基于时空双分支注意力的视频语义分割

黄丹丹[1)], 王一雯[1)], 陈广秋[1)], 胡奇[2)], 于斯宇[1)], 崔志瑜[3)]

[1)] (长春理工大学电子信息学院 长春 130022)
[2)] (长春理工大学人工智能学院 长春 130022)
[3)] (长春理工大学计算机科学技术学院 长春 130022)
(hdd@cust.edu.cn)

摘 要: 视频语义分割广泛应用在自动驾驶、交通管理、无人机驾驶等方面. 针对视频语义分割中由于分割结果不精确和分割时间过长导致不能达到实时性等问题, 提出基于时空双分支注意力机制的记忆网络分割算法. 该算法由记忆存储模块、特征整合模块和时空双分支注意力模块 3 部分构成. 其中, 记忆存储模块存储历史帧特征信息和序列时间信息; 特征整合模块引用大核分解卷积, 在不增加计算量的基础上扩大卷积感受野, 很好地捕获长距离依赖信息和上下文信息, 达到信道压缩和特征缩减的作用; 时空双分支注意力模块捕捉相邻帧中同一物体的特征信息并进行特征融合, 利用视频帧时序性提高分割精度. 在公开数据集 Cityscapes 和 CamVid 上的实验结果表明, 所提算法的平均交并比分别达到 76.92%和 73.68%, 计算速度达到 38 帧/s; 与其他视频语义分割算法相比, 该算法的分割精度和计算效率达到最优状态.

关键词: 视频语义分割; 记忆网络; 注意力机制; 特征融合
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2023-00046

Video Semantic Segmentation Based on Spatiotemporal Dual Branch Attention

Huang Dandan[1)], Wang Yiwen[1)], Chen Guangqiu[1)], Hu Qi[2)], Yu Siyu[1)], and Cui Zhiyu[3)]

[1)] (*Electronics and Information Engineering, Changchun University of Science and Technology, Changchun* 130022)
[2)] (*School of Artificial Intelligence, Changchun University of Science and Technology, Changchun* 130022)
[3)] (*School of Computer Science and Technology, Changchun University of Science and Technology, Changchun* 130022)

Abstract: Video semantic segmentation methods are widely used in autonomous driving, traffic management, drone driving and other fields. Aiming at the problems of inaccurate segmentation results and too long segmentation time to achieve real-time performance in video semantic segmentation, a memory network segmentation algorithm based on spatio-temporal dual-branch attention mechanism is proposed. The algorithm is mainly composed of memory storage module, feature integration module and spatio-temporal dual branch attention module. The Historical frame feature information and the sequence time information are stored in the memory storage module. The feature integration module uses large-core decomposition convolution to expand the convolution receptive field without increasing the amount of calculation, and captures long-distance dependent information and context information well to achieve channel compression and feature reduction. The spatio-temporal dual-branch attention module captures the feature information of the same object in adjacent frames and performs feature fusion, using the timing of video frames to improve segmentation accuracy. The

收稿日期: 2023-04-24; 修回日期: 2023-12-12. **基金项目**: 吉林省科技发展计划(20230201071GX); 国家自然科学基金重大仪器专项(62127813). **黄丹丹**(1984—), 女, 博士, 讲师, 硕士生导师, 主要研究方向为图像处理、计算机视觉、机器学习等; **王一雯**(1999—), 女, 硕士研究生, 主要研究方向为语义分割、视频语义分割; **陈广秋**(1977—), 男, 博士, 副教授, 硕士生导师, 主要研究方向为图像处理、机器视觉等; **胡奇**(1981—), 男, 博士, 副教授, 硕士生导师, 主要研究方向为视觉信息处理、智能感知等; **于斯宇**(1999—), 女, 硕士研究生, 主要研究方向为目标跟踪; **崔志瑜**(1999—), 男, 硕士研究生, 主要研究方向为图像分类.

algorithm achieves 76.92% and 73.68% mIoU on the public datasets Cityscapes and CamVid, and the calculation speed reaches 38 frames/s. The experiment proves that this method has reached the optimal state in segmentation accuracy and calculation efficiency compared with other video semantic segmentation algorithms.

Key words: video semantic segmentation; memory network; attention mechanism; feature fusion

语义分割是计算机视觉领域的基础任务之一[1], 在自动驾驶、医学影像、VR 技术等多个领域得到广泛应用. 作为像素级别[2]的分割任务, 图像语义分割的目标是为图像中的每个像素点进行分类[3]. 随着视频数据的爆发式增长与处理手段的飞速发展, 图像语义分割已经不能满足日常生活的需要. 相应地, 视频语义分割是计算机视觉领域未来的发展趋势, 该项任务需要对视频中每一帧的每个像素都进行标签分类, 因此更具有挑战性.

