Computer Science and Application 计算机科学与应用, 2020, 10(11), 1944-1951
Published Online November 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.1011205

Hans汉斯

基于双重注意力特征增强网络的语义分割方法

赵　芮，于晓艳，荣宪伟

哈尔滨师范大学，黑龙江　哈尔滨
Email: 844212391@qq.com

收稿日期：2020年10月22日；录用日期：2020年11月6日；发布日期：2020年11月13日

摘　要

语意分割作为计算机视觉领域的研究热点之一，在地理信息系统、医疗影像分析和机器人等领域有广泛应用。然而现有的语义分割方法主要面临两个挑战，即类内不一致和类间难区分问题。为此，我们提出了一种基于双重注意力特征增强网络的方法来实现语义分割。该方法采用位置注意力模块与通道注意力模块来获取丰富的空间信息与上下文信息，并且在网络末端添加金字塔池化模块来聚合不同区域的上下文信息，提高网络捕获全局信息的能力。最终在标准数据集上的实验结果验证了本文方法的有效性。

关键词

语义分割，双重注意力特征增强网络，位置注意力模块，通道注意力模块

Dual Attention Based Feature Enhanced Networks for Semantic Segmentation

Rui Zhao, Xiaoyan Yu, Xianwei Rong

Harbin Normal University, Harbin Heilongjiang
Email: 844212391@qq.com

Received: Oct. 22nd, 2020; accepted: Nov. 6th, 2020; published: Nov. 13th, 2020

Abstract

As one of the research hotspots in the field of computer vision, semantic segmentation has been widely applied in various fields such as geographic information systems, medical image analysis and robotics. However, contemporary semantic segmentation tasks generally face two challenges,

文章引用: 赵芮, 于晓艳, 荣宪伟. 基于双重注意力特征增强网络的语义分割方法[J]. 计算机科学与应用, 2020, 10(11): 1944-1951. DOI: 10.12677/csa.2020.1011205

namely intra-class inconsistency problem and inter-class indistinction problem. To this end, we solve the semantic segmentation by proposing Dual Attention based Feature Enhanced Networks. In this method, the position attention module and channel attention module are used to obtain rich spatial and context information, and the pyramid pooling module is added at the end of the network to aggregate the context information of different regions, which could improve the capability of the networks to capture global information. Finally, the experimental results on the standard dataset demonstrate the effectiveness of the proposed method.

Keywords

Semantic Segmentation, Dual Attention Based Feature Enhanced Networks, Position Attention Module, Channel Attention Module

1. 引言

语意分割作为像素级别的分类任务，在自动驾驶，人机交互，增强现实等领域有广泛应用[1]。由于深度学习具有拟合能力及表征能力强、灵活性高、应用范围广等优点[2]，因此近年来许多图像语义分割问题使用深度学习来解决。美国伯克利大学研究团队提出了全卷积网络[3]，该网络由编码器和解码器组成，编码器是利用卷积层及池化层对图像进行下采样，解码器是利用反卷积层实现上采样，恢复图像分辨率，实现端到端训练的全卷积网络。为了保证输入的空间分辨率同时增加感受野，Fisher Yu 等人提出了扩张卷积[4]，相对于传统卷积核各个像素是紧密相连的，扩张卷积依据扩张率来控制卷积核中各个像素之间的间隔，将最后几层池化层替换成了扩张率逐渐升高的扩张卷积层来进一步提高算法的精度。华中科技大学王兴刚等人通过改进了非局部神经网络[5]提出了交叉关注语意分割算法[6]，该算法能在更好地捕获上下文语意的同时减少 GPU 运行内存，提高计算效率。然而上述算法并不能在获取足够上下文信息的同时保证精准的空间信息。为了在不损失空间信息的前提下获取充分的上下文信息，一些研究[7] [8]在 U-Net [9]结构的基础上连接来自高阶和低阶的特征来捕获空间信息和上下文信息。此外，BiSeNet [10]设计了包含小步幅卷积层的空间路径来获取空间信息，同时提出了包含下采样策略的上下文路径来获得较大的感受野及上下文信息，最后，引入特征融合模块来融合由上面两路径生成的特征。尽管这些方法可以确保空间及上下文特征的获得，但是它们难以分辩外观不同但是具有相同语义标签的对象。

