视频序列中的每帧图像的每个像素点进行操作, 不仅帧数多, 而且还要考虑相邻帧之间的关系, 视频语义分割是动态的. 视频语义分割在自动驾驶、无人机导航、档案影像识别和可穿戴计算等领域有重要意义.

为解决视频分割问题, 研究者做了不同方面的尝试. 研究解决时空“超体素”[2,3]、无监督和运动驱动的对象分割[4–6], 或对标记视频进行弱监督分割[7–9]等, 这些方法不适用于实时或复杂的多类、多对象场景的语义分割. 利用 3D 场景结构解决视频语义分割的方法, 如: 在文献[10–12]中用在运动结构获得的三维点云构建模型, 基于这些几何的与/或运动特征, 改进语义分割; 在文献[13,14]中提出把在视频数据中得到的 2D 语义估计与 3D 场景重建结合起来, 虽然 3D 信息很丰富, 但信息的获取代价昂贵, 并且得到的预测错误很难解决. 还有一些采用快速滤波技术的方法, 对每帧图像先计算出语义标签信息再进行传播. 如: 在文献[15]中通过学习连续帧像素之间的相似函数去传递预测; 文献[16]实现了一个利用可学习的双边滤波器[17], 实现视频帧间信息的长距离传播.

还有一些方法利用帧序列之间的相关性. 因为视频帧的冗余性、还有数量的庞大性, 如果直接对整个视频序列进行处理, 不仅影响分割的结果, 而且耗费资源. 目前对视频语义分割的研究主要分为两个方面: 一是利用视频帧之间的时序信息提高图像分割的精度; 二是利用视频帧间的相似性确定关键帧, 减少计算量, 提高模型的运行速度和吞吐量. 本文介绍一些比较突出的视频语义分割方法.

1 数据集

目前用于视频语义分割的数据集主要有 2 个: CamVid 数据集[18]和 Cityscapes 数据集[19].

1.1 CamVid 数据集

CamVid 数据集[18]是第一个包含对象类语义标签的视频数据集, 由从白天和黄昏拍摄的驾驶视频中选取, 包含 701 张彩色图像, 并带有 11 个语义类的注释. 该数据集提供 ground truth 标签, 将每个像素与 32 个语义类中的其中一个相关联. 该数据集有 4 个视频片段组成, 每个视频片段平均包含 5000 帧, 分辨率为 720×960, 大约有 40 K 帧组成.

1.2 Cityscapes 数据集

Cityscapes 数据集[19], 即城市景观数据集, 用于城市场景理解和自动驾驶. 该数据集由从 50 个城市采集的街道场景视频片段组成, 帧率为 17 fps. 训练集、验证集和测试集分别包含 2975、500、1525 个视频片段. 每个视频片段有 30 帧, 只对第 20 帧做像素级的标注, 用于语义分割. 该数据集包含 30 个类, 数据集的标注分为 fine 和 coarse: fine 是对从 27 个城市中选择 5000 幅图像进行密集的像素级标注, 在 30 帧视频片段的第 20 帧上完成的, 目的是为了实现前景对象、背景和整体场景布局的高度多样性, 通过完整的注释提供上下文信息; coarse 是在剩下的 23 个城市中对每隔 20 s 或 20 m 的行驶距离 (无论哪个先到) 选择一张图像进行粗注释, 总共生成 2 万张图像, 以支持利用大量弱标记数据的方法.

2 提升精度的视频语义分割方法

现有 CNN 网络只能提取空间特征, 不能提取时序信息, 所以不能直接用现有的语义分割方法处理视频集. 根据视频是由连续的帧序列组成, 序列中包含时序信息, 利用时序信息可以把具有相同空间特征的不同类别对象区别开, 所以在视频语义分割任务中一般利用视频帧之间的时序信息提升分割精度.

