计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(3):58−69 [doi: 10.15888/j.cnki.csa.009008]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

多尺度多阶段特征融合的带噪图像语义分割①

黄　琳[1], 陈　飞[1], 曾勋勋[2]

[1](福州大学 计算机与大数据学院/软件学院, 福州 350108)
[2](福州大学 数学与统计学院, 福州 350108)
通信作者: 陈　飞, E-mail: chenfei314@fzu.edu.cn

摘　要: 在图像的采集过程中, 图像往往会带有一定的噪声信息, 这些噪声信息会破坏图像的纹理结构, 进而干扰语义分割任务. 现有基于带噪图像的语义分割方法, 大都是采取先去噪再分割的模型. 然而, 这种方式会导致在去噪任务中丢失语义信息, 从而影响分割任务. 为了解决该问题, 提出了一种多尺度多阶段特征融合的带噪图像语义分割的方法, 利用主干网络中各阶段的高级语义信息以及低级图像信息来强化目标轮廓语义信息. 通过构建阶段性协同的分割去噪块, 迭代协同分割和去噪任务, 进而捕获更准确的语义特征. 在 PASCAL VOC 2012 和 Cityscapes 数据集上进行了定量评估, 实验结果表明, 在不同方差的噪声干扰下, 模型依旧取得了较好的分割结果.

关键词: 语义分割; 图像去噪; 协同任务; 特征融合; 注意力机制

引用格式: 黄琳,陈飞,曾勋勋.多尺度多阶段特征融合的带噪图像语义分割.计算机系统应用,2023,32(3):58–69. http://www.c-s-a.org.cn/1003-3254/9008.html

Semantic Segmentation of Noisy Images with Multi-scale and Multi-stage Feature Fusion

HUANG Lin[1], CHEN Fei[1], ZENG Xun-Xun[2]

[1](College of Computer and Data Science/College of Software, Fuzhou University, Fuzhou 350108, China)
[2](School of Mathematics and Statistics, Fuzhou University, Fuzhou 350108, China)

Abstract: In the process of image acquisition, the image often contains certain noise information, which will destroy the texture structure of the image and thus interfere with semantic segmentation tasks. Most of the existing semantic segmentation methods based on noisy images adopt models featuring first denoising and then segmentation. However, they often lead to the loss of semantic information in denoising tasks, which thus affects segmentation tasks. To solve this problem, this study proposes a multi-scale and multi-stage feature fusion method for semantic segmentation of noisy images, which uses the high-level semantic information and low-level image information of each stage in the backbone network to enhance the semantic information of target contours. By constructing a staged collaborative segmentation denoising block, collaborative segmentation and denoising tasks are iterated, and then more accurate semantic features are captured. In addition, quantitative evaluation is carried out on PASCAL VOC 2012 and Cityscapes datasets. The experimental results show that the model still achieves positive segmentation results under the noise interference of different variances.

Key words: semantic segmentation; image denoising; collaborative task; feature fusion; attention mechanism

1　引言

语义分割的目标是确定每个像素点的类别 (如属于背景、人或车等), 从而将一些原始图像转换为具有突出显示的感兴趣区域的掩模. 目前, 许多先进的分割

① 基金项目: 国家自然科学基金 (61771141); 福建省教育厅中青年教师教育科研项目 (JAT190020); 福建省自然科学基金 (2021J01620)
收稿时间: 2022-08-15; 修改时间: 2022-09-15; 采用时间: 2022-09-30; csa 在线出版时间: 2022-12-16
CNKI 网络首发时间: 2022-12-19

方法已经应用在各个领域, 如自动驾驶、场景解析、目标检测和人机交互等领域[1].

最近, 深度神经网络在语义分割任务中已经取得显著的成功. 例如 PSANet[1]、DenseAPP[2] 以及 DANet[3] 等. 但是这些网络成功的前提, 依赖于高质量的训练数据集, 即干净无噪声的图像. 然而, 在实际应用中, 由于拍摄时的环境、聚焦失败和机子的抖动, 图像采集设备采集到的图像具有不同程度的噪声信息, 例如, 高斯噪声、短噪声和热噪声等[4]. 这些噪声信息是不可控的, 再精密的图像采集设备也无法控制拍摄现实图像时的环境. 由于噪声信息的复杂性以及不可控性, 在遥感图像、医学图像、视频建模、灾害监测等现实场景下的语义分割面临很大的挑战. 例如, 遥感图像中, 由于传感器的周期性偏移、电磁干扰, 所产生的随机噪声都会掩盖数字图像中真正的辐射信息; 在医学图像中, 以最常见的 CT 图像为例, 实验观测、硬件系统限制将不可避免地带来高斯噪声, 使得图像具有斑驳、颗粒状、纹理的外观, 覆盖并降低图像内某些特征的可见性, 减低模型分割的准确性; 在视频建模和灾害监测中, 无人机航拍成为较为常见的监测方式, 但拍摄角度引起的亮度不均以及传感器长期工作、温度过高, 都会产生高斯噪声, 影响建模和监测的质量. 以上问题都可概述为噪声信息往往会覆盖图像的一些小纹理结构, 从而降低了语义分割模型的能力. 当用带有噪声的数据集去训练当前主流的语义分割模型的时候, 分割精度明显下降.

为了降低噪声对语义信息的干扰, 最直接的方法是进行语义分割前, 完成图像去噪任务, 采取一步到位去噪加一步到位分割的串联方法. CNSOLT[5] 基于不可分过采样重叠变换进行扩展, 结合迭代收缩/阈值算法, 作为一种稀疏表示对雷达图像进行去噪. DCNNM[6] 采用多通道跨层聚合图像的高维全局语义信息与局部特征细节, 实现气象雷达噪声图像语义分割. 但是该方法未考虑到去噪对下游任务的促进作用. CNN DAE[7] 通过卷积层构建去噪自编码器用于小样本的医学图像去噪. Eformer[8] 利用 Transformer[9], 将可以学习的 Sobel 滤波器用于边界增强, 提高医学图像去噪的性能. 但是, 目前去噪网络都追求于去噪后的视觉效果, 并没有考虑到语义分割等下游任务. 该方法虽然可以去除噪声, 但是也造成了一些小纹理语义信息的缺失, 仍然不能达到使用干净图像训练分割模型的结果. 图 1 是关于语义分割的结果说明. 其中, 图 1(a) 和图 1(b) 分别表示噪声图像和 ground truth, 图 1(c) 和图 1(d) 分别表示 DANet 用干净图像分割和带噪图像分割的结果, 图 1(e) 表示为 DANet 加上去噪模块的串联方法的分割结果. 从图 1(d) 可以看出, 当用带有噪声的数据集去训练 DANet 的时候, 噪声信息严重干扰了图像纹理结构, 不仅了影响目标边界部分, 使得上下文信息无法准确获取, 还导致目标区域被错误划分, 语义分割精度明显下降. 从图 1(e) 可以看出, 为语义分割模块加上去噪模块, 形成简单串联架构, 虽然可以削弱噪声信息对目标纹理结构的破坏, 但不可避免的造成了部分纹理信息的缺失, 导致在后续一步到位的语义分割中产生错误定位并使得目标区域分割不完整.