视频语义分割现存 2 大问题: 一是如何充分地利用视频帧之间的时序性信息提高分割精度; 二是连续帧分割带来的计算量过大, 在实际生产生活中无法达到实时分割的要求. 针对第 1 个问题, 最直接的解决方案是将语义分割模型应用在视频的每一帧, 但是简单的逐帧分割并没有充分挖掘视频帧之间的时间关系和帧间相似性等信息, 因此在分割精度上仍有很大提升空间. 例如, 直接应用 DeepLabV3 算法[4]分割视频帧时, 得到的分割结果精度过低且引入了大量计算. 对于第 2 个问题, 则常利用具有记忆功能或传播功能的网络来减少计算量. Fayyaz 等[5]提出的 STFCN 将长短期记忆(long short-term memory, LSTM)模块[6]与 FCN[7]相结合, 构成端到端的时空卷积网络结构, 目的是利用 LSTM 网络存储时空序列记忆信息; 2017 年, NetWarp[8]模型采用光流法将前一帧的特征搬移到当前帧, 在一定程度上起到特征增强的作用; 同年, Zhu 等[9]提出视频识别深度特征流(deep feature flow for video recognition, DFF)模型, 采用双分支网络选取关键帧和非关键帧, 对关键帧进行语义分割得到特征图, 再利用光流法将关键帧特征传播到非关键帧中, 达到视频帧特征复用, 并减小计算量和计算时间的目的. 但上述方法中, 简单的特征提取模型循环神经网络(recurrent neural network RNN)[10]能够存储的特征信息有限, 限制了算法的使用范围; 而光流法在传播特征时遇到遮挡、光线变化、视频帧失真时会产生较大的匹配误差, 并且由于分割时间和特征传播时间的不同而造成延迟, 引起分割错误. 针对上述问题, 在视频问答和视频对象分割中, 通常采用记忆网络增强模型的分割能力. Gao 等[11]利用运动-外观协同记忆网络(motion-appearance co-memory networks, MAC)使用情景记忆, 通过将问题与当前记忆条件下的视频特征相互作用进行多个周期的推理; Oh 等[12]提出了一个时空记忆注意机制模型去捕获时间信息: 这 2 种方法通过引入记忆网络, 有效地提升了视频分割的精度.

受以上工作的启发, 本文提出一种基于时空双分支注意力机制的记忆网络分割算法. 首先构造记忆网络并搭建记忆存储器, 充分地提取并保存帧间相关信息; 然后通过时空双分支注意力机制计算当前帧和记忆序列帧的相关性, 检索特征信息, 增强当前帧特征; 最后针对记忆存储器计算量大、细节信息丢失等问题, 设计了特征整合模块, 采用大核分解卷积(large kernel decomposition convolution, LDC)实现信道缩减和特征压缩的目的. 通过上述 3 个模块, 实现了视频语义分割性能和效率的提升.

1 相关工作

1.1 视频语义分割

现有的视频语义分割方法大致分为 2 类: 一是根据连续帧具有相似性的特点, 重用前一帧的特征信息, 利用光流跨帧传播关键帧的特征达到特征复用的目的; 二是利用帧间信息的时序性, 使帧与帧之间同一语义分割对象的特征可以关联匹配. 这 2 类方法中, 利用光流传播虽然能够减小计算量, 但是光流传播特征计算的速度和关键帧语义分割提取特征的速度差别很大, 会造成计算的延迟; 此外, 光流估计很难精准传播, 遇到遮挡、光线变化等问题会影响分割的精确度. 如何充分利用帧间的时序性, 使分割精度和计算速度达到最优的平衡, 是视频语义分割研究的重点.

1.2 记忆网络

现代计算机的优点是可以对信息进行存储, 但是大多深度学习模型都缺少存储查询的结构. 对于 RNN, LSTM 这样的神经网络, 虽然原则上可

以实现记忆存储，但由于隐藏状态和权重编码包含的记忆空间太小，因此不能保留足够的信息．针对这一问题，记忆网络需要记忆组件保存记忆场景环境信息，达到长时的记忆，以便查询和提取特征信息．STM 首次将记忆网络应用在视频对象分割中，其将记忆网络和空间注意力相结合，将当前帧的信息与之前帧的所有信息在空间与时序维度上进行特征匹配，得到超过之前方法的性能表现．如图 1 所示，因为连续的视频帧一般具有相同的分割对象，可视为具有高度相关性，所以视频语义分割需要充分地提取和利用视频帧间的时序性以及相似性．由于记忆网络可以存储连续帧的信息，因此本文引入记忆网络存储历史帧信息，以提高分割性能．

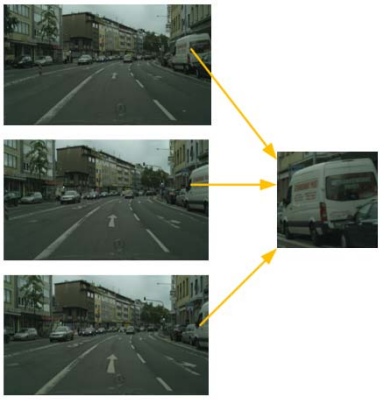


图 1　记忆帧序列中的相同对象

1.3　注意力机制

注意力机制最早应用于文本翻译．在认知学中，人们会对接收到的信息选择性地关注．之后，注意力机制被引入到计算机视觉领域，其可以看作一个自适应选择的过程，根据输入特征选择鉴别相似特征并自动忽略噪声响应．注意力机制的关键步骤是生成注意力相似特征图，注意力模块通过所有位置的特征加权及各个位置类似特征的选择性聚合将彼此相关化．通道注意力和空间注意力是 2 种常用的注意力机制，其中，通道注意力通过整合所有通道图之间的相关性，选择地增强相互依赖的通道特性．SENet[13]利用通道注意力显式地建模特征通道之间的相互依赖关系，自动获取每个通道的重要程度．CBAM[14]表示卷积模块的注意力机制模块，是一种结合了空间和通道的注意力机制模块．用于场景分割的双注意力网络[15](dual attention network for scene segmentation, DANet)自适应地集成局部特征和全局依赖，在传统的 FCN 之上附加 2 种类型的注意力模块，分别模拟空间和通道维度中的语义相互依赖性．本文算法借鉴 DANet，并在其基础上引入时间维度，构成时间-空间注意力(temporal and spatial attention, TSA)模块和时间-通道注意力(temporal and chanel attention, TCA)模块．首先将记忆存储模块中的历史帧信息自适应地进行特征整合，得到空间和通道的注意力图；然后将当前帧与记忆序列注意力图相乘，进行特征融合和信息增强，通过重用视频帧间的相似性信息提高分割精度．