为此，通过借鉴特征鉴别网络[11]，本文提出了双重注意力特征增强网络用于实现语义分割。如图 1 所示，该网络包括平滑网络与边界网络，前者用于处理类内不一致问题，后者用于解决类间难区分问题。为获取更精准的图像细节信息，本文提出在平滑网络中加入位置注意力模块用于捕获由网络低阶产生的有效的空间特征，依据图像处理中的非局部均值原理，该模块在计算某一特定像素点的特征时，采用对图像中所有点的特征值进行加权平均，其中各个特征分配的权值取决于各个像素之间的依赖关系，同时，利用通道注意力模块来获取由网络高阶产生的精准的上下文信息，通道注意力模块与位置注意力模块构成双重注意力机制来捕获网络不同阶段所产生的有效语义特征。另一方面，我们采用修正残差模块统一实验过程中平滑网络的内部通道数及进一步细化各个阶段产生的语义信息。最后，在残差网络的末端加入金字塔池化模块[12]，以此获取局部及全局有效的语义信息。

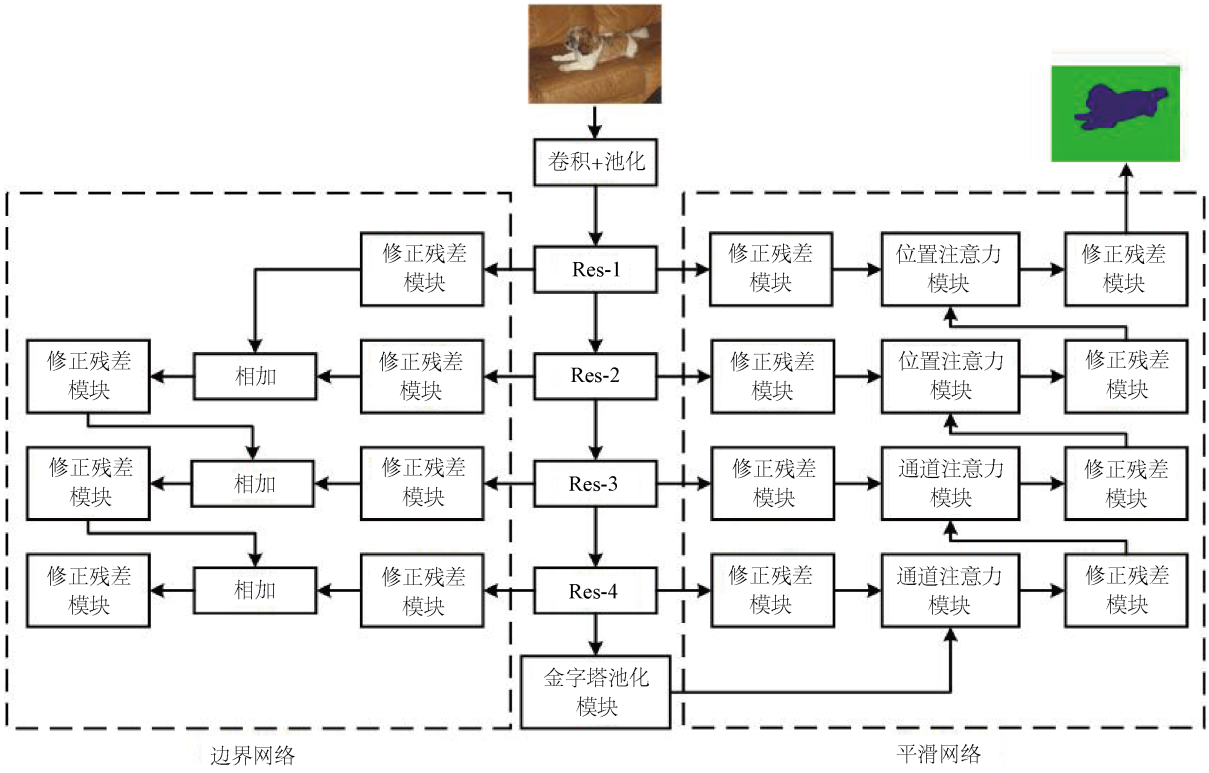


Figure 1. Dual attention based feature enhanced networks
图 1. 双重注意力特征增强网络

2. 位置注意力模块

在本文中，我们引入位置注意力模块来筛选更加有用的特征，通过学习长距离特征之间的依赖关系，位置注意力模块可以弥补由于连续卷积和池化所带来的图像细节信息的损失。受非局部神经网络[5]的启发，我们提出了位置注意力模块，其实现功能如下：(1) 连接高阶特征和低阶特征，实现从高阶到低阶特征的指导功能。(2) 学习特征之间的依赖关系来丰富空间信息，并且改善网络的辨别能力。接下来，我们进一步介绍聚合特征的过程。