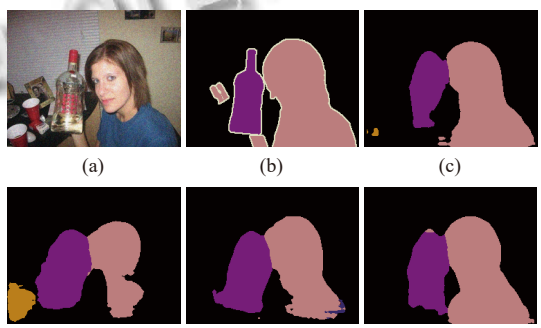


图 1 语义分割结果说明

为了解决上述问题, 本文提出了一种端到端的多尺度多阶段特征融合的带噪图像语义分割的方法. 该方法结合多尺度思想, 并引入空洞卷积, 有效解决了小纹理被噪声覆盖的问题. 并且, 该方法利用了主干网络每阶段所提取的语义信息, 通过阶段性分割促进去噪模块 (staged segmentation guided denoising module, SSGDM), 将分割结果、阶段语义特征跟噪声信息有效结合, 增强边缘纹理的语义信息, 促使分割进一步帮助去噪. 之后, 将去噪图像输入到语义分割模块中, 生成阶段性分割结果. 最后, 将多阶段的分割结果进行叠加, 输出最终的分割结果, 结果如图 1(f) 所示. 可以看出, 该方法的分割质量趋近于干净图像分割的结果, 模型不仅能识别更多正确的语义信息, 而且目标边界较为清晰明确. 通过一系列实验, 证明了该方法的优越性.

本文的主要贡献总结如下.

(1) 结合多尺度的思想, 并利用空洞卷积, 较好地解决了图像小纹理结构被噪声覆盖的问题. 并且, 提出阶段性分割促进去噪模块 (SSGDM), 将每一阶段的语义信息、去噪环节中的低层特征以及分割结果进行有效结合, 使得语义分割促进图像去噪.

(2) 提出一种利用主干网络中各个阶段提取到的高级语义特征以及低级图像特征来强化目标轮廓语义信息的方法. 该方法通过分阶段对带噪图像进行去噪和分割, 从不同层次上帮助去噪与分割任务.

(3) 提出了一种多尺度多阶段特征融合的带噪图像语义分割网络, 该方法避免了多次训练模型的过程. 实验结果表明, 在不同噪声水平的数据集上, 该方法有效提升了带噪图像语义分割的精度.

2 相关方法

2.1 图像去噪

图像去噪是诸多学者一直以来致力于解决的问题. 目前图像去噪方法主要包括基于模型的方法和基于深度学习的方法. 基于模型的方法常常利用一些先验知识, 其中包括非局部自相似性[10–12]、稀疏性[13,14]、局部平滑[15] 等. 而基于深度学习的去噪方法正在迅速发展, 成为主流趋势. DnCNN[16] 引入了残差学习和批量归一化来实现端到端去噪. FFDNet[17] 引入噪声水平图作为输入, 增强了网络对非均匀噪声的灵活性. CBDNet[4] 训练两个子网络, 使用噪声估计子网络估计噪声水平图, 再通过非盲去噪子网络去除噪声. PD[18] 采用 pixel-shuffle 下采样策略将真实噪声近似为加性高斯白噪声. RIDNet[19] 提出了一种基于特征关注的单级去噪网络用于真实图像去噪. SADNet[20] 提出了一种结合可变卷积的带有上下文块的编码器-解码器结构, 以用来捕获多尺度信息. MalleNet[21] 提出了一种基于可塑卷积 (malleable convolution) 的快速高质量图像去噪网络. Neighbor2Neighbor[22] 提出了一种仅需含噪图像的训练策略, 考虑到图像相邻的像素来设计采样器, 构造相似的噪声图像, 最后引入正则项的方式解决了采样过程中图像过于平滑的问题. DeamNet[23] 将传统一致性先验和非线性滤波算子相结合, 提出自适应一致性先验, 再将其引入到最大后验中, 实现基于模型去噪方法. WINNet[24] 基于小波方法原理, 实现用于去噪的稀疏编码过程, 并通过软阈值来自适应调整估计噪声水平, 提高模型的泛化能力. Restormer[25] 基于 Transformer 进行改进, 通过多个深度分离卷积在通道维度上进行自注意力, 以此聚合局部和非局部的交互信息. 以上这些网络都只关注自身任务, 没有考虑与下游任务的彼此协同性.

2.2 语义分割

最近, 基于全卷积网络和编码器-解码器结构相结合的方法, 在各种基准上取得了很好的效果. 但是, 卷积滤波器的局部性质限制了对图像中的全局信息的访问. 为了解决这个问题, DeepLab[26–28] 提出空洞卷积扩大卷积的感受野, 设计空间金字塔池化捕捉多尺度上下文信息. DANet 提出双重注意力模块, 分别在空间维度和通道维度上对语义相关性进行建模, 来聚合高级语义信息. Channel-attention U-Net[29] 提出跨层特征融合模块, 结合多尺度特征, 促使低级特征向高级特征学习, 以解决模型识别小目标的难题. SCARF[30] 引入类学习权值, 自适应平衡全局上下文和局部特征, 通过降低类别噪声来细化语义分割模型. SAGNN[31] 基于图神经网络, 提出了一种新的自节点协作机制, 通过多尺度特征图节点间的聚合特征, 捕获高级的语义依赖, 以此判断节点的特征表示. DPL-Dual[32] 提出了一种双路径学习的框架, 在训练时采用两个互补交互的模型相互促进学习. BiseNet V2[33] 提出了双边分割网络, 并行处理空间细节和类别语义信息, 平衡分割精度和推理速度. FLANet[34] 提出了完全注意力模块, 通过单个相似性图来结合空间和位置编码, 来解决在非局部的自注意模块中特征缺失问题. CIRKD[35] 结合知识蒸馏的方法, 通过跨图像来构建像素与像素之间以及像素与类别区域之间的相似性矩阵, 构建图像间的依赖关系, 以此捕获更广上下文信息, 产生更好的结构化语义信息. 由于卷积神经网络无法对输入图像进行全局理解, 受 VIT[36] 的启发, SETR[37] 使用 Transformers 取代基于堆叠卷积层的编码器, 将输入图像视为图像 patch 序列进行提取特征, 提出了直接、渐进式、多级特征聚合的上采样方式. 并且, 从 sequence-to-sequence的角度重新定义了图像语义分割问题. MPViT[38] 提出了多路径结构的多尺度嵌入方法, 实现了对多尺度信息的利用, 同时将图像分为多个尺度后进行标记, 实现全局自我注意, 通过聚合所生成的特征, 对相同的像素表示精细和粗糙特征. 以上的方法, 都是通过增强卷积感受野来捕获更丰富的上下文信息, 只针对干净图像进行训练, 对于噪声图像来说, 噪声信息可能会影响模型的表征学习能力.

2.3 协同去噪和分割

目前, 如何更好地对噪声图像进行语义分割正面临着非常大的挑战. DMS[39] 最先将低级图像任务和高级图像任务结合在一起. 该方法将图像去噪和语义分割任务的两个模块分别级联, 并使用联合损失通过反向传播更新去噪网络, 促使去噪效果更好. 可以看出,

该方法虽然考虑了分割对去噪的影响, 但是并没有考虑到一步到位分割会导致错误定位的情况, 从而不能很好地协助去噪任务. 在此基础上, SDABN[40] 基于交替 Boosting 的思想来实现语义分割与图像去噪的协同任务, 将 SFT 层嵌入去噪网络中, 使得语义分割的结果信息对去噪产生影响. 该方法虽然在去噪和分割之间找到了协同点, 但是其依赖于多个子模型, 需分别训练去噪和分割模型, 模型数量多, 参数量大. DDep[41] 对语义分割的解码器进行分析, 将去噪用于解码预训练, 结合迁移学习连接到扩散概率模型中, 通过随机初始化来获得去噪的显著受益. 该方法侧重于研究分割对去噪的促进作用, 同样未考虑对下游分割任务的影响. FECC-Net[42] 将协同任务应用于医学脑部图像, 通过可分离卷积的叠加组合构建增强编码器, 保留浅层特征和图像属性, 与深度语义特征融合, 有助于病灶边界的恢复.

