文章编号:1007－6735(2026)02－0149－11　DOI: 10.13255/j.cnki.jusst.20250212002

基于分流存储网络的半监督视频对象分割

郭旭丰[1]，　王朝立[1]，　孙占全[1]，　蒋纯国[2]
（1. 上海理工大学 光电信息与计算机工程学院，上海 200093；2. 上海出版印刷高等专科学校 体育部，上海 200093）

摘要：半监督视频分割的目标是在给定首帧视频标注的条件下，持续跟踪并分割出后续视频帧中的目标。许多基于时空存储方法的网络，利用过去存储的视频帧背景与目标特征来引导后续帧目标的分割，已经取得了显著的效果。但这些网络多数是按时间顺序存储视频帧，导致存储库不断扩大，增加了匹配分割时间，且存储特征中不断累积的扰动与噪声也降低了分割的精确度。受记忆曲线启发，提出一种分流存储网络来解决上述问题。网络将存储库分为 3 个级别：初级、中级和高级，并引入分流存储模块作为分流准则，控制特征存储的流向。通过采用不同策略，使网络在提升精确率的同时，缩短了分割时间。在训练数据保持一致的情况下，本文网络与同类型网络相比，在单对象数据集 DAVIS16 与多对象数据集 DAVIS2017 上均取得了较好效果。

关键词：计算机视觉；深度学习；视频分割；半监督学习
中图分类号：TP 18　　**文献标志码：**A

Semi-supervised video object segmentation based split-flow memory network

GUO Xufeng[1], WANG Chaoli[1], SUN Zhanquan[1], JIANG Chunguo[2]
(1. *School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China*; 2. *Ministry of Sports, Shanghai Publishing and Printing College, Shanghai 200093, China*)

Abstract: The goal of semi-supervised video object segmentation is to continuously track and segment out the target objects in subsequent video frames given the first frame video object labeling. Many networks based on space-time memory have achieved significant results by utilizing the background and object features of past stored video frames to guide the segmentation of objects in subsequent frames. However, most of these networks simply store video frames in chronological order, resulting in an ever-expanding memory bank that increases the matching segmentation time, while the accumulating

收稿日期: 2025−02−12
基金项目: 国家自然科学基金资助项目 (62173232)
第一作者: 郭旭丰（1999—），男，硕士研究生．研究方向：计算机视觉．E-mail：18383090068@163.com
通信作者: 王朝立（1965—），男，教授．研究方向：非线性控制、图像分割．E-mail：clwang@usst.edu.cn

引文格式: 郭旭丰，王朝立，孙占全，等．基于分流存储网络的半监督视频对象分割[J]．上海理工大学学报，2026，48(2)：149-158.
Citation: GUO Xufeng, WANG Chaoli, SUN Zhanquan, et al. Semi-supervised video object segmentation based split-flow memory network[J]. Journal of University of Shanghai for Science and Technology, 2026, 48(2): 149-158.

perturbations and noises in the stored features also reduce the segmentation accuracy. Inspired by the memory curve, a split-flow memory network was proposed to solve the above problems. The network divided the memory bank into three levels: primary, middle and senior, and introduced a split-flow memory module to set standards and control the feature memory flow. By adopting different strategies, the network improved the accuracy rate while shortening the segmentation time. With consistent training data, the network achieved optimal results on both single-object dataset DAVIS16 and multi-object dataset DAVIS2017 compared with the same type of networks.

Keywords: *computer vision*; *deep learning*; *video object segmentation*; *semi-supervised learning*

半监督视频分割是视频分割中一个重要分支，它解决的问题是在给定一段视频序列以及视频第一帧中特定对象分割掩码的情况下，在后续视频序列中逐帧跟踪并分割出首帧标注的对象。目前半监督视频分割已经有了广泛的实际运用，比如视频编辑、视频分析与自动驾驶等[1]。

半监督视频分割是一项具有挑战性的任务，其难点在于只有首帧对象信息，后续视频帧中对象的外观位置都会不断地发生改变，同时视频帧中背景信息也会影响对象的识别，例如背景中的遮挡物可能会遮挡目标对象，或者与对象相似的物体可能会误导识别过程等。最近，一些基于深度学习的方法在这个任务上已经取得了较好的效果，主要有基于传播的方法[2−3]、基于检测的方法[4−5]与基于匹配的方法[6]等。其中，有一种方法是基于时空存储匹配的方法[7]，其原理是在视频分割过程中，网络会把每一帧中原始与分割得到的背景与对象特征存储下来，后续视频帧将利用这些存储特征进行分割，不断分割和存储，直到视频分割完成。目前，这种时空存储匹配网络凭借其高分割精度已被视为视频分割的主流网络框架，许多新的网络变体基于此框架被提出，如 STCN[8]、MiVOS[9]、MRP[10] 等。