2.1 STFCN

基于之前视频语义分割方法没有考虑视频序列中时序信息的特点, Fayyaz 等提出把 LSTM 模块[20,21]与 FCN[1]相结合, 构成端到端的时空卷积网络结构 (Spatio-Temporal FCN)[22], 用 STFCN 表示该时空模型. 把 LSTM 模块嵌入到 FCN-8 的 fc7 层中, 因为 fc7 是最深的全卷积层, 与较浅的层相比, 该层提取的特征比其它层提取的语义信息多. LSTM 网络是一种特殊的循环神经网络, 由一个记忆单元、一定数量的输入和一个输出门组成, 输出门用来控制序列中的信息流, 避免丢失重要信息. LSTM 可以用来解决梯度消失问题和记忆长时间的信息.

STFCN[22]模型的操作分为 4 步: 首先, 视频帧经过 FCN[1]提取空间特征, 得到该帧各区域的特征图; 其次, 将各区域特征图送入时空模型得到时空特征; 然后, 把时空特征送入时空分类器得到各区域基本的预测; 最后经过反卷积上采样操作, 恢复到与输入尺寸相同大小. 该模型在 CamVid[18]数据集上进行验证, 在一定程度上提升了分割性能. 但该模型没有充分考虑帧间的相关性, 且模型过于复杂, 无法达到实时的要求.

2.2 Netwarp 模型

基于视频数据帧数多, 静态语义分割模型不适用, Gadde 等提出一个可以处理视频数据的新技术, 即构造 Netwarp 模型[23], 将该模型与静态 CNN 相结合. 文献[24]中展示通过 CNN 中间层的相邻帧之间的特征变化缓慢, 尤其是在更深的卷积层中, 与文献[25]中基于运行时间的 Bilateral inception 模型为 Netwarp 提供理论依据. Netwarp 利用相邻帧之间的光流信息, 把通过 CNN 中间层的前一帧的特征 warp 到当前帧的相应位置, 光流定义为两张图像之间对应像素移动的向量. 该模型的输入是连续的两帧, 当前帧用 t 表示, 前一帧用 t–1 代表. 具体操作分为 3 步: 首先, 把用 Dis-Flow[26] 方法得到光流信息 $F(t)$ 送入 FlowTransformation 模块, 该模块用一个小的卷积神经网络 FlowCNN 传输信息, 表示为 $\wedge(F_t)$; 其次, 将 $\wedge(F_t)$ 与前一帧第 k 层的特征 warp 到当前帧得到 $\overset{\vee}{z}{}^{k}_{(t-1)}$; 最后把当前帧在第 k 层的特征与 $\overset{\vee}{z}{}^{k}_{(t-1)}$ 通过式 (1) 得到 z^k_t. 嵌入 Netwarp 模块的网络可以在线进行端到端的训练, 与逐帧操作相比计算开销更小. 并且该模块可以对网络的中间层进行优化, 在网络中可多次使用.

$$\tilde{z}\frac{k}{t} = w_1 \odot z\frac{k}{t} + w_1 \odot \overset{\vee}{z}\frac{k}{(t-1)} \tag{1}$$

2.3 时空变压器门控递归单元

基于视频数据集缺少高质量的标注与相邻视频帧之间包含大量冗余信息, 差异性显著区域的信息尤为重要, 所以 Nilsson 和 Sminchisescu 提出基于时空变压器门控递归单元 STGRU (Spatio-Temporal Transformer Gated Recurrent Unit) 的 GRFP 模型[27], 结合多帧未标注信息提高分割性能. 该模型由基于 Gated Recurrent Unit[28] 的 STGRU、基于 Spatial Transformer Network[29]的时空扭曲变压器和前向后向传播操作组成, 只对分割后的结果进行处理. STGRU 的本质是通过光流信息把当前帧的前后两帧的标签映射结合到当前帧, 考虑到前面帧的一些信息对当前帧的分割没有帮助, 所以使用门控思想让网络学习结合不同的语义图. 相邻帧的局部信息用卷积 GRU 学习, 可以把不同时间点的信息很好的融合.