在本文研究中, 提出了一种新颖高效的带噪图像语义分割方法, 通过融合多尺度特征, 扩大感受野捕获上下文信息的同时, 减少了噪声的干扰, 使得模型的高层特征和低层特征有效融合, 以实现分割任务和去噪任务相互促进的目的.

3 带噪图像语义分割

3.1 整体网络架构

传统的带噪图像语义分割, 使用简单的串联架构. 在先完成图像去噪的基础下, 再进行语义分割. 这种架构相比于直接进行带噪图像语义分割, 降低了噪声影响, 能够识别较准确的特征. 但却忽略了在去噪的同时, 也减少了类语义信息, 导致在分割任务中, 目标区域被错误划分. 为了解决以上问题, 不仅结合多尺度特征融合思想, 并通过分割协助去噪的形式, 促使去噪模块关注于小目标以及边界特征, 进一步确保类语义信息的完整性; 同时, 通过分阶段进行分割, 融合主干网络多阶段提取的目标语义特征, 以提高类分割准确度. 结合以上思想, 本文提出了多尺度多阶段特征融合的带噪图像语义分割网络 (noisy images segmentation network, NISNet). 网络结构如图 2 所示.

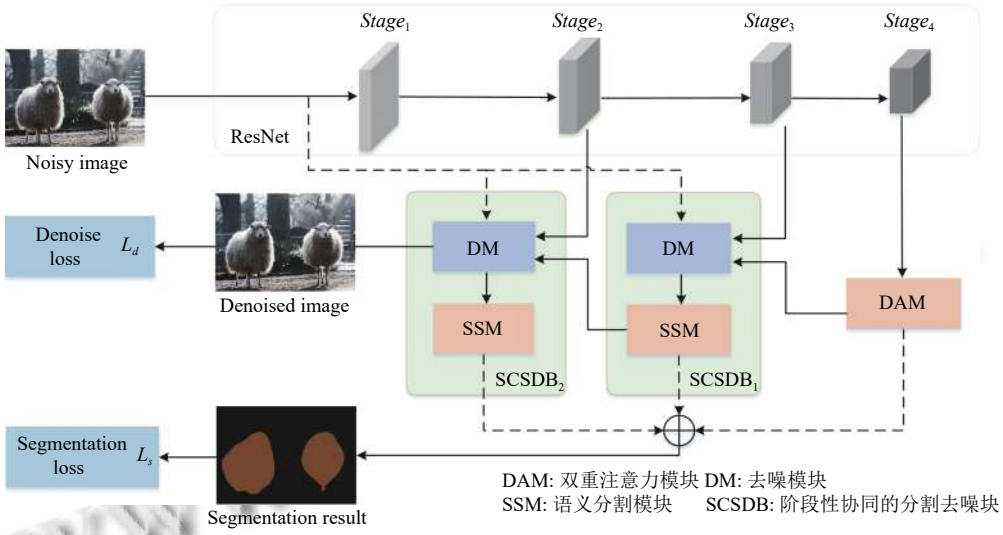


图 2 NISNet 结构示意图

整体网络架构可以看成 3 部分, 主干网络、双重注意力模块 (DAM) 以及阶段性协同的分割去噪块 (staged collaborative segmentation and denoising block, SCSDB). 为了解决低级去噪任务和高级分割任务之间的协同作用, 提出了 SCSDB 模块. 其中, SCSDB 由去噪模块 (denoising module, DM) 和语义分割模块 (semantic segmentation module, SSM) 组成. 由于图像的噪声信息对分割产生了极大的干扰, 而随着网络模型的加深, 层次越深的卷积能检测到更高级的语义信息. 因此, 将高级语义特征跟主干网络的$Stage_2$、$Stage_3$所提取的特征相结合, 获取上下文信息, 依次输入到 SCSDB 中. 首先, 将噪声图像$I \in R^{3\times H\times W}$输入到主干网络中, 经过卷积提取 4 个阶段的特征图$Stage_i$, 然后将$Stage_i$的特征图输入到 DAM 模块中, 得到初步的分割特征S_1, 再将$Stage_3$和S_1作为监督信息输入到 SCSDB中, 协同去噪和分割任务. 在 SCSDB 中, DM 模块对I进行去噪, 得到去噪特征图D_1, 再通过语义分割模块 SSM 得到进一步的分割特征S_2, 完成 SCSDB 模块. 然后, 将$Stage_2$

和S_2输入到 SCSDB 中, 重复以上操作, 得到去噪特征D_2和分割特征S_3. 最终, 将S_i进行加权, 输出最后的特征图S, 完成语义分割任务.

3.2 阶段性协同的分割去噪块 (SCSDB)

本节分析阶段性协同的分割去噪块 (SCSDB), 结构如图 2 中所示. SCSDB 主要分为两个部分, 去噪模块 (DM) 和语义分割模块 (SSM).

SCSDB 的主要目的是促进分割任务和去噪任务之间的协同作用. 在 DM 中, 将上一阶段语义分割所得到的结果跟噪声图像进行融合, 使得去噪过程更关注于分割目标区域的低级特征, 解决去噪环节中导致目标语义信息丢失的问题. 同时, 为了捕捉未被分割任务识别到的区域信息, 网络根据主干网络的各个阶段进行划分, 以此聚合多阶段特征. 最后, 将 SCSDB 的多阶段结果进行融合, 目的是为了减少不同阶段产生的噪声干扰, 尽可能捕获到更准确的目标轮廓信息.

3.2.1 去噪模块 (DM)

本节主要介绍去噪模块 (DM), 该模块使用编码--解码架构, 即 U 型网络架构, 模块结构如图 3(a) 所示.

在编码阶段, 对于输入图像$R^{3\times H\times W}$, 通过多次卷积核为 3, *padding*为 1 的卷积操作以及采用全局平均池化的下采样方式, 得到特征图$R^{4C\times H/4\times W/4}$, 获取到噪声图像的低级特征. 在解码阶段, 以往的去噪网络往往直接通过卷积以及反卷积进行解码, 完成去噪任务.

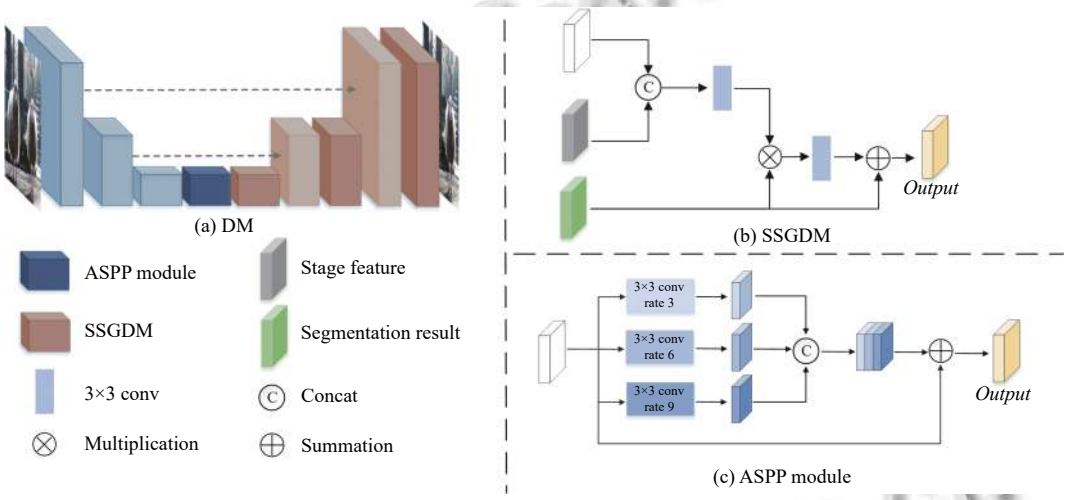


图 3 去噪模块示意图

然而, 为了使高级的语义特征能融入去噪模块, 提出了阶段性分割促进去噪模块 (SSGDM), 利用 SSGDM 来帮助去噪模块更好的降低噪声对目标轮廓语义信息的干扰. 通过该方法, 聚合底层图像特征, 保证低级信息的完整性. 同时, 通过分割结果作为监督, 使得网络更关注于目标区域内的部分. 该方法所生成的新的特征图, 精准的提取目标边缘信息, 减少无关上下文所带有的噪声对目标区域的影响.