非局部模块[5]的思想是位置 i 的特征 x_i 取决于 x_i 和 x_j 之间的依赖性，其中 x_j 代表特征图中所有点的特征。在本文中，采用嵌入式高斯函数来获得依赖关系，如公式(1)所示，并且包含特征图中任意两点之间依赖关系的特征图可以由公式(2)得到。

$$f\left(x_i, x_j\right) = \mathrm{e}^{\theta(x_i)^{\mathrm{T}}\varphi(x_j)} \tag{1}$$

$f\left(x_i, x_j\right)$ 代表位置 i 和 j 位置的特征之间的依赖关系，θ 和 φ 是卷积操作，其中 $\{i, j\} \in R^{C\times W\times H}$。

$$y_{ij} = \frac{1}{C(x)}\sum_{\forall j} f\left(x_i, x_j\right) \tag{2}$$

y_{ij} 代表特征 x_j 对 x_i 的影响，且 $C(X) = \sum_{\forall j} f\left(x_i, x_j\right)$，根据 softmax 函数的定义，公式(2)可以进一步转化为公式(3)。

$$y_{ij} = \mathrm{softmax}\left(\theta(x_i)^{\mathrm{T}}\varphi(x_j)\right) \tag{3}$$

如图 2 所示，通过上述 softmax 函数获得包含各个位置特征之间依赖关系的注意力图，通过公式(4)获得位置注意力模块的最终输出结果。

$$Z = \mathrm{CBR}\left(A_z L\right) + H \tag{4}$$

其中，Z 代表位置注意力模块最终输出结果，A_z 代表注意力图，L 代表经过卷积处理的低阶特征，H 代表来自高阶的特征，CBR 为由卷积、批归一化、Relu 激活函数组成的处理模块。

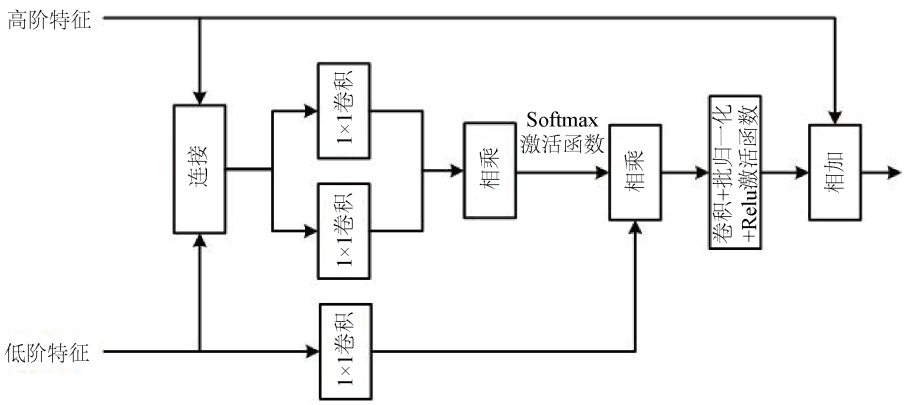


Figure 2. Components of the position attention block
图 2. 位置注意力模块

3. 双重注意力特征增强网络

对于给定分割图像，语义分割效果通常会受到光照及图像尺寸大小的影响，因此捕获更加有效的特征对于实现精准分割至关重要。同时，网络的低级阶段虽然可以获取更多的空间信息，但是因其感受野小而导致上下文信息不足；然而网络的高级阶段因其感受野较大，所以可以获取准确的上下文信息，但是连续的卷积会导致空间信息的损失[13]。因此，为了获得更加有效的信息，我们需要不同的注意力模块来处理不同网络阶段产生的特征。