STGRU 具体操作为: 首先, 计算相邻两帧的语义分割图以及光流; 其次利用光流把前一帧的结果 warp 到当前帧; 最后把 warp 后的结果与当前帧的分割图一起送入 GRU, 得到当前帧最终的分割结果. 该模型与多个图像分割网络相结合在标准数据集上进行验证, 试验表明 STGRU 可以在不增加额外标签、占用很少计算量的情况下, 只用标签视频帧就能向邻近无标签视频帧传播信息, 并且在提升语义分割性能的同时保证时间的一致性.

2.4 新 PERAL 模型

基于视频序列的时间连续性, Jin 等提出通过预测未来帧学习判别特征, 并结合预测结果和当前特征来解析帧的新 PEARL (Parsing with prEdictive feAtuRe Learning) 模型[30]. 与之前的场景预测学习模型 PEARL 最大的不同是增加了预测学习网络, PEARL 包含 2 个预测学习阶段. 第一阶段 (无监督学习) 中采用类似 GAN[31]网络结构, 在未标记的视频数据中预测未来帧, 实现对时间特征的学习. 预测学习网络作为生成器 G, 通过特征提取器将输入的视频序列映射到时间表示上; 再用上采样层对其进行空间放大, 最后反馈给卷积层, 生成像素级的 RGB 值; 判别器 D 对 G 生成的图像与真实图像进行判别. 第二阶段将预测学习网络转移到预测解析任务中, 把 G 的生成图片和输入视频序列的下一帧相结合, 通过上采样和卷积操作实现时间平滑和结构保持, 得到最终的分割结果. 与之前对单帧的简单分割相比, 该模型的分割的效果更好.

3 减少计算量

视频数据具有连续帧之间大部分区域不变性与局部区域相对变化明显的特点, 变化明显的视频帧中往往包含着丰富的目标运动信息, 关键帧的选取是解决目前视频任务的重点. 选取关键帧的依据是视频的变化程度, 而不依赖于视频数据的长短, 在变化明显的视频中选取较多的关键帧, 没有明显变化的视频数据中选取较少的关键帧. 所以, 关键帧的选取主要依据视频帧之间的变化程度.

3.1 Clockwork FCN

Darrell 等人依据以下 2 点提出 Clockwork FCN[32]: 一是视频帧序列之间的像素点变化迅速, 但是帧的场景语义内容变化缓慢; 二是把执行视为结构的一方面, 为网络生成特定的计算时间表. 受 Clockwork 循环网络[33]的影响, 作者定义一个由固定或自适应信号驱动的新的 Clockwork 卷积族. 把新的 Clockwork 与全卷积网络 FCN[1]相结合形成 Clockwork FCN 模型, 完成

跨帧传播任务, 网络结构如图 1 所示. 该模型把网络层分为不同的阶段, 每个阶段有不同的更新率. 时钟控制网络的具体操作: 经过第一阶段后, 计算只在特定的时钟信号点执行; 静态场景期间缓存一直持续, 当遇到动态场景时开始新的计算, 并且输出与前面的静态特征相结合, 以此达到减少计算量的目的.

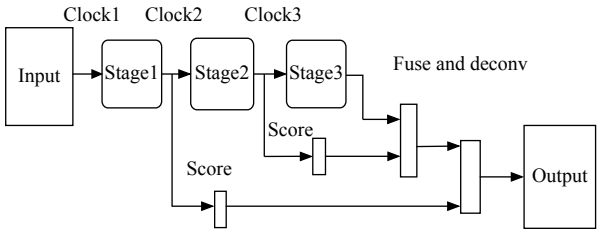

图 1 Clockwork FCN 框架

3.2 Deep Feature Flow (DFF)

CNNs 中间卷积特征图与输入图像有相同的空间范围, 保持低层图像内容与中高层语义概念[33]之间的空间对应关系, 类似于光流[34,35]提供通过空间扭曲在邻近帧之间廉价传播特征的机会. 基于光流估计与特征传播比通过卷积计算特征速度更快, 且相邻帧之间的差别不大, 所以 Wei Y 等提出 DFF[36]. DFF 基于视频数据中相邻帧之间差异性小, 采用固定间隔 k 选取一帧为关键帧, 通过把关键帧的深度特征映射传给其他帧的方法减少网络计算量. DFF 网络结构如图 2 所示.