在 SSGDM 中, 如图 3(b) 所示, x表示去噪网络每个阶段所生成的特征图$R^{C\times H\times W}$, *Stage*为主干网络所提取的特征, S_{out}为上一阶段语义分割的输出$R^{C_2\times H\times W}$, 其中C_2为类别数, *output*为融合后输出的特征图$R^{C\times H\times W}$. SSGDM 可以用式 (1) 表示为:

$$output = F\left(F\left(cat\left(x, Stage\right)\right)\cdot S_{\text{out}}\right) + S_{\text{out}} \tag{1}$$

其中, $F(\cdot)$表示卷积运算, $cat(\cdot)$表示维度拼接.

噪声信息往往会破坏图像的纹理特征, 而纹理特征在语义分割中有着至关重要的作用. 由于去噪网络作为浅层网络, 几何表征能力强, 容易破坏目标的语义信息, 影响分割质量. 因此, 为了进一步获取更多的上下文信息, 引入了空洞空间卷积池化金字塔模块 (ASPP module), 结构如图 3(c) 中所示. 利用不同膨胀因子 (rate) 的空洞卷积, 融合多尺度信息, 通过不同尺度组合出更多的感受野区域, 引入上下文信息. 同时, 将高层特征和低层特征进行融合, 丰富了预测结果.

3.2.2 语义分割模块 (SSM)

本节主要介绍语义分割模块 (SSM), 此模块主要利用双重注意力机制[3], 通过捕捉丰富的上下文相关性来实现语义分割任务. 结构如图 4 所示.

首先, 将经过 SSGDM 后的去噪结果图D和主干网络阶段特征图进行拼接操作, 通过 1×1 卷积来细化语

义特征, 然后将其分别传入空间注意力机制和通道注意力机制中, 进一步捕获语义特征的全局依赖关系. 最后将两个注意力机制的输出进行相加, 由此输出全局上下文语义信息.

空间注意力机制如图 4(a) 所示, 首先将原始特征图$R^{C\times H\times W}$通过*reshape*分别得到特征图$A^{C\times HW}$, 特征图$B^{HW\times C}$, 再进行矩阵乘法, 经过*Softmax*层计算得到空间注意力特征图$M^{HW\times HW}$, 可表示为式 (2):

$$M_{ji}=\frac{e^{B_i\cdot A_j}}{\sum_{i=1}^{HW}e^{B_i\cdot A_j}} \tag{2}$$

其中, M_{ji}表示特征图中第i个位置和第j个位置的联系. 之后, 重塑原始特征图$R^{C\times H\times W}$并与M矩阵相乘, 并将其重塑为与原始特征图大小一致的特征图$P^{C\times H\times W}$. 最后, 将其乘以一个尺度参数λ, 将λ初始化为 0, 并通过学习不断分配更多的权值. $P^{C\times H\times W}$用式 (3) 表示:

$$P_j=\lambda\sum_{i=1}^{HW}\left(C_iM_{ji}\right) \tag{3}$$

通过以上操作, 可以选择性的加强全局区域下相似语义特征之间的关系, 进而在全局区域内融合相似特征.

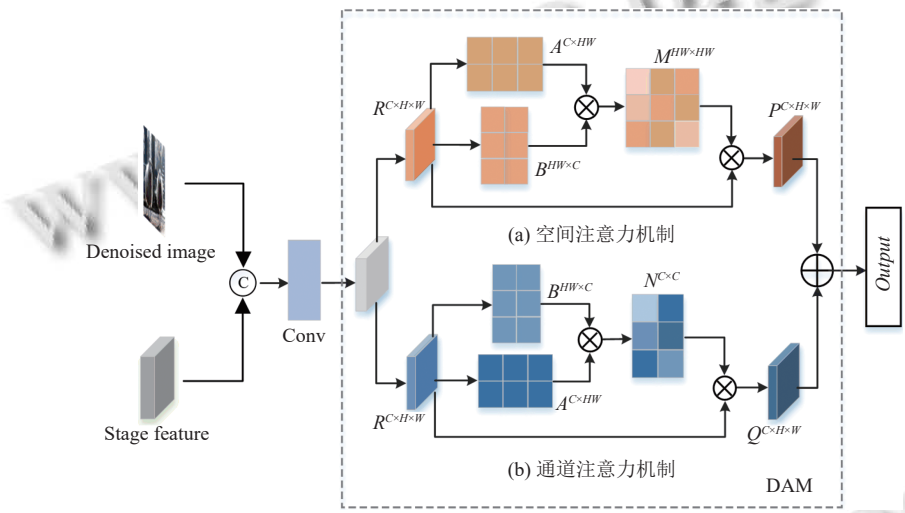


图 4 语义分割模块结构示意图

通道注意力机制如图 4(b) 所示, 与空间注意力机制相似, 通过*reshape*和矩阵乘法后, 经过*Softmax*层计算得到通道注意力特征图$N^{C\times C}$, 可表示为式 (4):

$$N_{ji}=\frac{e^{A_i\cdot B_j}}{\sum_{i=1}^{HW}e^{A_i\cdot B_j}} \tag{4}$$

其中, N_{ji}表示了第i个通道对第j个通道的影响. 同理, 与原特征图相乘, 引入尺度参数之后, 将其乘以一个尺度参数μ, 最终的输出$Q^{C\times H\times W}$可用式 (5) 表示:

$$Q_j=\mu\sum_{i=1}^{HW}\left(N_{ji}C_i\right) \tag{5}$$

通过以上操作, 使得各个通道之间能产生全局的关联, 获得更强的语义响应的特征.

3.3 损失函数

为了使去噪后的图像更接近 ground truth, 进一步促进分割, 利用均方差损失 (mean square error) 来优化去噪模块的参数. 均方差损失L_d可表示为:

$$L_d=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{6}$$

其中, y_i表示像素i的 ground truth, $\hat{y}_i$表示像素i的概率估计, n表示像素点的数量.

在噪声图像中, 由于噪声信息的干扰, 容易导致前景和背景误识别, 从而导致信息不平衡问题, 使得训练网络陷入局部最优的情况. 因此, 为了解决该问题, 在语义分割模块中, 提出了混合损失来学习语义分割模型的最优参数. 由于交叉熵损失可以用来评判预测结果和真实结果的相似性, 并通过*Sigmoid*函数来避免去噪模块中均方差损失带来的影响, 因此, 选用交叉熵损失来避免局部最优问题. 同时, 平均交并集 (mean intersection over union, *mIoU*) 作为分割模块的评判指标, 可

以计算图像中的预测区域和真实区域之间的 *IoU*, 并使用 *IoU loss* 来对模型进行迭代优化.