2　本文算法

2.1　网络模型整体结构

本文算法的结构如图 2 所示．首先输入含有多帧图像的视频序列，其中，将带有标签信息的视频帧称为查询帧，将不带标签信息的视频帧作为记忆序列送入记忆存储器．记忆序列和查询帧通过共享主干网络提取特征信息，输出高维度特征；然后将得到的特征送入特征整合模块进行信道和特征缩减，其中，特征整合模块采用 LDC，在增大感受野的同时降低计算量．将记忆序列从特征整合模块输出的 2 个不同的记忆特征记为 $\boldsymbol{M}_1$ 和 $\boldsymbol{M}_2$，查询帧输出的查询特征记为 $\boldsymbol{D}_1$ 和 $\boldsymbol{D}_2$．两者共同送

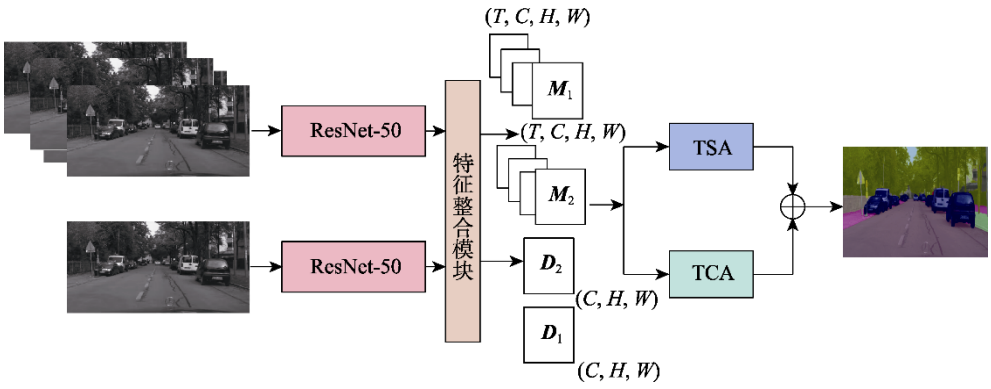


图 2　本文算法结构

入时空双分支注意力模块提取长时间上下文信息，生成空间和通道特征图；再将当前帧特征与注意力图进行特征聚合，以增强特征表示，最后送入分割头获得分割结果.

2.2 记忆存储模块

记忆存储模块的功能是利用连续的视频帧组成记忆序列，并将其送入共享的骨干网络中提取特征. 该模块从连续的视频片段中生成高度汇总的记忆状态，以便查询帧使用相同物体的相似性信息进行特征校正. 输入连续特征图为三通道彩色图像，通过骨干网络得到 T 帧的记忆序列特征；再按照时间维度进行串联后获得四维特征 $\boldsymbol{M}(T,C,H,W)$，其中，T 表示选择历史帧的数量，C 表示特征的通道数，H 和 W 分别表示特征的宽和高，$\boldsymbol{M}$ 特征包含时间信息维度；同时，将查询帧送入共享的骨干网络提取特征，得到查询特征 $\boldsymbol{D}(G,H,W)$. 本文采用 ResNet-50 作为共享主干网络，由于输出维度过高会导致计算量过大，因此将记忆存储模块输出特征 $\boldsymbol{M}(T,G,H,W)$ 和当前帧输出特征 $\boldsymbol{D}(G,H,W)$ 送入特征整合模块，进行信道缩减和特征压缩.

2.3 特征整合模块

特征整合模块的作用是对骨干网络输出的高维特征信息进行信道压缩和特征缩减，减小计算量并节约计算时间. 本文采用 LDC 对特征进行处理，分为深度卷积、深度可分离卷积和 1×1 通道卷积 3 部分. 由视觉注意网络[16](visual attention network, VAN)可知，卷积核大小为 $K\times K$ 的卷积操作，可以分解成卷积核大小为 K/d 的深度可分离卷积，其中，d 表示可分离系数. 记忆序列和查询帧通过共享主干网络输出特征 $\boldsymbol{M}_1,\boldsymbol{M}_2,\boldsymbol{D}_1,\boldsymbol{D}_2$ 后进入特征整合模块，其结构如图 3 所示. 首先经过卷积核大小为 7×7、步长为 4、填充系数为 3 的卷积操作；然后经过 1×1 卷积调整通道数后，送入 LDC 部分.

LDC 模块中，首先进行卷积核 k 大小为 5×5、填充系数 p =2 的深度卷积，然后经过卷积核 k 大小为 7×7、步长 s 为 1、填充系数 p =9，可分离系数 d =3 的深度可分离卷积：再经过 1×1 通道卷积，得到输出矩阵后进行矩阵变换，最后和输入矩阵相乘，自适应地调整相关性特征权重. 上述计算过程的公式为

$$\boldsymbol{o}_1=\mathrm{Conv}\big(L(H(\boldsymbol{i}))\big),$$

$$\boldsymbol{o}=\boldsymbol{i}\otimes\boldsymbol{o}_1.$$

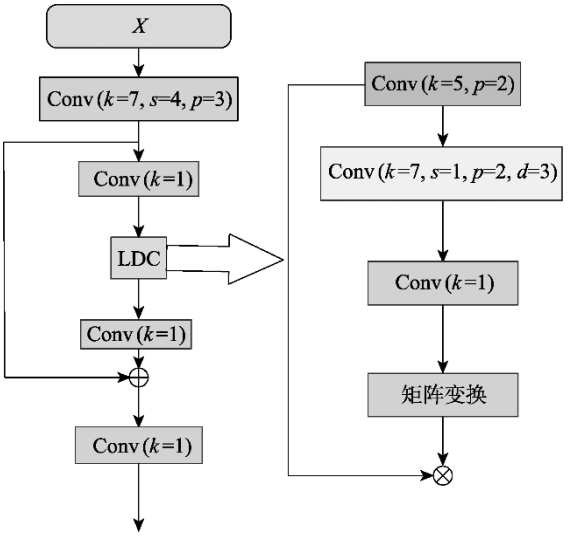


图 3 LDC 结构流程图

其中，$\boldsymbol{i}$ 表示输入到 LDC 的特征矩阵；L 表示深度卷积；H 表示深度可分离卷积；Conv 表示卷积计算；⊗表示矩阵相乘；$\boldsymbol{o}$ 表示 LDC 模块的输出. LDC 可以获得特征的长程上下文信息和联系特征之间的依赖性. 最后对特征图进行上采样，恢复通道数，并输出特征.