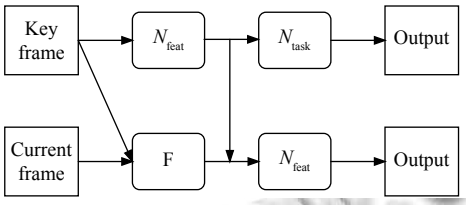

图 2 Deep Feature Flow

DFF 可以分为两个连续的子网络, N_{feat} 特征网络与 N_{task} 任务网络. I 代表输入图片, 连续帧序列用 I_i 表示, i=0, 1, …, N 代表卷积神经网络, 图片的输出为 $y=N(I)$.N_{feat} 特征网络由全卷积网络组成, 输入为每帧图像, 输出为多个中间特征图, $f=N_{feat}(I)$.N_{task} 任务网络, 根据不同的任务有不同的结构, 输入是特征图 f, 输出为 $y=N_{task}(f)$. 因为连续视频帧之间的相似性, 与经过深层网络后得到的特征图之间的强相关性, 特征网络 N_{feat} 只对关键帧进行操作, 非关键帧 I_i 的特征图通过前面关键帧 I_k 的特征传播得到. 通过对稀疏关键帧的操作, 在极大的减少视频序列计算量的同时, 精确度损失适中. 该方法为减少计算量提供了一个新的方向, 因为是固定间隔的选取关键帧, 所以该方法适用在环境变化缓慢的场景. 如果用在剧烈变化的场景中会丢失信息, 产生低精度的分割结果.

3.3 Low-Latency 视频语义分割

为增加视频语义分割在现实世界中的应用, 要求在减少计算量的同时最大程度的减少时延, Li Y 等提出 Low-Latency[37]的视频语义分割框架. 该框架有 2 个组成部分: 一是特征传播模型, 通过空间变量卷积 (spatially variant convolution) 自适应的融合特征, 减少对每帧的计算量; 二是基于精确度预测自动分配计算量的自适应调度程序, 选择关键帧. 该方法首先将视频序列的第一帧设为关键帧, 以后每隔 t 时刻选一帧与前一关键帧进行比较, 确定是否为关键帧. 该模块对关键帧和非关键帧的特征传播方式不同, 对于关键帧直接通过 $S(h)$ 得到深层特征; 对于非关键帧用前一关键帧的特征进行传播.

3.3.1 自适应特征传播模块

空间变量卷积 (spatially variant convolution) 是用卷积表示领域的线性组合, 卷积核在不同的位置上有不同变化. 空间变量卷积核的权值通过权值预测器 (weights predictor) 确定, 该网络由三层相互嵌套的卷积层与 Relu 组成. 输入为当前帧与前一关键帧的低层特征图 F_l^t 与 F_l^k, 最后一个卷积层的输出尺寸为 $H_K^2 \times H \times W$, 其中 $H \times W$ 是高层特征图的尺寸, 即每个位置输出一个 H_K^2 通道向量, 再将其转换为该位置 $H_K \times H_K$ 尺寸的卷积核. 卷积层的输出通过 Softmax 层进行归一化处理, 使得每个卷积核权值之和为 1. 特征传播流程图如图 3 所示, $S(l)$ 代表网络提取低层特征 (low level feature), $S(h)$ 代表获得深度特征 (deep feature), I^k 为关键帧, I^t 是 I^k 隔 t 时刻后选取的一帧, WeightPredictor 为权值预测器, Spatially Variant Convolution为空间变量卷积, F_h^k 为 I^k 的高层特征图, F_h^t 为 I^t 的高层特征图.