因此, 分割模块是通过交叉熵损失 (cross entropy loss)、以及 *mIoU* 损失 (*mIoU loss*) 得到的混合交叉熵损失L_S来学习语义分割模型的最优参数, 交叉熵损失L_{S_1}为式 (7) 所示:

$$L_{S_1} = -\frac{1}{p}\sum_{i=1}^{p}\log f_i\left(y_i^*\right) \tag{7}$$

其中, p表示一张图片的像素数量, y_i^*表示像素i的 ground truth 类别, $f_i(y_i^*)$表示像素i的概率估计. *mIoU* 损失L_{S_2}为式 (8) 所示:

$$L_{S_2} = 1 - \frac{|X \cap Y|}{|X| + |Y| - |X \cap Y|} \tag{8}$$

其中, X表示预测的像素集合, Y表示 ground truth 的像素集合. 混合交叉熵损失L_S为式 (9) 所示:

$$L_S = L_{S_1} + L_{S_2} \tag{9}$$

3.4 算法流程

算法 1 展示了所提出网络结构的训练步骤. 首先通过主干网络提取各阶段特征, 再使用语义分割模块 SSM 得到初步分割结果. 最后经过多次 SCSDB, 将多次迭代分割结果融合, 得到最终的分割结果. 为了训练去噪模块, 使用干净图像y与最后一次 SCSDB 中的去噪模块输出y_2求均方差损失. 语义分割的损失函数使用混合交叉熵损失. 算法流程如算法 1 所示.

算法1. 训练NISNet

输入: 噪声图像x, 干净图像y, 分割标签z

输出: 分割结果$\hat{z}$

1) 将噪声图像x输入到主干网络ResNet50中, 提取各阶段的特征$Stage_1$–$Stage_4$;

2) 将主干网络$Stage_4$生成的特征图f_4输入到双重注意力模块DAM中, 细化特征, 输出初步的分割结果z_1;

3) 在块$SCSDB_1$中, 以噪声图像x作为输入图像, 将主干网络$Stage_3$生成的特征图f_3、初步分割结果z_1作为去噪模块DM的辅助信息, 在解码阶段, 设计SSGDM模块, 通过级联拼接、卷积以及残差连接等操作, 聚合多阶段特征, 强调高级语义信息所在区域, 以此对去噪任务起到监督作用, 生成新的去噪特征图y_1, 再通过SCSDB中的分割模块SSM, 聚合阶段特征和去噪特征, 强调图像的边界信息, 最后经过双重注意力模块, 生成新的分割结果z_2, 完成第一次SCSDB迭代.

4) 在块$SCSDB_2$中, 重复步骤3)操作, 将主干网络$Stage_2$生成的特征图f_2初步分割结果z_2作为去噪模块DM的辅助信息, 通过去噪模块DM后会生成新的去噪特征图y_2, 通过分割模块SSM会生成新的分割结果z_3, 完成第2次SCSDB迭代.

5) 将阶段性的分割结果$z_1z_2z_3$进行叠加, 生成多阶段特征融合的语义分割结果$\hat{z}$.

6) 最后, 将生成的去噪特征图y_2与干净图片y求均方差损失, 将语义分割结果$\hat{z}$与分割标签z求混合交叉熵损失, 通过损失反传更新模型参数.

4 实验

4.1 实验细节

在本节中, 主要介绍算法的实现细节. 在 PASCAL VOC 2012[43] 官方数据集中, 共有 4 369 张图像. 其中, 有 1 464 张图像用于训练, 1 449 张图像用于验证, 1 456 张图像用于测试. 通过设置 1 个背景类和 20 个前景类别来评估所提出的网络架构. 为了证明模型的鲁棒性, 还在 Cityscapes[44] 数据集上进行了实验. Cityscapes 数据集有 5 000 张在城市环境中驾驶场景的图像, 包含 2 975 张训练集图像, 500 张图像验证集, 1 525 张测试集图像, 具有 19 个类别的密集像素标注.

在实验中, 为 PASCAL VOC 2012 数据集和 Cityscapes 数据集随机添加了标准差范围为 [0, 30] 的高斯噪声, 并使用 ResNet50 作为主干网络. 在训练过程中, 将所有的图片裁剪为 480×480 作为网络输入. 在实验过程中, 分别提取主干网络的$Stage_2$、$Stage_3$以及$Stage_4$的语义信息, 将其作为 SSGDM 的输入. *Batch_size* 设为 4, 优化器使用 SGD, 学习率策略使用 Warmup, momentum 为 0.09. 当训练 PASCAL VOC 2012 数据集时, epoch 为 50, 初始的学习率为 0.000 1. 当训练 Cityscapes 数据时, epoch 为 120, 初始的学习率为 0.01.

4.2 评判指标

最后, 实验的评判指标使用平均交并集 (mean intersection over union, *mIoU*) 以及像素准确度 (pixel accuracy, *PixAcc*) 作为评估指标. 其中, *PixAcc*表示的是标记正确的像素占总像素的比例, *mIoU*[45] 是语义分割的标准指标, 表示两个集合的交集和并集之比, 公式如下:

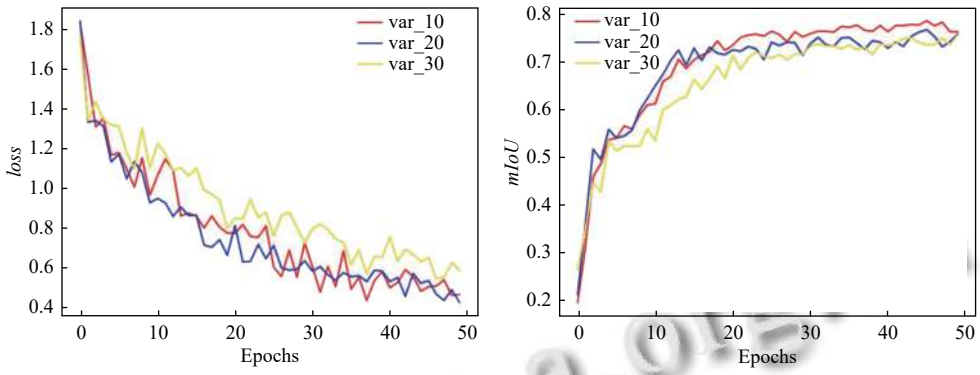
$$PixAcc = \frac{\sum_{i=0}^{k} P_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k} P_{ij}} \tag{10}$$

$$mIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{P_{ii}}{\sum_{j=0}^{k} P_{ij} + \sum_{j=0}^{k} P_{ji} - P_{ii}} \tag{11}$$

其中, k表示类别数, P_{ij}表示原属于第i类但被预测为类j的像素数量, P_{ii}表示识别正确的像素数量.

图 5 给出了 NISNet 模型在 PASCAL VOC 2012 数据集上的训练过程曲线图, 图中存在损失函数和验证集 *mIoU* 指标的变化曲线. 图 5(a) 所示, 是损失函数

随着 epoch 的变化曲线, 横坐标为迭代次数 epochs, 纵坐标为训练时每个 epoch 的平均损失. 图 5(b) 所示, 是评判指标 *mIoU* 随着 epoch 的变化曲线, 纵坐标为每次训练 epoch 所保存模型在验证集上的精度. 可以看出, 随着 epoch 的增长, 损失函数最终趋于收敛状态, 评判指标 *mIoU* 也趋近于收敛, 说明模型训练的可行性.