2.4 时空双分支注意力模块

通过对卷积网络输出的特征重新分配权重，注意力机制能够增强或抑制部分权重比例，提高分割精度. 在空间注意力中，所有位置特征通过加权求和进行聚合，权重由相似性信息决定，相似特征的位置都能够增强特征而与其在空间上的距离无关，因此可获取全局信息. 通道注意力则通过自适应地获得特征图通道的重要程度，在特征通道之间赋予权重来增强或抑制比重，提高特征的表达能力和准确性，使不同语义的响应互相关联.

由于视频语义分割任务中包含大量的帧间信息，相对于空间信息，这些帧间信息可视为时间信息，有助于提高分割准确性，因此本文将空间注意力、通道注意力加以改进，在原有结构的基础上增加时间维度信息，构成 TSA 模块和 TCA 模块. 这 2 个模块可以对视频序列不同帧的相同物体提取长程上下文信息和物体边缘的分割细节，它们的结构分别如图 4 和图 5 所示. 如前文所述，记忆网络将记忆的信息分别嵌入到 $\boldsymbol{M}$ 和 $\boldsymbol{D}$ 中. 通过特征整合模块输出 $\boldsymbol{M}_1,\boldsymbol{M}_2,\boldsymbol{D}_1,\boldsymbol{D}_2$，4 组特征送至注意力模块中.

从图 4 中可知，TSA 模块首先将 $\boldsymbol{M}_2$ 进行矩阵变换后矩阵大小为 $(C,T\times H\times W)$，将 $\boldsymbol{D}_2$ 进行矩阵

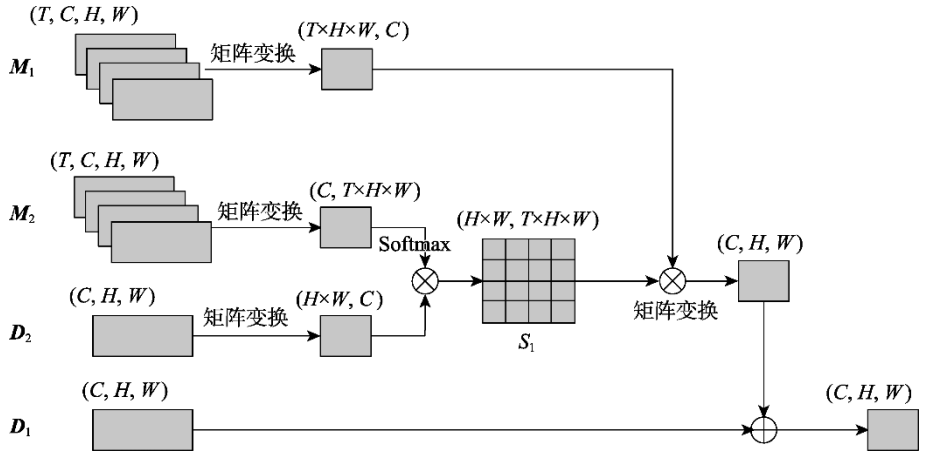


图 4　TSA 结构

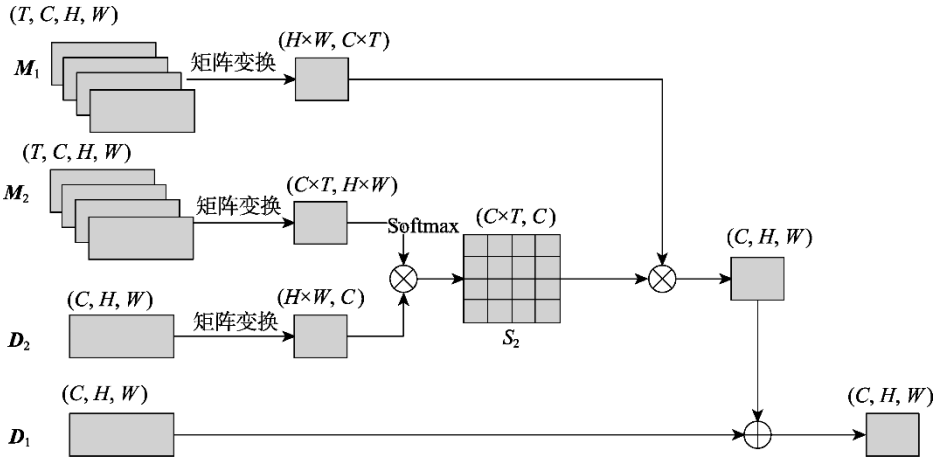


图 5　TCA 结构

变换后矩阵大小为 $(H\times W, C)$，矩阵变换是为了后续生成不同帧、相同对象之间的注意力图；然后将 $\boldsymbol{M}_2$ 和 $\boldsymbol{D}_2$ 进行矩阵相乘，应用一个 Softmax 层得到时间空间记忆注意力图 $\boldsymbol{S}_1$，其中，$\boldsymbol{S}_1 \in \boldsymbol{M}_2 \times \boldsymbol{D}_2$，矩阵大小为 $(H\times W, T\times H\times W)$，记忆序列特征第 j 个位置对查询帧第 i 个位置的影响的计算公式为

$$\boldsymbol{S}_{ij} = \frac{\exp(\boldsymbol{M}_2^i \cdot \boldsymbol{D}_2^j)}{\sum_j^N \exp(\boldsymbol{M}_2^i \cdot \boldsymbol{D}_2^j)}.$$