为解决上述问题，我们提出了双重注意力特征增强网络，该网络主要包括平滑网络与边界网络，其中，平滑网络用于解决将具有相同标签但外观不同的目标分为不同类别的情况，即类内不一致问题，而边界网络主要改善将外观相似不同类别的目标划分为同一类别的现象，即类间难区分问题。同时，在网络中加入双重注意力机制，即位置注意力模块与通道注意力模块，来获取充分的语义信息与上下文信息以便得到精准的分割结果。

平滑网络部分利用位置注意力模块与通道注意力模块来获取充分的语义信息，如图 1 所示，在平滑网络的前两层加入位置注意力模块来捕获更加有效的图像细节及边界等空间信息，在后两层加入通道注意力模块通过对贡献更大的通道特征赋予更大的权值来捕获更加精确的上下文信息。另一方面，如图 2 及图 3 所示，位置注意力模块与通道注意力模块连接相邻阶段的特征，以此实现利用高级别特征来指导低级别特征。与特征鉴别网络相比，我们对于低阶特征不使用通道注意力模块，因为低阶特征的语义在不同的通道上几乎没有差别，本方法使用位置注意力模块替代网络低阶的通道注意力模块以便网络可以更好的捕获有效的空间特征，同时，为了获取不同感受野下的全局信息，本算法在残差网络末端加入金字塔池化模块从而进一步增强网络获取语义信息的能力。最后，如图 4 所示，引入了残差修正模块用于统一平滑网络内部通道数量及进一步整合语义信息。

本文采用特征鉴别网络中的边界网络，该部分的功能是放大类与类之间的特征区别，从而解决类别

之间模型不清晰问题。边界网络采用语义边界来指导特征学习，增强网络对不同类别的区别能力，鉴于网络在不同的阶段具有不同的识别能力，边缘网络利用在低级阶段获得边缘信息进一步细化在高级阶段捕获语义信息，同时利用局部损失函数[14]来监督该部分的输出，以达到更好的提取语义边界的效果。