(a) Train_*loss* (b) Validation_*mIoU*

图 5 训练过程曲线图

4.3 消融实验

本节主要介绍 NISNet 模型的消融实验, 总结出模型的最佳设置. 本节在高斯噪声标准差σ为 25 的 PASCAL VOC 2012 数据集上进行一系列消融实验.

利用分割和去噪相互协同的方法进一步细化语义特征, SSGDM 模块正是去噪和分割模块特征融合的核心. 实验结果如表 1 所示, 其中 $Stage_{256}$、$Stage_{128}$ 和 $Stage_{64}$ 分别指在解码阶段中通道数为 256、128 以及 64 的最后一层卷积层后引入 SSGDM 模块.

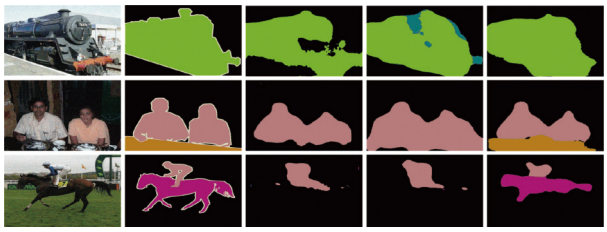
表 1 关于 SSGDM 的消融实验, 标准差σ为 25

$Stage_{256}$	$Stage_{128}$	$Stage_{64}$	*mIoU*	*PixAcc*
√			74.9	94.2
√	√		75.5	93.7
√	√	√	**75.7**	**94.9**

根据表 1 可以得出, 分别在去噪网络的每一个解码阶段使用 SSGDM, 分割精度显著提升. 图 6 进一步对比了不同阶段使用 SSGDM 的分割效果图, 可以看出 SSGDM 只迭代 1 次和 2 次时, 不仅出现类别丢失, 而且目标轮廓分割不完整. 而在去噪网络的各个解码阶段都应用 SSGDM 时, 可以较好地弥补在去噪过程中丢失的语义信息, 增强目标区域的语义信息, 有效提升了分割效果.

NISNet 通过 SCSDB 实现分割与去噪协同工作, 以迭代的形式去细化语义特征. 实验结果如表 2 所示.

根据表 2 可知, SCSDB 只迭代 1 次, *PixAcc*/*mIoU* 达到 94.1/74.5, 而 SCSDB 迭代 3 次, *mIoU* 反而下降 1.2. 当 SCSDB 迭代 2 次时, 精度达到最高, 为 94.9/75.7.



(a) Noise image (b) GT (c) SSGDM×1 (d) SSGDM×2 (e) SSGDM×3

图 6 不同阶段使用 SSGDM 方法的分割结果对比图

表 2 关于 SCSDB 的消融实验, 标准差σ为 25

方法	*mIoU*	*PixAcc*
SCSDB×1	74.5	94.1
SCSDB×2	**75.7**	**94.9**
SCSDB×3	73.3	94.2

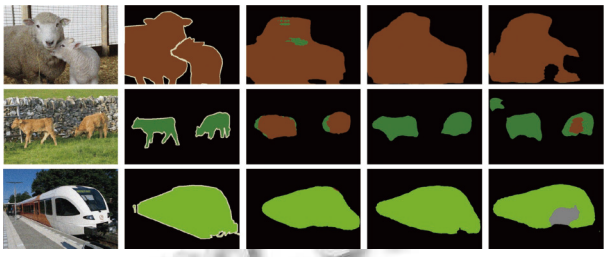
为了更加直观地展示 SCSDB 的迭代次数对分割结果的影响, 图 7 给出了分割结果图. 从前两张图像可以看出, 当 SCSDB 只迭代一次时, 噪声信息还没有被很好地抑制, 破坏了像素点的语义信息, 从而使得两个相近类别的错误识别. 当 SCSDB 迭代 3 次时, 在去噪过程中, 去除了过多的语义信息, 造成目标区域分割不完整. 从第 3 行图像可以看到, 当 SCSDB 迭代两次时, 模型能得到较好的去噪效果, 从而获取更精确的上下文信息, 使得分割区域边界清晰, 更趋近于 ground truth.

4.4 对比实验

为了验证模型的有效性和鲁棒性, 在 PASCAL VOC 2012 数据集和 Cityscapes 数据集上与目前先进的语义分割网络和带噪图像语义分割网络进行了一系列对比实验. 所有对比方法都采用相同的数据集进行

训练与测试.

首先在 PASCAL VOC 2012 数据集上与目前一些先进的语义分割网络进行定量比较, 所有方法控制相同的实验参数. 表 3 展示了提出的 NISNet 模型和其他语义分割方法的分割精度. 从表 3 可以看出, 在直接使用噪声图像进行训练时, 精度都取得了比较欠佳的效果, 高精度网络 DANet 在标准差 10、20、30 的数据集中, 其 *PixAcc/mIoU* 精度仅仅达到 91.3/75.9、90.5/74.4、89.8/72.3. 而具备 ASPP 的 DeepLabv3 效果更加不尽人意, 表现为 86.4/71.0、85.9/70.1、86.2/64.7, 这是由于噪声信息覆盖并破坏了图像的语义特征, 从而导致聚合上下文模块无法得到准确的语义依赖, 所以即使是高精度网络, 也不能取得较好的分割效果. 而且, 用这些网络对不同标准差噪声图像进行验证时, 当噪声标准差越强, 分割精度越低. 而提出的 NISNet 模型, 在标准差为 10、20 以及 30 的噪声图像中都取得了最优的性能.



(a) Noise image (b) GT (c) SCSDB×1 (d) SCSDB×2 (e) SCSDB×3

图 7 SCSDB 的不同迭代次数对应的分割结果图

表 3 PASCAL VOC 2012 数据集的语义分割结果

σ	FCN16s	PSPNet	DeepLabv3	PSANet	DANet	DenseASPP	NISNet (ours)
10	86.2/61.2	89.3/74.1	86.4/71.0	84.5/69.5	91.3/75.9	91.6/74.6	**94.0/78.1**
20	84.3/58.0	90.2/73.5	85.9/70.1	84.3/69.1	90.5/74.4	92.7/72.9	**93.3/77.4**
30	81.9/54.1	87.3/71.8	86.2/64.7	84.3/68.5	89.8/72.3	89.1/70.3	**92.1/74.9**

注: */*表示分割结果*PixAcc/mIoU*; 加粗字体为最优结果.

为了进一步保证实验的公平性, 为部分先进网络加入了去噪模块, 引入噪声损失, 与语义分割模块进行串联连接, 达到先去噪再分割的目的, 再与之进行实验对比. 表 4 给出了一系列方法的分割精度. 从表 4 可以看出, 经过去噪模块后, 分割精度得到了一定的提升. 相比于噪声图像, 使用去噪图像进行模型训练的 DANet, 在标准差为 10、20、30 的数据集中, 其 *mIoU* 提升了 0.7、1.0、1.0, 达到 76.6、75.4、73.3, 同时 DenseASPP 达到了 75.3、73.3、71.2 的 *mIoU*. 但是在所有噪声标准中, 提出的 NISNet 模型仍然领先于这些网络, *PixAcc/mIoU* 达到了 94.0/78.1、93.3/77.4、92.1/74.9. 相比于 DANet 模型, 提升了 2.7/1.5、1.6/2.0、2.9/1.6, 这是一个很可观的提升. 而且, 随着噪声的增强, NISNet 模型领先的精度更加明显. 同时, 还对协同分割去噪模型 DMS 方法进行了实验, 该方法是首个提出将去噪任务与语义分割任务进行结合的网络. 在不同标准的噪声图像中, DMS 的 *PixAcc/mIoU* 达到 92.6/76.9、93.7/76.2、89.5/74.1, 相比于 DANet, 精度提升了 1.3/0.3、2.0/0.8、0.3/0.8. 可以发现, 相较于先去噪后分割的模型架构, 协同分割去噪模型更具有优越性. 即便如此, 提出的 NISNet 模型的分割精度仍领先于 DMS, *mIoU* 在标准上 10、20、30 的数据集上, 提升了 1.2、1.2、0.8, 由此可以看出 NISNet 模型的先进性和鲁棒性.