其中，$\boldsymbol{S}_{ij}$ 越大，说明二者的关系越大. 将计算全部位置的 $\boldsymbol{S}_{ij}$ 集成为 $\boldsymbol{S}_1$ 注意力图；然后将 $\boldsymbol{M}_1$ 进行矩阵变换为矩阵大小为 $(T\times H\times W, C)$；然后将 $\boldsymbol{M}_1$ 和 $\boldsymbol{S}_1$ 矩阵相乘，尺寸为 (C, H, W)，最后与 $\boldsymbol{D}_1$ 进行矩阵拼接，尺寸为 $(2\times C, H, W)$，即为 TCA 分支的输出结果. 从图 5 中可知，TCA 将矩阵 $\boldsymbol{M}_2$ 进行矩阵变换操作，矩阵大小变为 $(C\times T, H\times W)$，再将 $\boldsymbol{D}_2$ 重塑为 $(H\times W, C)$；最后将 $\boldsymbol{M}_2$ 和 $\boldsymbol{D}_2$ 进行矩阵相乘，得到的特征向量经过 Softmax 层得到时间空间记忆注意力图 $\boldsymbol{S}_2 \in \boldsymbol{M}_2 \times \boldsymbol{D}_2$，矩阵大小为 $(C\times T, C)$，其计算方法与 $\boldsymbol{S}_1$ 的计算方法相同. 得到特征图 $\boldsymbol{S}_2$ 后，先将 $\boldsymbol{M}_1$ 进行矩阵变换，变换后大小为 $(C\times T, H\times W)$；后与 $\boldsymbol{S}_2$ 进行矩阵相乘后大小为 (C, H, W)；最后与 $\boldsymbol{D}_1$ 进行矩阵变换，输出矩阵大小为 $(2\times C, H, W)$，即为时间-通道注意力分支的输出结果. TCA 改变了通道维度的权重比例，能够增强或抑制特征，使分割结果更加准确.

2.5　分割头

本文在训练阶段添加辅助分割模块以计算辅助分割损失. 辅助分割模块采用 FCN，将骨干网络底层 stage3 输出特征 $\boldsymbol{X}$ 送入 FCN，其中，$\boldsymbol{X}$ 的通道数为 1 024. FCN 先经过卷积核大小为 3×3，扩展系数为 1 的卷积操作，输出通道数为 256；再经过 2 次卷积核大小为 3×3，扩展系数为 11 的卷积操作进行通道拼接；最后送入分割头进行分类计算辅助损失函数.

本文还运用注意力网络得到视频帧的长程上下文信息，将当前帧和上下文信息进行特征融合，采用特征拼接操作后得到融合的信息，最终送入分割头得到视频帧的掩码图，并计算损失函数，以便进行反向传播.

3 实验及结果分析

3.1 数据集和评价指标

为了评估本文算法的分割效果，在2个基准数据集 Cityscapes[17]和 CamVid[18]上进行实验. 其中，Cityscapes 数据集包含 5 000 张精细标注注释的图片，分为训练集 2 970 个片段，验证集 500 个片段，测试集 1 275 个片段，每个片段包含 30 帧图片，仅在第 20 帧注释了 19 个类用于语义分割；CamVid 数据集包含 4 个视频帧，注释标签一共用 11 个类别，训练集共有 467 个片段，验证集有 100 个片段，测试集有 233 个片段，每个片段包含 30 帧图片，带有标签注释的为片段第 1 帧.

实验中，采用类别平均像素准确率(mean pixel accuracy, mPA)和平均交并比(mean intersection over union, mIoU)作为评价指标，它们的计算公式分别为

$$\mathrm{mIoU}=\frac{1}{k+1}\sum_{i=0}^{k}\frac{\mathrm{TP}}{\mathrm{FN}+\mathrm{FP}+\mathrm{TP}},$$

$$\mathrm{mPA}=\frac{1}{i}\left(\sum_{i=0}^{i}\frac{\mathrm{TP+TN}}{\mathrm{FN}+\mathrm{FP}+\mathrm{TP+TN}}\right).$$

其中，k 为数据集类别数；i 为类别；真正(truth positive, TP)表示预测正确，预测结果为正类；假正(false positive, FP)表示预测错误，预测结果为正类；假负(false negative, FN)表示预测错误，预测结果为负类；真负(truth negative, TN)表示预测正确，预测结果为负类. 通过 mIoU 能够比较客观地评价语义分割算法的分割性能.

3.2 实验参数设置

本文实验环境为 Ubuntu18.04, RTX 3090Ti 显卡，Python 环境为 Python3.6, Pytorch1.7.0, CUDA 版本为 11.0. 网络训练中，Cityscapes 数据集输入尺寸调整为 768×768 像素，CamVid 数据集输入尺寸为 640×640 像素，并采用随机翻转，裁剪、图像失真、图片标准化等数据增强操作. 选择交叉熵损失函数作为网络的损失函数，公式为

$$L=-\frac{1}{N}\sum_{i}\sum_{i}^{c}y_{ic}\ln(\rho).$$

其中，ρ 表示概率分布；N 表示类别数量；y_{ic} 表示符号函数(0 或者 1)，如果样本 i 的真实类别等于 c，则 y_{ic}=1，否则 y_{ic}=0. 优化器选择随机梯度下降法进行网络参数的更新. 设置初始学习率为 0.01，动量为 0.9，权重衰减为 0.000 1，并采用 poly 学习率策略加快训练进程. 在测试阶段，采用滑动窗口策略生成视频片段，并且在骨干网络的底层(如 ResNet 的 stage3 输出)添加辅助分割损失. 如果没有特别说明，本文所有实验均采用上述设置.

3.3 消融实验

为了验证本文算法的有效性，在 CamVid 和 Cityscapes 数据集上进行消融实验.

