{max} which have the most corresponding pixels in the superpixel S_i and assign C_{max} to this superpixel.
   end
5. We get the final $\widetilde{L}$ which refers to the semantic segmentation result after superpixel refinement and output the $\widetilde{L}$.

图 5 为使用上述超像素优化算法融合高层语义特征和超像素物体边缘信息的语义分割结果，如图所示，左侧部分为融合前的语义分割结果，即 DeepLabV3＋输出的粗糙的语义分割图；

右侧部分为本文所提算法的语义分割结果，即经过超像素优化的语义分割图；中间部分为局部细节放大对比图，由对比结果可看出，DeepLabV3＋的高层语义信息和超像素的边缘信息被很好地融合了，在物体边缘处的语义分割得到了优化。

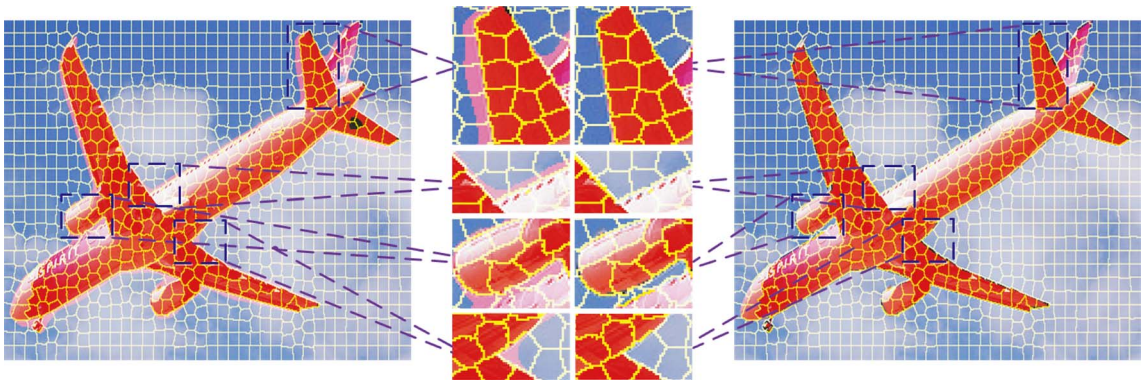

图 5　超像素优化结果

Fig. 5　Result of superpixel refinement

3　实验结果分析

为验证所提算法的有效性，本文对上述算法进行了实验验证。本节首先介绍实验的平台及相关参数设置，然后介绍实验所使用的数据集。接下来对所提出的算法进行了定性和定量分析，并与其它先进的语义分割算法进行了对比。

3.1　实验平台及相关参数设置

本文使用 ImageNet 预训练的 ResNet-101[19]网络进行微调并使用扩张卷积来提取密集的图像语义特征；使用随机梯度下降（Stochastic Gradient Descent，SGD）法训练模型；参照文献[16]，本文使用“poly”学习策略[21]，设置初始化学习率为 0.007，然后在训练过程中将学习率乘以 $\left(1-\frac{iter}{max\_iter}\right)power$，其中 $power$ 值为 0.9；在数据增强方面，以 0.5～2 的比例随机缩放输入图像并在训练过程中将图像进行随机翻转。实验基于 Ubuntu 18.04 操作系统，CPU 为 Inter Core i7-6700，GPU 为 NVIDIA GTX 1080Ti，使用 TensorFlow 深度学习框架来训练并测试本文的语义分割模型。

3.2　数据集

PASCAL VOC 2012[22]数据集是常用的训练和评价语义分割算法性能的数据集，该数据集拥有 1 464 张训练图片，1 449 张验证图片和1 456 张测试图片，包括 20 个前景类别和 1 个背景类别共 21 个语义分类，该数据集中的大部分图像的分辨率接近 500×500。在此基础上，Hariharan[23]对 PASCAL VOC 2012 数据集作了增强，提供了额外的像素级标注，使数据集拥有 10 582 张训练图像。本文使用这些图像来训练并测试网络模型。