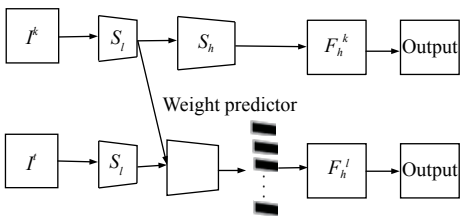

图 3 Low-Latency 自适应传播流程图

3.3.2 关键帧选择模块

关键帧的选取基于每帧的低层特征 (low level features). 因为当当前帧的内容发生巨大的变化时, 低层特征比如边缘位置信息的差异性一定会很大, 而且获取底层特征比获取高层特征的代价更小. 该预测网络有两个 256 通道的卷积层、全局池化层与全连接层组成, 输入为当前帧与前关键帧的低层特征图, 输出是当前帧为关键帧的可能性. 如果输出值大于设定的阈值就输出 1, 表示当前帧的内容和前关键帧的差距比较大, 更新关键帧; 否则输出为 0, 代表不更新关键帧.

3.4 DVSNet

为快速高效的对视频数据进行语义分割操作, Yu S 等人提出了动态视频语义分割网络 DVSNet (Dynamic Video Segmentation Network)[38]. 该网络由两个子网络组成: 分割网络和光流网络. 分割网络生成高精度的语义分割图, 但是运行速度慢、层数多; 流网络运行速度快, 但其输出结果需要进一步的处理, 才能得到低精确度的语义分割图. 该网络的操作分为 3 步: 首先, 将视频帧分为 4 个相同尺寸的帧区域, 为避免切分处信息的丢失, 每个区域增加 64 个重复的像素值; 其次, 将当前帧与前一关键帧相应位置的帧区域对送入决策网络 DN, DN 根据期望置信分数 (expected confidence score) 与设定阈值进行比较, 决定把当前帧区域送入哪个网络: 如果期望置信分数比设定阈值小, 对应区域送入分割网络; 反之, 对应区域送入包含空间扭曲的流网络; 最后, 不同路径的帧区域得到不同的分割图.

3.4.1 DN 决策网络

DN 通过估计该区域的期望置信分数与阈值的比较, 决定是否把该区域送入分割网络. 阈值是提前设定的, 可根据不同的任务设置不同的阈值, 并且阈值的大小与分割的精度和帧率有关: 较小的阈值产生低的精确度和高的帧率, 大部分输入帧区域送入空间扭曲路径; 较大的阈值产生高的精确度, 但速度有所降低.

DN 网络是 1 个轻量级的 CNN, 由 1 个卷积层、3 个全连接层组成. DN 网络的作用是评估空间扭曲路径产生的分割结果 (O_C) 与分割路径的结果 (S_C) 之间的相似性. DN 的输入是光流网络第 6 层的输出特征图, 输出为期望置信分数. 在训练阶段, DN 网络的目标是学习预测一个帧区域的期望置信值, 尽可能的接近真实的置信值. 首先, 将预测得到的期望置信分数与真实置信分数进行比较, 计算均方误差 (MSE) 损失; 其次, 根据 MSE 用 Adam optimizer 更新 DN 中的参数. 在执行 (测试) 阶段, DN 和光流网络不访问真实置信分数. DN 首先分别对 4 个帧区域对进行处理, 得到 4 个期望置信分数; 然后把期望置信分数与预先设置的阈值比较: 如果比设定阈值小, 对应区域送入分割网络; 如果比设定阈值大, 对应区域送入包含空间扭曲的流网络.

4 总结与展望

当前对于视频语义分割的研究主要分为 2 类, 一是高层建模, 二是特征传播.

4.1 高层建模

在高层建模方面, 一般是在已有网络结构上增加额外的层, 提升分割精确度. 通过设计不同的模块并与现有 CNNs 网络相结合, 如: STFCN[22]模型是在 FCN[1] 结构上增加 LSTM 模块, 利用视频序列的时序信息提升精度, 但该模型的实现过程太过复杂, 且没有考虑相邻帧之间的相关性, 如果每秒的帧数过高, 则需要对每帧进行处理, 计算成本高; Netwarp[23]模块是利用相邻帧的光流信息实现跨时间的内在网络特征的 warping (扭曲), 该模块可以与现存的 CNNs 相结合实现端到端的训练, 并提升性能; 时序门控循环流组件 STGRU[27] 可以嵌入到静态语义分割结构, 将其转化为弱监督的视频处理结构, 在 Cityscapes[19]和 CamVid[18]数据集上都取得了比原有网络好的结果, 在一定程度上提升了性能; 新 PERAL 模型在原有基础上增加预测学习网络, 用类似 GAN[31]的结构把预测结果与待分割的视频帧进行微调, 利用时间一致性提升分割精度.