3.3.1 LDC 模块

记忆网络特征整合模块是进行信道缩减和特征压缩的关键步骤，如果单独使用 3×3 卷积会使特征丢失全局信息和边缘信息，导致分割对象边界分割不清晰和错误分割的问题. 考虑到计算成本，本文引入 LDC 进行矩阵压缩. 在 2 个数据集上，不同大小的 LDC 模块性能对比如表 1 所示. 表 1 中的 K 和 D 分别代表卷积核大小和可分离卷积的分离系数. 可以看出，LDC 下，不同大小的卷积核和不同的可分离系数也会对计算精度和计算量造成影响，当 K=14 且 D=3 以及 K=21 且 D=3 时，mIoU 最高；当 K 和 D 过大时分割精度反而降低，可能是当卷积分离系数过大使特征丢失全局信息，导致分割精度的下降，而且卷积核过大计算量也相应变大，失去了特征整合模块压缩特征、减小计算量的优势.

表 1 2 个数据集不同大小的 LDC 模块性能对比

模型	K	D	CamVid			Cityscapes		
			mIoU/%	mPA/%	分割速度/(帧·s^{-1})	mIoU/%	mPA/%	分割速度/(帧·s^{-1})
CNN	3	0	70.52	78.75	26	74.27	80.04	59
LDC	7	2	71.68	80.34	58	75.93	83.48	56
	14	3	73.68	81.42	38	76.92	85.21	46
	21	3	72.54	81.59	25	75.89	84.90	37
	28	4	71.60	79.90	12	75.04	83.69	33

3.3.2　双分支注意力模块

在 2 个数据集上，对 TSA 和 TCA 这 2 个模块的有效性进行实验，结果如表 2 所示. 其中，选取 ResNet50 作为骨干网络，特征整合模块只采用普通的 3×3 卷积. 可以看出，采用 TSA 和 TCA 并行连接的模块的分割精度最好.

表 2　2 个数据集上 TSA 和 TCA 模块有效性验证　%

模块	CamVid		Cityscapes	
	mIoU	mPA	mIoU	mPA
TSA	69.84	76.79	73.65	81.28
TCA	69.03	76.52	72.09	80.54
TSA+TCA	70.20	77.32	74.89	83.21
TCA+TSA	68.17	75.32	73.90	82.84
TCA/TSA	**73.68**	**83.21**	**76.92**	**85.21**

注. 粗体表示最优值.

实验结果证明，引入 TSA 和 TCA 模块，对于提取记忆序列不同帧中相同物体的特征，以及同一帧的同一对象特征的有效性. 本文提出的双分支注意力机制可以很好地提取长程上下文信息，对查询帧的特征进行校正和增强，从而提高分割精度.

3.3.3　记忆存储模块

作为本文算法的主体框架，记忆存储模块存储的记忆帧的数量(简称为记忆帧数)对分割结果尤为重要，本文通过设置不同的数值选取最佳的结果. 当设置记忆帧数为 4 和 5 时，分割精度最高，效果最好，但是，记忆帧数为 5 时的分割速度远小于记忆帧数为 4 时. 因此，本文将记忆帧数设置为 4. 确定了记忆帧数后，通过设置不同的通道数 P 进行实验，结果如表 3 所示. 可以看出，当 P =128 时本文算法的分割性能最好，可更有效地进行特征缩减并减少计算量.

表 3　2 个数据集上不同记忆帧数性能对比

记忆序列帧数	P	CamVid			Cityscapes		
		mIoU/%	mPA/%	分割速度/(帧·s^{-1})	mIoU/%	mPA/%	分割速度/(帧·s^{-1})
2	128	72.11	81.02	65	74.27	82.58	58
3	128	72.89	81.95	54	75.17	84.05	48
4	128	73.68	83.21	38	76.92	85.21	46
5	128	73.57	83.16	32	76.43	85.38	32
4	64	72.90	81.98	54	76.05	84.78	52
4	128	73.68	83.21	38	76.92	85.21	46
4	256	73.67	82.5	29	76.86	85.18	32

3.3.4　输出融合模块

本文选取 4 种输出融合模块的方法对 TSA 和 TCA 并行连接后输出的特征进行融合，其余地方采用 LDC 和双分支注意力模块，表 4 所示为不同融合方法的准确度对比. 可以看出，特征相加的融合方法分割精度最高；池化和卷积操作都会导致特征信息丢失；特征相加可以关联更多特征之间的信息，保存更完整的细节信息，所以采用特征相加的融合方法获得的分割精度最高，效果最好.

表 4　不同模块融合方法性能对比　%

模块融合方法	CamVid		Cityscapes	
	mIoU	mPA	mIoU	mPA
3×3 卷积	73.02	81.45	76.25	85.01
平均池化	72.24	79.63	75.62	84.93
最大池化	72.19	79.30	75.36	84.59
特征相加	**73.68**	**83.21**	**76.92**	**85.21**

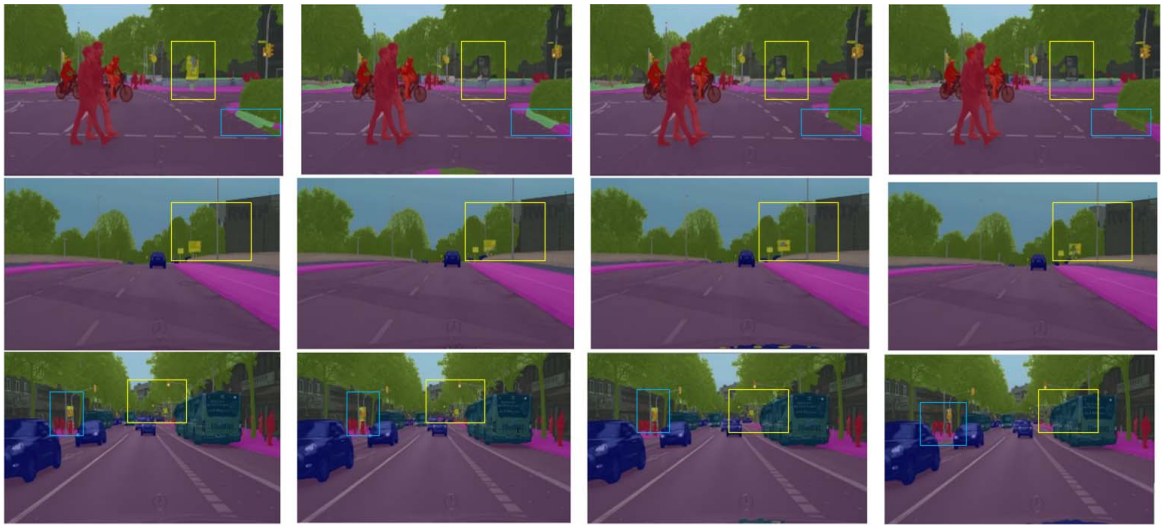
注. 粗体表示最优值.