为了直观比较表 4 方法的分割结果, 图 8 给出了在高斯噪声标准差为 20 的 PASCAL VOC 2012 数据集上提出的方法与其他方法的定性对比结果图. 从图 8(c)–图 8(e) 可以看到, 加完去噪模块后, 语义分割质量还是不尽人意, 目标范围内被错误识别, 以及目标的边界不精确和不平滑, 图 8(f) 显示了 DMS 的结果, 错误识别的问题得到了一定的改进, 但是在边界区域没有很好地被识别. 图 8(g) 显示了 NISNet 模型的结果, 可以明显看出, 无论是多类别还是小目标, 都能很好地进行分割. 而且, 从第 4 行和第 5 行图像可以看出, NISNet 模型在目标边界上达到了显著的提升. 因此, 提出的 NISNet 模型比其他去噪后再分割的串联网络更具有优势.

表 4 具有去噪模块的网络在 PASCAL VOC 2012 数据集的语义分割结果

σ	Deno.+DeepLabv3	Deno.+DenseASPP	Deno.+DANet	DMS	NISNet (ours)
10	84.9/71.7	94.2/75.3	91.3/76.6	92.6/76.9	**94.0/78.1**
20	86.2/70.3	90.0/73.3	91.7/75.4	93.7/76.2	**93.3/77.4**
30	83.9/67.8	88.4/71.2	89.2/73.3	89.5/74.1	**92.1/74.9**

注: */*表示分割结果*PixAcc/mIoU*; Deno.表示去噪; *+*表示去噪+分割的串联方法; 加粗字体为最优结果.

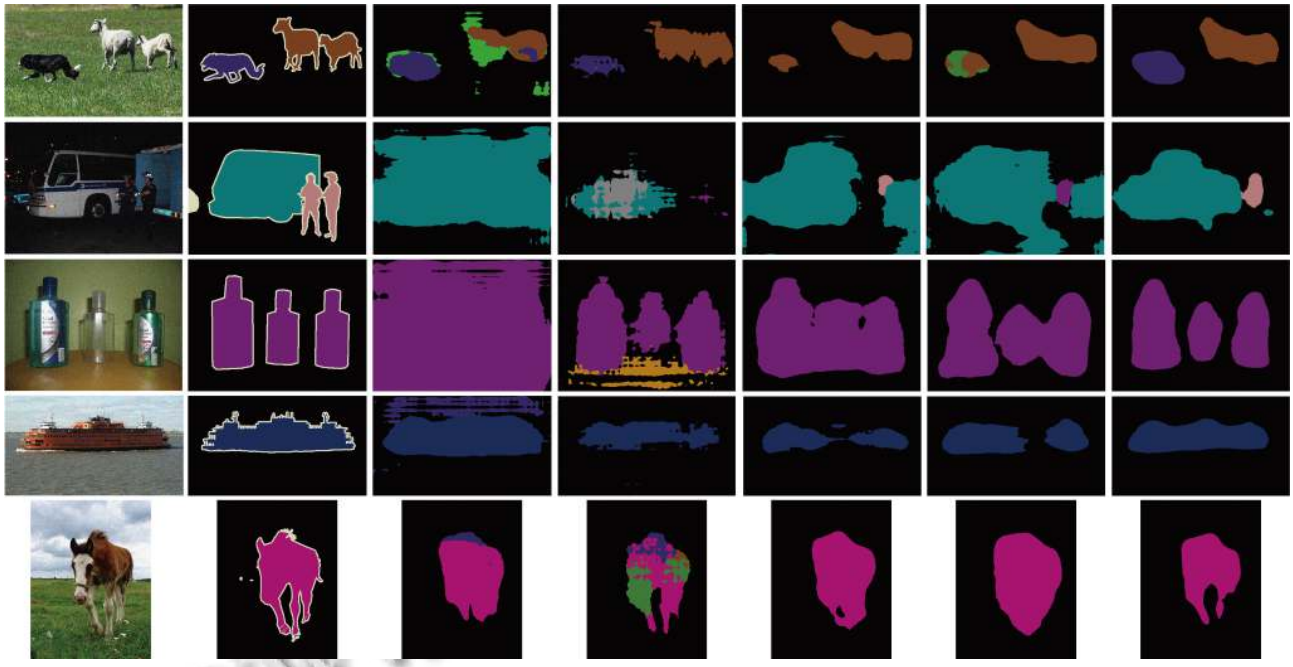


图 8 在噪声水平$\sigma = 20$的 PASCAL VOC 2012 数据集上的分割结果图

为了进一步验证模型的有效性, 在 Cityscapes 数据集上与其他语义分割网络进行了定量比较, 如表 5 所示. 在先去噪后分割的架构中, DANet 仍取得了较好的精度, *PixAcc/mIoU* 达到 92.3/67.9、91.7/66.5、88.3/64.8, 而协同分割去噪模型 DMS 达到 92.7/69.8、92.9/68.1、90.9/66.8 的精度, 可以发现, 协同任务模型比串联模型效果更佳. 而本研究提出的 NISNet 的精度比 DMS 更具有优越性, 相比于 DMS, *PixAcc/mIoU* 提升了 1.4/2.4、0.6/2.3、2.0/2.1. 综上所示, NISNet 不仅能够较好地识别目标边界区域, 并且具备很强的泛化能力.

表 5 具有去噪模块的网络在 Cityscapes 数据集的语义分割结果

σ	Deno.+DeepLabv3	Deno.+DenseASPP	Deno.+DANet	DMS	NISNet (ours)
10	92.2/68.6	92.2/67.7	92.3/67.9	92.7/69.8	**94.1/72.2**
20	91.3/66.1	91.4/64.4	91.7/66.5	92.9/68.1	**93.5/70.4**
30	88.6/62.7	88.2/61.9	88.3/64.8	90.9/66.8	**92.9/68.9**

注: */*表示分割结果*PixAcc/mIoU*; Deno.表示去噪; *+*表示去噪+分割的串联方法; 加粗字体为最优结果.

5 结论与展望

为了解决当前语义分割模型不能很好地捕捉噪声图像目标区域语义信息的难题, 提出了多尺度多阶段特征融合的带噪图像语义分割方法. 该方法应用了阶段性分割促进去噪模块 (SSGDM), 使得分割任务可促进去噪任务, 弥补在去噪环节中丢失语义信息的缺陷, 进而提高整体方法的分割精度. 在此基础上, 利用主干网络不同阶段提取的特征区别, 构成基于阶段性协同的分割去噪块 (SCSDB), 通过迭代融合多阶段语义特征, 使得网络更关注于目标区域的语义信息, 降低噪声对语义分割模型的干扰. 同时, 在公开数据集 PASCAL VOC 2012 和 Cityscapes 进行了验证, 对比其他语义分割方法, 提出的方法分别在不同噪声水平的噪声图像上取得了较好的分割精度, 进一步证明了方法的可行性.

然而, 如何对带有噪声的图像更好地进行分割仍然是一个具有挑战性的问题. 优化模型架构、减少参数量和提高分割速度等都是需要进一步探讨的问题. 在未来的工作中, 将进一步研究去噪和分割之间的协同性, 改善带噪图像语义分割算法性能.