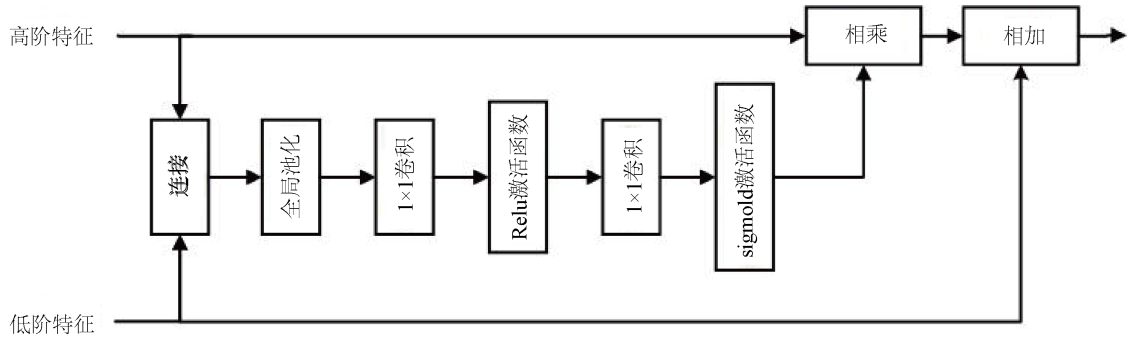


Figure 3. Components of the channel attention block
图 3. 通道注意力模块

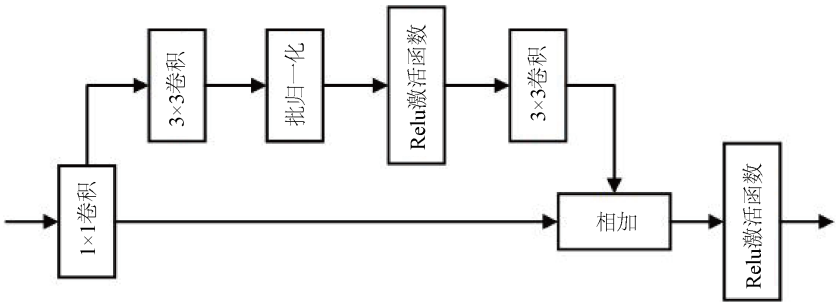


Figure 4. Components of the refinement residual block
图 4. 修正残差模块

在本文中，我们采用深度监督来获得更好的实验结果。交叉熵损失函数可以更好的衡量真实分布和预测的分布的差异情况，其具体计算公式如公式(6)所示。因此对于平滑网络，采用交叉熵损失函数来计算除了全局池化层以外每个阶段的损失，如公式(7)所示。对于边界网络，如公式(8)所示，采用局部损失函数来监督每一阶段的输出。此外，如公式(9)所示，我们引入参数 σ 来平衡平滑网络和边界网络的损失，通过实验表明，当 σ 取值为 0.5 时，可以达到最好的语义分割效果。

$$Loss = -\sum_{i=1}^{m} y_i \log\left(p_i\right) \tag{6}$$

其中 p_i 是该图第 i 类的预测概率值， y_i 是该图第 i 类的标注， y_i 可取值为 0 或 1，其中标注 0 代表该像素点属于第 i 类，反之，标注 1 则代表该像素点不属于第 i 类。

$$SNLOSS_i = CrossEntropyLoss\left(y_{si}; w\right) \tag{7}$$

$$BNLOSS_i = FocalLoss\left(y_{bi}; w\right) \tag{8}$$

$$L = \sum_{i=0}^{3} SNLOSS_i + \sigma \sum_{i=0}^{3} BNLOSS_i \tag{9}$$

其中，i 代表平滑网络与边界网络的层数，其中 $i \in \{0,1,2,3\}$， $SNLOSS_i$ 和 $BNLOSS_i$ 是平滑网络和边界网络第 i 层的损失。

4. 性能评价

4.1. 数据集及参数设置

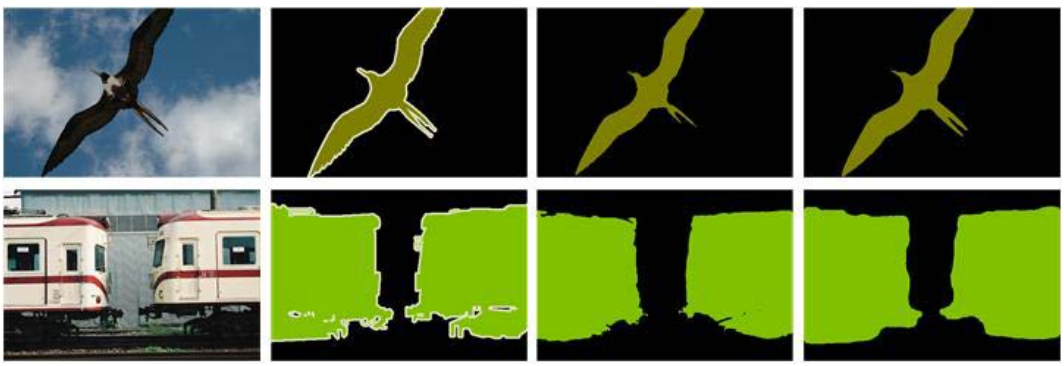
我们使用交并比均值(Mean IOU)作为性能指标来评价语义分割的性能，并使用了标准数据集PASCAL VOC 2012 验证本文所提出方法的性能。

PASCAL VOC 2012：作为语义分割标准数据库，PASCAL VOC 2012 包括 20 个类别以及一个背景，其中包含 1464 张训练图像和 1449 张验证图像。通过使用语义边界数据集[15]对 PASCAL VOC 2012 进行扩充，扩充后的 PASCAL VOC 2012 数据集包含 10582 张训练数据集。

为了防止由于实验数据过少而导致训练过程中出现过拟合的情况，对于输入的训练图像，我们采用均值减法及水平翻转对图像进行预处理。同时，实验中对训练图像进行均值减法和随机水平翻转，同时，为了实现实验数据的扩充，在我们的实验中训练图像被随机按比例缩放，缩放比例设置为{0.5, 0.75, 1, 1.5, 1.75}，最后统一将图像调整为 512 × 512 大小作为最终训练数据输入到网络中。本文采用随机梯度下降[16]作为梯度下降算法来进行网络训练，其中批量大小为 4，动量参数为 0.9，学习率衰减值为 0.0001。实验中采用 poly 学习速度策略，其中动量值设为 0.9，初始学习率为 $1e^{-4}$ 并且在每次迭代后将它乘以 $\left(1-\frac{\text{当前迭代次数}}{\text{总迭代次数}}\right)^{0.9}$。