3.3.5　模块可视化

LDC，TSA 和 TCA 模块的可视化结果如图 6 所示. 可以看出，单独添加不同模块时，分割结果大体上都十分精准，能够准确地识别树木、道路、行人、建筑和自行车等；只应用 TSA 或者 TCA 模块，对广告牌交通灯等小物体分割不准确；只应用 LDC 模块，分割效果有所提升；应用全部模块，分割效果提升明显，针对小物体和远处物体都能大体分割清晰. 这是因为卷积操作会丢失分割对象边缘细节信息，本文应用 LDC 可有效地扩大感受野，保留边缘信息；双分支注意力模块整合记忆帧特征，完善查询帧的分割细节部分，因此组合后分割效果更好.

3.4　对比实验

为了更好地体现本文算法的综合性能，将其与经典的视频语义分割算法[9,19-30]在 Cityscapes 和 CamVid 数据集上进行分割精度的对比. 结果如表 5



a. 完整模型 b. 只包含 LDC c. 只包含 TCA d. 只包含 TSA

图 6 3 个模块的有效性验证

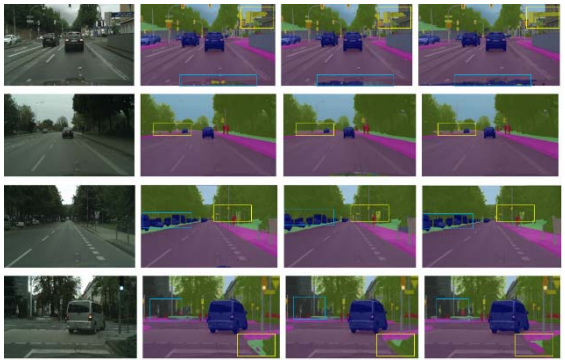
表 5 2 个数据集上不同算法 mIoU 对比 %

算法	Cityscapes	CamVid
Accel-50[24]	72.70	67.70
TDNet[25]	74.91	71.30
TMANet[26]	73.64	70.20
DVSNet[27]	73.50	
DFF[9]	69.20	66.00
CFFM[28]	75.10	
AuxAdapt[29]	76.60	65.80
CReff[30]	70.72	71.60
GRFP[23]	73.60	66.10
AGNet[19]		73.60
LERNet[20]	69.50	
Noisy-LSTM[21]	73.00	
LVSNet[22]	75.30	
本文	**76.92**	**73.68**

注. 粗体表示最优值.

所示. 由于 AGNet[19], LERNet[20], Noisy-LSTM[21], LVSNet[22]和 GRFP[23]算法无公开代码, 因此直接引用该算法中的原始数据, 表 5 中的其余数据均是在同一平台进行训练和测试的. 实验结果表明, 本文引入的记忆存储模块可以有效地提高语义分割精度, 设计的 TSA 和 TCA 模块能够有效地对特征进行融合和校正, 使不同帧中相同对象的特征通过记忆网络实现有效传递; 引入的 LDC 模块可以获得更大的感受野, 使分割图的细节边缘部分更加清晰准确, 具有更高的计算效率.

从表 5 可以看出, TDNet 和 TMANet 是经典算法中分割精度最高的 2 种算法, 且与本文算法一样, 都是利用视频帧间时序性去提高算法分割精度, 所以选取这 2 种算法的分割效果图与本文算法的效果图进行对比, 3 种算法分割结果的可视化如图 7 所示. 从图 7 可以看出, 整体上看, 3 种算法的分割结果都基本精准, 但是针对小物体和分割细节处, 本文算法的分割效果更好; 对于第 1 行图像黄框中的广告标识牌, 本文算法能够分割出完整的个体, 而其余 2 种算法只能够分割出一部分; 对于第 3 行的树木和车辆穿插处, 本文算法可以具体分割出树的形状和车的边缘, 相比之下, 其他 2 种算法只能整块识别成树木; 对于第 4 行中马路边和草丛连接处, 本文算法也能精准分割, 比其他 2 种算法分割更准确. 实验结果表明, 本文算法针对物体边缘的分割效果更好; 结合表 5 可知, 在同等条件下, 该算法的精度高于其他算法.



a. 原图 b. TDNet[25] c. TMANet[26] d. 本文

图 7 3 种算法在 Cityscapes 数据集上的分割结果对比

4 结 语

针对视频语义分割中存在的问题, 本文提出

一种基于时空双分支注意力机制的记忆网络分割算法，设计了记忆存储模块、特征整合模块和时空双分支注意力模块实现视频语义分割．实验结果表明，时空记忆存储模块能够存储足够的历史帧信息用于挖掘帧间关联；时空双分支注意力模块能够充分地利用历史视频帧信息，增强当前帧的特征表示，提高分割精度；特征整合模块引入的 LDC 可以扩大感受野，并且在有效地捕捉长时间上下文信息的同时不增加计算量；本文算法在精度和计算时间方面达到平衡，与其他视频语义分割算法相比取得了最好的效果．未来将致力于通过无监督和弱监督的方式减少模型对数据的依赖性，并且关注在提高计算速度的同时不降低精度，为实际应用中的实时性问题提供解决方案．