4.2 特征传播

在特征传播方面, Clockwork Net[32]采用多级 FCN[1], 对网络中不同层次的特征映射使用不同的更新周期, 并在一定的网络层直接重用前一帧的第二级或第三级特性, 节省计算量. 虽然其高层特性相对稳定, 但这种简单的复制并不能得到最佳结果, 特别是当场景发生巨大改变时. DFF[36]通过流网络[35]中学习到的光流信息, 将高层特征从关键帧传播到当前帧, 获得了较好的性能. 但单独的流网络增加计算成本, 光流逐像素的位置变换可能会丢失视频帧中的空间信息. 在 Clockwork Net[32]与 DFF[36]中关键帧的选择对整体性能至关重要, 但以上两种方法只简单的使用固定帧间间隔调度[25, 31] 或启发式阈值方案[25]选择关键帧, 并没有提供详细的研究.

由表 1 可知 Low-Latency 与其他方法相比所花费的时间和延时都是最低的. 就特征传播而言, Clockwork Net [32]直接重用特征, 在一定程度上减少了计算量, 但精度不高. GRFP 是通过预测学习对分割结果过进行微调, 与 Clockwork Net 相比精确度略高, 但耗费的时间更长, 属于用时间换精确度. DFF[36]与 Clockwork Net 相比的分割精度更高, 但后者采用固定帧间间隔的方法选择关键帧, 用光流传播特征, 忽略视频帧的空间对应关系, 延时效果最差. 并且当视频快速变化时, 会丢失一些重要信息, 影响分割结果. 与以上方法相比, Low-Latency 在延时和精确度方面得到了权衡, 在降低延时的同时, 保证精确度的稳定, 并且该方法可以在线设置中保证低延时.

表 1 几种方法的实验结果

Method	mIoU (%)	Average runtime (ms)	Latency (ms)
ClockworkNet	67.7	141	360
GRFP	69.4	470	470
DFF	70.1	273	654
Low-Latency	75.89	119	119

为提升视频语义分割在自动驾驶等领域的可能性, 解决 DFF 固定间隔确定关键帧方法的弊端, Low-Latency[37]与 DVSNet[38]采用动态的关键帧更新方法, 即每隔时间间隔 t 选一帧与前一关键帧比较, 决定是否更新关键帧. Low-Latency[37]根据当前帧与前一关键帧低层特征之间的偏移量与阈值的比较结果确定该帧是否为关键帧, 实现关键帧的自适应调度. 选取低层特征作为比较依据是因为与获取高层特征相比, 低层特征所需的时间少, 并且只对关键帧提取高层特征可以减少计算量. 而 DVSNet[38]是将视频帧分为 4 个相同的尺寸, 把对应帧区域对送入 DN 网络, DN 根据期望置信分数与阈值的比较结果确定是否更新关键帧, 实现关键帧的动态调度. 这 2 种方法都选择每隔时间间隔 t 选一帧与前一关键帧进行比较, 前者依据低层特征之间的偏移量确定是否更新关键帧, 后者则根据帧区域对的置信分数确定是否更新关键帧.

Low-Latency 的目的是在最大程度减小时延的同时保证精确度的稳定, 尽可能的满足视频语义分割实时性的要求. DVSNet[38]通过改变阈值的大小实现精度与速度的调整, 在标准数据集 Cityscapes[19]上验证, 证明该方法在帧率为 19.8 fps 的情况下 mIoU 为 70.4%、30.4 fps 情况下得到 63.2% 的 mIoU. 以上方法都是在视频语义分割道路上的探索, 利用视频数据的时间一致性, 通过特征传播、信息重用、更新关键帧等方式减少计算量、提升分割精度、降低时延, 但目前的研究还不能完全满足自动驾驶等领域对实时性以及精确度的要求.