4.2. 实验结果

借鉴文献[17]的方法，本实验的训练过程使用 PASCAL VOC 2012 的扩充数据集在 ImageNet [18]预训练的 ResNet-101 模型上进行训练，在训练时，采用扩充 PASCAL VOC 2012 的 trainval 数据集作为我们的训练集，同时输入图像的大小裁剪为 512 × 512。在验证过程中，我们使用原始 PASCAL VOC 2012 的 trainval 数据集来验证我们的方法，采用多尺度输入以及水平翻转进一步优化验证数据集。最后，我们将基准网络与本论文提出方法的实验结果进行可视化，如图 5 所示，通过与基准网络 ResNet-101 的输出结果进行对比，本实验在物体边界及图像细节部分的分割结果更加清晰与准确，这是由于注意力模块及金字塔池化模块帮助网络获取更加有效的语义信息，细化分割结果的边界及细节部分。本文方法在 PASCAL VOC 2012 数据集上取得了在比较算法中最佳的性能，如表 1 所示，即最高的 88.4%的交并比均值，这说明本文方法在比较算法中取得了最好的分割效果。



(a) 输入图像　(b) 语义分割标签图像　(c) 基准网络输出图像　(d) 本文网络输出图像

Figure 5. Comparison of segmentation results between Dual Attention based Feature Enhanced Networks and benchmark network

图 5. 双重注意力特征增强网络及基准网络分割结果对比

Table 1. The comparison of the performance in Mean IOU between different algorithms
表 1. 不同算法的交并比均值性能比较

Method	Backbone	Mean IOU (%)
FCN [3]	VGG-16	62.2
ParseNet [19]	VGG-16	69.6
PSPNet [12]	ResNet-101	85.4
Deeplab v3 [20]	ResNet-101	85.7
EncNet [21]	ResNet-101	85.9
DFN [11]	ResNet-101	86.2
Exfuse [22]	ResNet-101	86.2
SDN [23]	Dense-Net-161	86.6
DIS [24]	ResNet-101	86.8
EMANet [25]	ResNet-101	87.7
Ours	ResResNet-101	88.4

5. 结论

本文提出了一种双重注意力特征增强网络来解决当前语义分割面临的挑战任务。该网络引入了位置注意力模块来增强长距离特征之间的依赖关系并且进一步细化了空间特征，利用位置注意力模块与通道注意力模块来增强特征的表征能力，以便获得更加有效的语义特征，同时在网络末端添加金字塔池化模块来提高网络捕获全局场景信息的能力。为了验证提出方法的有效性，我们在 PASCAL VOC 2012 标准数据集上进行了比较实验，实验结果表明位置注意力模块及金字塔池化模块的添加可以明显提高语义分割的效果。为了扩展本文方法的应用领域，未来的工作将着重研究降低双重注意力特征增强网络的计算复杂度方法。

基金项目

本课题是在国家自然科学基金资助项目 61401127 及黑龙江省自然科学基金资助项目 F2018022 的支持下完成的。

参考文献

[1] Garcia-Garcia, A., *et al*. (2017) A Review on Deep Learning Techniques Applied to Semantic Segmentation. *International Conference on Computational Linguistics*, Spain, 22 April 2017, 2132-2144.

[2] 陈一鸣, 彭艳兵, 高剑飞. 基于深度学习的遥感图像新增建筑物语义分割[J]. 计算机与数字工程, 2019, 47(12): 3182-3186.

[3] Long, J., Shelhamer, E. and Darrell, T. (2014) Fully Convolutional Networks for Semantic Segmentation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **39**, 640-651. https://doi.org/10.1109/TPAMI.2016.2572683

[4] Yu, F. and Koltun, V. (2016) Multi-Scale Context Aggregation by Dilated Convolutions.