参考文献(References):

[1] Chen Shu, Xu Lei, Zou Beiji, *et al*. Semantic relocation parallel network for semantic segmentation[J]. Journal of Computer-Aided Design & Computer Graphics, 2022, 34(3): 373-381(in Chinese)
(陈姝, 徐蕾, 邹北骥, 等. 基于语义重定位的语义分割并行网络[J]. 计算机辅助设计与图形学学报, 2022, 34(3): 373-381)

[2] Qin Feiwei, Shen Xile, Peng Yong, *et al*. A real-time semantic segmentation approach for autonomous driving scenes[J]. Journal of Computer-Aided Design & Computer Graphics, 2021, 33(7): 1026-1037(in Chinese)
(秦飞巍, 沈希乐, 彭勇, 等. 无人驾驶中的场景实时语义分割方法[J]. 计算机辅助设计与图形学学报, 2021, 33(7): 1026-1037)

[3] Yan Chengliang, Chen Guangzhu, Yi Jia, *et al*. Lightweight semantic segmentation of intelligent workshop scene objects combining multi-scale and attention mechanisms[J]. Journal of Computer-Aided Design & Computer Graphics, 2022, 34(10): 1626-1636(in Chinese)
(严成良, 陈光柱, 易佳, 等. 融合多尺度与注意力机制的智能车间场景目标轻量级语义分割[J]. 计算机辅助设计与图形学学报, 2022, 34(10): 1626-1636)

[4] Chen L C, Papandreou G, Schroff F, *et al*. Rethinking atrous convolution for semantic image segmentation[OL]. [2023-04-24]. https://arxiv.org/abs/1706.05587

[5] Fayyaz M, Saffar M H, Sabokrou M, *et al*. STFCN: spatio-temporal FCN for semantic video segmentation[OL]. [2023-04-24]. https://arxiv.org/abs/1608.05971

[6] Yu Y, Si X S, Hu C H, *et al*. A review of recurrent neural networks: LSTM cells and network architectures[J]. Neural Computation, 2019, 31(7): 1235-1270

[7] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 3431-3440

[8] Gadde R, Jampani V, Gehler P V. Semantic video CNNs through representation warping[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 4463-4472

[9] Zhu X Z, Xiong Y W, Dai J F, *et al*. Deep feature flow for video recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 4141-4150

[10] Cabessa J, Villa A E P. The expressive power of analog recurrent neural networks on infinite input streams[J]. Theoretical Computer Science, 2012, 436: 23-34

[11] Gao J Y, Ge R Z, Chen K, *et al*. Motion-appearance co-memory networks for video question answering[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 6576-6585

[12] Oh S W, Lee J Y, Xu N, *et al*. Video object segmentation using space-time memory networks[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 9225-9234

[13] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 7132-7141

[14] Woo S, Park J, Lee J Y, *et al*. CBAM: convolutional block attention module[C] //Proceedings of the 15th European Conference on Computer Vision. Heidelberg: Springer, 2018: 3-19

[15] Fu J, Liu J, Tian H J, *et al*. Dual attention network for scene segmentation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 3141-3149

[16] Guo M H, Lu C Z, Liu Z N, *et al*. Visual attention network[J]. Computational Visual Media, 2023, 9(4): 733-752

[17] Cordts M, Omran M, Ramos S, *et al*. The Cityscapes dataset for semantic urban scene understanding[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 3213-3223

[18] Brostow G J, Fauqueur J, Cipolla R. Semantic object classes in video: a high-definition ground truth database[J]. Pattern Recognition Letters, 2009, 30(2): 88-97

[19] Li J Y, Zhao Y K, Fu J, *et al*. Attention-guided network for semantic video segmentation[J]. IEEE Access, 2019, 7: 140680-140689

[20] Wu J R, Wen Z Z, Zhao S Y, *et al*. Video semantic segmentation via feature propagation with holistic attention[J]. Pattern Recognition, 2020, 104: Article No.107268

[21] Wang B W, Li L Z, Nakashima Y, *et al*. Noisy-LSTM: improving temporal awareness for video semantic segmentation[J]. IEEE Access, 2021, 9: 46810-46820

[22] Li Y L, Shi J P, Lin D H. Low-latency video semantic segmentation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 5997-6005

[23] Nilsson D, Sminchisescu C. Semantic video segmentation by gated recurrent flow propagation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec-

ognition. Los Alamitos: IEEE Computer Society Press, 2018: 6819-6828

[24] Jain S, Wang X, Gonzalez J E. Accel: A corrective fusion network for efficient semantic segmentation on video[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 8858-8867

[25] Hu P, Caba F, Wang O, *et al*. Temporally distributed networks for fast video semantic segmentation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 8815-8824

[26] Wang H, Wang W N, Liu J. Temporal memory attention for video semantic segmentation[C] //Proceedings of the IEEE International Conference on Image Processing (ICIP). Los Alamitos: IEEE Computer Society Press, 2021: 2254-2258

[27] Xu Y S, Fu T J, Yang H K, *et al*. Dynamic video segmentation network[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 6556-6565

[28] Sun G L, Liu Y, Ding H, *et al*. Coarse-to-fine feature mining for video semantic segmentation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 3116-3127

[29] Zhang Y Z, Borse S, Cai H, *et al*. AuxAdapt: stable and efficient test-time adaptation for temporally consistent video semantic segmentation[C] //Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Los Alamitos: IEEE Computer Society Press, 2022: 2633-2642

[30] Hu Y B, He Y Z, Li Y H, *et al*. Efficient semantic segmentation by altering resolutions for compressed videos[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2023: 22627-22637