这类网络的核心部分是存储库，然而大部分网络中的存储库只是简单地将视频帧特征直接存入，更新方式也是采用简单的排列方法先进先出。这种设计存在几个缺陷：a. 简单的逐帧存储视频特征会导致内存消耗迅速增加，一方面会增加设备硬件的需求，一旦所需存储内存超出设备的内存容量，就会导致网络运行中断；另一方面不断扩充的数据会延长后续视频帧的分割时间，因为后续帧都将与存储库中所有特征进行相似度比较，所以后续分割时间会跟随存储数据量增加而增加。上述两方面限制了网络的使用条件与范围，降低了容错率。b. 通常情况下，视频连续几帧之间存在着一些相同的对象背景信息，比如一些始终没有移动的物体，因此逐帧存储特征会导致部分特征信息冗余，消耗内存的同时也延长了匹配分割的时间。c. 不加限制地存储分割掩码信息会导致其中的扰动与噪声不断累积，从而误导分割匹配，降低分割结果的精度[11]。d. 先进先出的简单存储更新机制会在具有挑战性的情况下受到限制[12]，比如对象长时间被遮挡的情况下更新删除了未被遮挡时对象特征而保留了大量被遮挡情况下的背景特征，这样就容易导致后续视频帧对象跟踪的丢失。

针对以上问题，本文网络主体结构采用时空存储网络的主干网络，每一视频帧经过编码器编码后，与 3 种存储库内的对象特征信息进行相似度比较以获取匹配输出，其中包含当前视频帧对象的详细语义信息。随后，解码器将这些匹配输出解码，生成最终的对象分割掩码，然后与当前视频帧一起被送到分流存储模块进行特征分析。分流模块会根据分析后得到的存储分流系数来判定当前视频帧存储的流向，将其存储到对应存储库内。不同的存储库则会对存储进来的特征进行特定的存储与更新操作，至此代表一帧的处理结束。上述操作过程会不断循环，直到视频中所有帧都完成分割。

本文设计了 3 种存储库：初级存储库、中级存储库和高级存储库，存储着不同数量和质量的视频背景特征信息，在视频分割中起到了不同程度的指导作用，并采用不同的存储与更新操作来控制内存消耗。提出了分流存储模块，制定了分流存储机制，以此来决定待存储视频帧特征的存储流向。本文方法与同类型时空存储网络相比，

在单对象数据集 DAVIS2016[13] 与多对象数据集 DAVIS2017[14] 上均取得了最优结果。

1　相关工作

时空存储网络的核心组件在于其存储库模块，该模块的信息存储机制与动态更新策略直接制约着网络的分割精度和计算时效性。因此，本研究重点集中于存储架构的优化设计。模仿人类大脑对新事物遗忘的规律，德国心理学家 Ebbinghaus[15] 提出了一种遗忘曲线，对此加以利用，可以提升自我记忆能力，该曲线对人类记忆认知研究产生了重大影响。该遗忘曲线提出人类记忆效能主要受控于 3 个关键维度：a. 信息暴露时长。记忆巩固程度与感官输入持续时间呈正相关，瞬时接触仅能形成语义轮廓的表征，而持续性输入可促进细节特征的神经编码。b. 信息置信水平。高置信度信息会触发大脑的稳态存储机制，这类信息被优先编码为长期记忆单元。当遭遇相似输入时，海马体–新皮层回路会启动模式匹配机制，通过贝叶斯推断对信息真实性进行验证，验证通过后将触发记忆突触的权重强化过程。c. 认知注意偏向。高兴趣度信息会激活前额叶–边缘系统的协同响应，促使多模态感知通道的同步编码，显著提升记忆编码速率和存储强度。此外，此类信息具有记忆触发因子的特性，可引发关联记忆网络的级联激活。基于上述认知神经学原理，本文对时空存储网络的存储库进行了结构化重构。具体来说，本文将原网络中的单一的存储库拆分为 3 个部分，命名为：初级存储库、中级存储库和高级存储库。3 个存储库内分别存储不同的内容，并在网络运行时协同工作实现视频分割。各存储层级的更新策略分别对应遗忘曲线理论中的 3 个关键维度，接下来将具体阐述各个存储库的设计思路。