参考文献

1 Zhao HS, Zhang Y, Liu S, *et al*. PSANet: Point-wise spatial attention network for scene parsing. Proceedings of the 15th European Conference on Computer Vision. Munich:

Springer, 2018. 270–286.
2 Yang MK, Yu K, Zhang C, *et al*. DenseASPP for semantic segmentation in street scenes. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 3684–3692.
3 Fu J, Liu J, Tian HJ, *et al*. Dual attention network for scene segmentation. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3146–3154.
4 Guo S, Yan ZF, Zhang K, *et al*. Toward convolutional blind denoising of real photographs. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 1712–1722.
5 Nagayama S, Muramatsu S, Yamada H, *et al*. Millimeter wave radar image denoising with complex nonseparable oversampled lapped transform. Proceedings of the 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. Kuala Lumpur: IEEE, 2017. 1824–1829.
6 Yang HY, Wang FY. Meteorological radar noise image semantic segmentation method based on deep convolutional neural network. Journal of Electronics & Information Technology, 2019, 41(10): 2373–2381. [doi: 10.11999/JEIT 190098.]
7 Gondara L. Medical image denoising using convolutional denoising autoencoders. Proceedings of the 2016 IEEE 16th International Conference on Data Mining Workshops. Barcelona: IEEE, 2016. 241–246.
8 Luthra A, Sulakhe H, Mittal T, *et al*. Eformer: Edge enhancement based transformer for medical image denoising. arXiv:2109.08044, 2021.
9 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: ACM, 2017. 6000–6010.
10 Buades A, Coll B, Morel JM. A non-local algorithm for image denoising. Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego: IEEE, 2005. 60–65.
11 Xu J, Zhang L, Zhang D, *et al*. Multi-channel weighted nuclear norm minimization for real color image denoising. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 1105–1113.
12 Gu SH, Zhang L, Zuo WM, *et al*. Weighted nuclear norm minimization with application to image denoising. Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 2862–2869.
13 Aharon M, Elad M, Bruckstein A. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Transactions on Signal Processing, 2006, 54(11): 4311–4322. [doi: 10.1109/TSP.2006.881199]
14 Mairal J, Bach F, Ponce J, *et al*. Non-local sparse models for image restoration. Proceedings of the 2009 IEEE 12th International Conference on Computer Vision. Kyoto: IEEE, 2009. 2272–2279.
15 Xu JJ, Osher S. Iterative regularization and nonlinear inverse scale space applied to wavelet-based denoising. IEEE Transactions on Image Processing, 2007, 16(2): 534–544. [doi: 10.1109/TIP.2006.888335]
16 Zhang K, Zuo WM, Chen YJ, *et al*. Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. IEEE Transactions on Image Processing, 2017, 26(7): 3142–3155. [doi: 10.1109/TIP.2017.2662206]
17 Zhang K, Zuo WM, Zhang L. FFDNet: Toward a fast and flexible solution for CNN-based image denoising. IEEE Transactions on Image Processing, 2018, 27(9): 4608–4622. [doi: 10.1109/TIP.2018.2839891]
18 Zhou YQ, Jiao JB, Huang HB, *et al*. When AWGN-based denoiser meets real noises. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 13074–13081. [doi: 10.1609/aaai.v34i07.7009]
19 Anwar S, Barnes N. Real image denoising with feature attention. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 3155–3164.
20 Chang M, Li Q, Feng HJ, *et al*. Spatial-adaptive network for single image denoising. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 171–187.
21 Jiang YF, Wronski B, Mildenhall B, *et al*. Fast and high-quality image denoising via malleable convolutions. arXiv:2201.00392, 2022.
22 Huang T, Li SJ, Jia X, *et al*. Neighbor2Neighbor: Self-supervised denoising from single noisy images. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 14781–14790.
23 Ren C, He XH, Wang CC, *et al*. Adaptive consistency prior based deep network for image denoising. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 8596–8606.
24 Huang JJ, Dragotti PL. WINNet: Wavelet-inspired invertible

network for image denoising. IEEE Transactions on Image Processing, 2022, 31: 4377–4392. [doi: 10.1109/TIP.2022.3184845]

25 Zamir SW, Arora A, Khan S, *et al*. Restormer: Efficient transformer for high-resolution image restoration. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 5718–5729.

26 Chen LC, Papandreou G, Kokkinos I, *et al*. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(4): 834–848.

27 Chen LC, Papandreou G, Schroff F, *et al*. Rethinking atrous convolution for semantic image segmentation. arXiv:1706.05587, 2017.

28 Chen LC, Zhu Y, Papandreou G, *et al*. Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 833–851.

29 Huang GH, Zhu JW, Li JJ, *et al*. Channel-attention U-Net: Channel attention mechanism for semantic segmentation of esophagus and esophageal cancer. IEEE Access, 2020, 8: 122798–122810. [doi: 10.1109/ACCESS.2020.3007719]

30 Ding XF, Shen CM, Che ZP, *et al*. SCARF: A semantic constrained attention refinement network for semantic segmentation. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 3002–3011.

31 Xie GS, Liu J, Xiong H, *et al*. Scale-aware graph neural network for few-shot semantic segmentation. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 5475–5484.

32 Cheng YT, Wei FY, Bao JM, *et al*. Dual path learning for domain adaptation of semantic segmentation. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 9062–9071.

33 Yu CQ, Gao CX, Wang JB, *et al*. BiSeNet V2: Bilateral network with guided aggregation for real-time semantic segmentation. International Journal of Computer Vision, 2021, 129(11): 3051 –3068. [doi: 10.1007/s11263-021-01515-2]

34 Song Q, Li J, Li CH, *et al*. Fully attentional network for semantic segmentation. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(2): 2280–2288. [doi: 10.1609/aaai.v36i2.20126]

35 Yang CG, Zhou HL, An ZL, *et al*. Cross-image relational knowledge distillation for semantic segmentation. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 12309–12318.

36 Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16×16 words: Transformers for image recognition at scale. arXiv:2010.11929, 2020.

37 Zheng SX, Lu JC, Zhao HS, *et al*. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 6881–6890.

38 Lee Y, Kim J, Willette J, *et al*. MPViT: Multi-path vision transformer for dense prediction. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 7277–7286.

39 Liu D, Wen BH, Liu XM, *et al*. When image denoising meets high-level vision tasks: A deep learning approach. arXiv:1706.04284, 2018.

40 Xu SX, Sun K, Liu D, *et al*. Synergy between semantic segmentation and image denoising via alternate boosting. arXiv:2102.12095, 2021.

41 Brempong EA, Kornblith S, Chen T, *et al*. Denoising pretraining for semantic segmentation. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 4174–4185.

42 Huang ZH, Zhang XC, Song YH, *et al*. FECC-Net: A novel feature enhancement and context capture network based on brain MRI images for lesion segmentation. Brain Sciences, 2022, 12(6): 765. [doi: 10.3390/brainsci12060765]

43 Everingham M, Van Gool L, Williams CKI, *et al*. The PASCAL visual object classes (VOC) challenge. International Journal of Computer Vision, 2010, 88(2): 303–338. [doi: 10.1007/s11263-009-0275-4]

44 Cordts M, Omran M, Ramos S, *et al*. The Cityscapes dataset for semantic urban scene understanding. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 3213–3223.

45 Garcia-Garcia A, Orts-Escolano S, Oprea S, *et al*. A review on deep learning techniques applied to semantic segmentation. arXiv:1704.06857, 2017.

(校对责编: 孙君艳)