3.3　定性分析

根据本文提出的语义分割算法，在 PASCAL VOC 2O12 数据集上进行了验证。本文算法与 DeepLabV3＋的语义分割结果对比图如图 6 所示，其中第 1 列为输入图像，第 2 列为 DeepLabV3＋的语义分割结果，第 3 列为本文算法的语义分割结果，第 4 列为 ground-truth，即真实的语义标签。

由图 6 可以看出，相比较 DeepLabV3＋语义分割算法，本文算法在融合了高层语义特征和超像素信息后，在物体边缘等细节部分有着更高的语义分割准确性，语义分割性能得到了提升。

图 6　语义分割结果图

Fig. 6　Results of semantic segmentation

3.4 定量分析

本文使用 mIoU(mean Intersection over Union)作为标准评价语义分割算法性能。在图像分割领域 mIoU 值是一个衡量图像分割精度的重要指标,其可解释为平均交并比,即在每个类别上计算 IoU(Intersection over Union)值。IoU 和 mIoU 的定义如下:

$$IoU=\frac{P_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}}, \tag{7}$$

$$mIoU=\frac{\sum_{i=0}^{k}IoU_i}{k+1}, \tag{8}$$

其中:$k+1$ 表示包括背景在内的语义类别总数,i 表示真实值,j 表示预测值,p_{ij} 表示将类别 i 预测为类别 j 的像素数量。

基于上述评价标准,用本文提出的算法和目前比较先进的语义分割算法在 PASCAL VOC 2012 测试集上进行了对比,结果如表 1 所示。由表 1 可知,相比较 DeepLabV3+,由于本文算法融合了超像素信息,所以在物体边缘等细节部分有着更好的分割性能。本文算法的语义分割 mIoU 值达到了 83.8%,性能得到显著提高并达到了目前领先的水平。

表 1 语义分割检测算法性能对比

Tab. 1 Comparison of semantic segmentation algorithms

Method	mIoU/%
FCN_8s[11]	62.2
SegNet[13]	59.1
DeepLab[14]	71.6
DeepLabV3[15]	77.2
DeepLabV3+(ResNet101)[16]	80.2
Our approach	83.4

4 结　论

本文提出了一种基于 DeepLabV3+与超像素优化的语义分割算法。首先使用 DeepLabV3+模型提取图像语义特征并得到粗糙的语义分割结果,然后使用 SLIC 超像素分割算法将输入图像分割成超像素,在此基础上,将高层的抽象语义特征和超像素的物体边缘细节信息进行融合并输出语义分割结果,以恢复在池化或下采样过程中丢失的物体边缘等图像细节信息,解决现有基于深度学习的语义分割算法在物体边缘部分表现不佳的问题。

参考文献:

[1] LADICKY L, SHI J, POLLEFEYS M. Pulling things out of perspective [C]. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014:89-96.

[2] XIAO J, QUAN L. Multiple view semantic segmentation for street view images [C]. 2009 *IEEE 12th international conference on computer vision*, 2009: 686-693.

[3] SHOTTON J, JOHNSON M, CIPOLLA R. Semantic texton forests for image categorization and segmentation [C]. 2008 *IEEE Conference on Computer Vision and Pattern Recognition*, 2008:1-8.

[4] TU Z, BAI X. Auto-context and its application to high-level vision tasks and 3d brain image segmentation [J]. *IEEE transactions on pattern analysis and machine intelligence*, 2009, 32(10):1744-1757.

[5] FULKERSON B, VEDALDI A, SOATTO S. Class segmentation and object localization with superpixel neighborhoods [C]. 2009 *IEEE 12th international conference on computer vision*, 2009:670-677.

[6] KRIZHENVSHKY A, SUTSKEVER I, HINTON G. Imagenet classification with deep convolutional networks [C]. *Proceedings of the Conference Neural Information Processing Systems (NIPS)*, 1097-1105.