[5] Wang, X., Girshick, R., Gupta, A., *et al*. (2018) Non-Local Neural Networks. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-22 June 2018, 7794-7803. https://doi.org/10.1109/CVPR.2018.00813

[6] Huang, Z., Wang, X., Huang, L., *et al*. (2019) CCNet: Criss-Cross Attention for Semantic Segmentation. *IEEE International Conference on Computer Vision*, Seoul, 27-28 October 2019, 603-612. https://doi.org/10.1109/ICCV.2019.00069

[7] Chao, P., *et al*. (2017) Large Kernel Matters—Improve Semantic Segmentation by Global Convolutional Network. *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 1743-1751. https://doi.org/10.1109/CVPR.2017.189

[8] Chen, L., *et al*. (2020) ANU-Net: Attention-Based Nested U-Net to Exploit Full Resolution Features for Medical Image Segmentation. *Computers & Graphics*, **90**, 11-20. https://doi.org/10.1016/j.cag.2020.05.003

[9] Ronneberger, O., Fischer, P. and Brox, T. (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Munich, 5-9 October 2015, 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

[10] Yu, C., *et al*. (2018) BiSeNet: Bilateral Segmentation Network for Real-Time Semantic Segmentation. *European Conference on Computer Vision*, Munich, 8-14 September 2018, 334-349. https://doi.org/10.1007/978-3-030-01261-8_20

[11] Yu, C., Wang, J., Peng, C., *et al*. (2018) Learning a Discriminative Feature Network for Semantic Segmentation. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-22 June 2018, 1857-1866. https://doi.org/10.1109/CVPR.2018.00199

[12] Zhao, H., Shi, J., Qi, X., *et al*. (2016) Pyramid Scene Parsing Network. *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 6230-6239. https://doi.org/10.1109/CVPR.2017.660

[13] 翟鹏博, 杨浩, 宋婷婷. 结合注意力机制的双路径语义分割[J]. 中国图象图形学报, 2020, 25(8): 1627-1636.

[14] Lin, T.Y., Goyal, P., Girshick, R., *et al*. (2017) Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **99**, 2999-3007.

[15] Hariharan, B., Arbelaez, P., Bourdev, L.D., *et al*. (2011) Semantic Contours from Inverse Detectors. *IEEE International Conference on Computer Vision*, Barcelona, 6-13 November 2011, 991-998. https://doi.org/10.1109/ICCV.2011.6126343

[16] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) Imagenet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, Vol. 2, 1097-1105.

[17] 熊炜, 童磊, 金靖熠. 基于卷积神经网络的语义分割算法研究[J/OL]. 计算机应用研究, 2020, 38(3): 1-5.

[18] Russakovsky, O., Deng, J., Su, H., *et al*. (2015) ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, **115**, 211-252. https://doi.org/10.1007/s11263-015-0816-y

[19] Liu, W., Rabinovich, A. and Berg, A.C. (2015) ParseNet: Looking Wider to See Better. arXiv preprint arXiv: 1506.04579

[20] Chen, L.C., Papandreou, G., Schroff, F., *et al*. (2017) Rethinking Atrous Convolution for Semantic Image Segmentation.

[21] Zhang, H., Dana, K., Shi, J., *et al*. (2018) Context Encoding for Semantic Segmentation. *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 7151-7160. https://doi.org/10.1109/CVPR.2018.00747

[22] Zhang, Z., Zhang, X., Peng, C., *et al*. (2018) ExFuse: Enhancing Feature Fusion for Semantic Segmentation. *European Conference on Computer Vision*, Munich, 8-14 September 2018, 269-284. https://doi.org/10.1007/978-3-030-01249-6_17

[23] Jun, F., *et al*. (2019) Stacked Deconvolutional Network for Semantic Segmentation. *International Conference on Image Processing*, Taipei, 22-25 September 2019, 3085-3089. https://doi.org/10.1109/TIP.2019.2895460

[24] Luo, P., Wang, G., Lin, L., *et al*. (2017) Deep Dual Learning for Semantic Image Segmentation. *IEEE International Conference on Computer Vision*, Venice, 22-29 October 2017, 2737-2745. https://doi.org/10.1109/ICCV.2017.296

[25] Xia, L., Zhong, Z.S., Wu, J.L., *et al*. (2019) Expectation-Maximization Attention Networks for Semantic Segmentation. *IEEE International Conference on Computer Vision*, Seoul, 27-28 October 2019, 9166-9175.