4.3 展望

以上方法在一定程度上促进了视频语义分割的发展, 简单模型的组合不能很好地适应时代的要求, 依据选择关键帧的方法在减少计算量的同时提升精确度.

(1) 关键帧选择

目前对关键帧的选择: 一是固定帧间间隔确定; 二是固定时间间隔选择一帧与前一关键帧比较确定是否更新关键帧. 目前的方法并没有明确给出详细的阈值计算方法, 而且阈值的设定太过主观. 未来可以在确定关键帧方面进行改进, 用相对客观的方法选择关键帧, 如可以考虑 Siamese[39]比较相邻帧间的相似性, 在减少计算量的同时, 更注重分割精确度, 使视频语义分割技术更好的应用在自动驾驶等领域.

(2) 数据集

用于视频语义分割的数据集精细标注很少, 训练过程中信息易缺失. 因为数据集的有限性, 使得模型的迁移能力差, 不能很好地适应未训练的数据集. 并且在现实世界中, 场景类型多变, 所面对的挑战也更大. 可以考虑将真实数据集与虚拟场景数据集相结合, 提升模型迁移学习的能力.

(3) 特征提取

现有视频语义分割方法都是在静态语义分割的基础上改进, 对图像语义分割的改进可以提升视频语义分割的性能. 如为使输出分割图与输入尺寸相同需对后三层特征图进行融合, 该方法可能丢失信息. 为尽可能多的提取特征可以把每一个卷积池化操作后的特征图都与后面各层特征图进行融合, 最后再通过跳跃结构融合特征, 提升分割质量. 提升图像特征提取能力可以进一步的提高视频语义分割精度, 这也是未来对视频语义分割研究的一个方向.

参考文献

1 Long L, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 3431–3440.

2 Grundmann M, Kwatra V, Han M, *et al*. Efficient hierarchical graph-based video segmentation. Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, CA, USA. 2010. 2141–2148.
3 Xu CL, Corso JJ. Evaluation of super-voxel methods for early video processing. Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI, USA. 2012. 1202–1209.
4 Shi JB, Malik J. Motion segmentation and tracking using normalized cuts. Proceedings of the 6th International Conference on Computer Vision. Bombay, India. 1998. 1154–1160.
5 Papazoglou A, Ferrari V. Fast object segmentation in unconstrained video. Proceedings of 2013 IEEE International Conference on Computer Vision. Sydney, NSW, Australia. 2013. 1777–1784.
6 Fragkiadaki K, Arbelaez P, Felsen P, *et al*. Learning to segment moving objects in videos. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 4083–4090.
7 Hartmann G, Grundmann M, Hoffman J, *et al*. Weakly supervised learning of object segmentations from web-scale video. Proceedings of European Conference on Computer Vision. Florence, Italy. 2012. 198–208.
8 Tang K, Sukthankar R, Yagnik J, *et al*. Discriminative segment annotation in weakly labeled video. Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, OR, USA. 2013. 2483–2490.
9 Liu X, Tao DC, Song ML, *et al*. Weakly supervised multiclass video segmentation. Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA. 2015. 57–64.
10 Brostow GJ, Shotton J, Fauqueur J, *et al*. Segmentation and recognition using structure from motion point clouds. Proceedings of the 10th European Conference on Computer Vision. Marseille, France. 2008. 44–57.
11 Floros G, Leibe B. Joint 2D-3D temporally consistent semantic segmentation of street scenes. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI, USA. 2012. 2823–2830.
12 Sturgess P, Alahari K, Ladicky L, *et al*. Combining appearance and structure from motion features for road scene understanding. Proceedings of the British Machine Vision Conference. London, UK. 2009.
13 Kundu A, Li Y, Dellaert F, *et al*. Joint semantic segmentation and 3D reconstruction from monocular video. Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland. 2014. 703–718.
14 Sengupta S, Greveson E, Shahrokni A, *et al*. Urban 3D semantic modelling using stereo vision. Proceedings of 2013 IEEE International Conference on Robotics and Automation. Karlsruhe, Germany. 2013. 580–585.
15 Miksik O, Munoz D, Bagnell JA, *et al*. Efficient temporal consistency for streaming video scene analysis. Proceedings of 2013 IEEE International Conference on Robotics and Automation. Karlsruhe, Germany. 2013. 133–139.
16 Jampani V, Gadde R, Gehler PV. Video propagation networks. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 3154–3164
17 Jampani V, Kiefel M, Gehler PV. Learning sparse high dimensional filters: Image filtering, dense CRFs and bilateral neural networks. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 4452–4461.
18 Brostow GJ, Fauqueur J, Cipolla R. Semantic object classes in video: A high-definition ground truth database. Pattern Recognition Letters, 2009, 30(2): 88–97. [doi: 10.1016/j.patrec.2008.04.005]
19 Cordts M, Omran M, Ramos S, *et al*. The cityscapes dataset for semantic urban scene understanding. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 3213–3223.
20 Gers FA, Schmidhuber J, Cummins F. Learning to forget: Continual prediction with LSTM. Neural Computation, 2000, 12(10): 2451–2471. [doi: 10.1162/089976600300015015]
21 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]
22 Fayyaz M, Saffar MH, Sabokrou M, *et al*. STFCN: Spatio-temporal FCN for semantic video segmentation. arXiv preprint arXiv: 1608.05971, 2016.
23 Gadde R, Jampani V, Gehler PV. Semantic video CNNs through representation warping. Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy. 2017. 4463–4472.
24 Shelhamer E, Rakelly K, Hoffman J, *et al*. Clockwork convnets for video semantic segmentation. Proceedings of European Conference on Computer Vision. Amsterdam, The Netherlands. 2016. 852–868.
25 Gadde R, Jampani V, Kiefel M, *et al*. Superpixel