[7] WU Z, SHEN C, VAN DEN HENGEL A. Wider or deeper: Revisiting the resnet model for visual recognition [J]. *Pattern Recognition*, 2019, 90:119-133.

[8] 李宇, 刘雪莹, 张洪群, 等. 基于卷积神经网络的光学遥感图像检索 [J]. 光学 精密工程, 2018, 26(1):200-207.
LI Y, LIU X Y, ZHANG H Q, *et al.*. Optical remote sensing image retrieval based of convolutional neural networks [J]. *Opt. Precision Eng.*, 2018, 26(1):200-207. (in Chinese)

[9] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition

[J]. *arXiv preprint* arXiv:1409.1556,2014.

[10] 方明，孙腾腾，邵桢. 基于改进 YOLOv2 的快速安全帽佩戴情况检测 [J]. 光学 精密工程，2019，27(5)，1196-1205.
FANG M, SUN T T, SHAO Z. Rapid helmet wear detection based on improved YOLOv2 [J]. *Opt. Precision Eng.*, 2019,27(5),1196-1205. (in Chinese)

[11] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation [C]. *Proceedings of the IEEE conference on computer vision and pattern recognition*,2015:3431-3440.

[12] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation [C]. *International Conference on Medical image computing and computer-assisted intervention*,2015:234-241.

[13] BADRINARAYANAN V, KENDALL A, CIPOLLA R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation [J]. *IEEE transactions on pattern analysis and machine intelligence*, 2017, 39(12):2481-2495.

[14] 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):834-848.

[15] CHEN L-C, PAPANDREOU G, SCHROFF F, *et al.*. Rethinking atrous convolution for semantic image segmentation [J]. *arXiv preprint* arXiv: 1706.05587,2017.

[16] CHEN L-C, ZHU Y, PAPANDREOU G, *et al.*. Encoder-decoder with atrous separable convolution for semantic image segmentation [C]. *Proceedings of the European Conference on Computer Vision (ECCV)*,2018:801-818.

[17] REN F, HE X, WER Z, *et al.*. Fusing appearance and prior cues for road detection [J]. *Applied Sciences*,2019, 9(5):996.

[18] WEI Z, YI F, WEI X, *et al.*. An improved image semantic segmentation method based on superpixels and conditional random fields [J]. *Applied Sciences*,2018, 8(5):837.

[19] 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*,2016:770-778.

[20] ACHANTA R, SHAJI A, SMITH K, *et al.*. SLIC superpixels compared to state-of-the-art superpixel methods [J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*,2012, 34(11):2274-2282.

[21] LIU W, RABINOVICH A, BERG AC. Parsenet: Looking wider to see better [J]. *arXiv preprint* arXiv:1506.04579,2015.

[22] EVERINGHAM M, VAN GOOL L, WILLIAMS CK, *et al.*. The pascal visual object classes (voc) challenge [J]. *International journal of computer vision*,2010, 88(2):303-338.

[23] HARIHARAN B, ARBELAEZ P, BOURDEV L, *et al.*. Semantic contours from inverse detectors [C]. 2011 *International Conference on Computer Vision*,2011:991-998.

作者简介：

任凤雷(1991－)，男，河北沧州人，博士研究生，2015 年于吉林大学获得学士学位，主要从事数字图像处理、自动驾驶方面的研究。E-mail: renfenglei15@mails.ucas.edu.cn

通讯作者：

吕　游(1988－)，男，吉林松原人，助理研究员，2011 年于吉林大学获得学士学位，2016 年于长春光机所获得博士学位，主要从事目标特性测量、自主导航技术方面的研究。E-mail: lvyou8863@163.com

导师简介：

何　昕(1966－)，男，吉林长春人，研究员，博士研究生导师，1988 年于哈尔滨工业大学获得学士学位，1991 年于长春光机所获得硕士学位，主要从事图像处理、光电测量等方面的研究。E-mail: hexin6627@sohu.com