convolutional networks using bilateral inceptions. Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands. 2016. 597–613.

26 Kroeger T, Timofte R, Dai DX, *et al*. Fast optical flow using dense inverse search. Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands. 2016. 471–488.

27 Nilsson D, Sminchisescu C. Semantic video segmentation by gated recurrent flow propagation. arXiv preprint arXiv: 1612.08871, 2016.

28 Ilg E, Mayer N, Saikia T, *et al*. FlowNet 2.0: Evolution of optical flow estimation with deep networks. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 1647–1655.

29 Jaderberg M, Simonyan K, Zisserman A, *et al*. Spatial transformer networks. Proceedings of Advances in Neural Information Processing Systems. Montreal, QB, Canada. 2015. 2017–2025.

30 Jin XJ, Li X, Xiao HX, *et al*. Video scene parsing with predictive feature learning. Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy. 2017. 5581–5589.

31 Goodfellow IJ, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial nets. Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, QB, Canada. 2014. 2672–2680.

32 Koutník J, Greff K, Gomez F, *et al*. A clockwork RNN. Proceedings of the 31st International Conference on International Conference on Machine Learning. Beijing, China. 2014. II-1863–II-1871.

33 Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland. 2014. 818–833.

34 Horn BKP, Schunck BG. Determining optical flow. Artificial Intelligence, 1981, 17(1–3): 185–203. [doi: 10.1016/0004-3702(81)90024-2]

35 Dosovitskiy A, Fischer P, Ilg E, *et al*. FlowNet: Learning optical flow with convolutional networks. Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile. 2015. 2758–2766.

36 Zhu XZ, Xiong YW, Dai JF, *et al*. Deep feature flow for video recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 4141–4150.

37 Li YL, Shi JP, Lin DH. Low-latency video semantic segmentation. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. 2018. 5997–6005.

38 Xu YS, Fu TJ, Yang HK, *et al*. Dynamic video segmentation network. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. 2018. 6556–6565.

39 Chopra S, Hadsell R, LeCun Y. Learning a similarity metric discriminatively, with application to face verification. Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, CA, USA. 2005. 539–546.