初级存储库主要关注与信息接触时间的长短。本文的研究对象是视频，视频中的信息变化迅速。初级存储库旨在存储那些不随时间显著变化的特征信息，比如视频中背景和对象的大致外观与位置信息，这些信息是比较粗糙的，但是会一直重复出现。初级存储库的主要功能是在后续帧与存储内容进行匹配时，能够在存储少量特征信息的情况下，快速确定视频的大致外观与位置，减少冗余的匹配操作。至于视频的具体语义信息，则由其他两个存储库进行补充。

中级存储库设计体现的是信息自身的置信度。该存储库内存储了多帧视频的细节纹理与语义信息，在后续帧与初级存储库内信息匹配后，中级存储库用于进一步细化和补充分割对象的具体信息。为了预防内存爆炸，中级存储库将只存储一定数量的对象细节纹理信息。当存储信息数量达到上限时，新增信息将会与已存储信息进行置信度评估，保留可靠性较高的信息，淘汰可靠性较低的信息。通过这种不断的更新迭代，最终存储库将包含极具真实性与合理性的对象信息。因此，后续帧在参考这些信息进行分割时，其精确性自然会得到提升。

高级存储库的设计遵循的是对信息感兴趣的强弱，视频分割核心任务是捕获目标对象的真实信息。与前两种存储库相比，高级存储库存储了更多关于对象本身的详细信息，在所有存储信息中其置信度最高，且信息存储时间最久，存储的信息数量有一定限制。当存储容量到达上限时，存储库将根据时间顺序，优先删除最早存储的信息。

虽然上述 3 种存储库的更新机制各不相同，但它们都需要协同工作、互不干扰地发挥各自的优点，合力完成视频帧的精确分割与存储任务。随着存储库设计的完成，如何合理地将视频帧特征存储到相应的存储库中，成为一个新的设计问题。通过分析上述存储库设计的差异后可以总结出影响视频帧存储流向的因素：视频帧特征信息的丰富程度。具体来说，3 种存储库内存储的视频纹理细节与语义特征随着级别的提高而逐级增多。其中，高级存储库拥有最丰富和最真实的对象特征，中级存储库次之，而初级存储库由于其只有对象大致的位置与外观信息而排在最后。因此，网络只需识别每个视频分割帧的特征信息丰富程度后，根据存储划分区间即可决定该视频帧的存储流向。受 Mask scoring R-CNN[16] 的启发，本文认为视频提取特征与对应标签特征的接近程度是衡量特征丰富程度的关键。那些特征提取程度较低的帧将导向初级存储库，用于更新对象外观与位置的粗略信息；特征提取程度居中的视频帧则流向到中级存储库，来更新对象的语义信息；而特征提取程度高的帧会被送到高级存储库

当中，用于指导分割对象边缘细节。这样的分流策略构成了一个合理的存储管理标准。

2 方 法

2.1 网络模型

本文的网络框架如图1所示，针对时空存储类型网络进行改进，以目前主流的时空存储(STM)网络框架[7]为基础进行了拓展。网络在给定一段视频序列与首帧对象标注掩码的情况下，逐帧地定位并分割出对象，尽可能地达到精确率高、分割速度快。首先，使用特征编码器提取视频中对象与背景特征信息得到查询键值对；随后，将其与多级存储库内的存储键值对依次进行像素级匹配，获得聚合匹配特征值，该值会被送入到解码器解码输出对象分割掩码；最后，由分流存储模块评估当前视频帧与其分割掩码，再将其更新到相应存储库当中，不断重复上述流程，直到所有视频帧分割结束。

2.2 编码器

特征编码器的输入为待分割的视频帧。整个编码器主干采用了预训练过的ResNet-50[17]来对视频帧进行下采样编码操作，丢弃了Res5层与全连接层，在Res4层后额外连接多层3×3卷积层，最终得到该视频帧编码后的特征映射键值对：查询键$K_q \in \mathbb{R}^{C^k \times H \times W}$和查询值$V_q \in \mathbb{R}^{C^v \times H \times W}$。其中：$H$、$W$分别为下采样后图像的高度和宽度；$C^k$、$C^v$分别为键和值的空间通道维度。

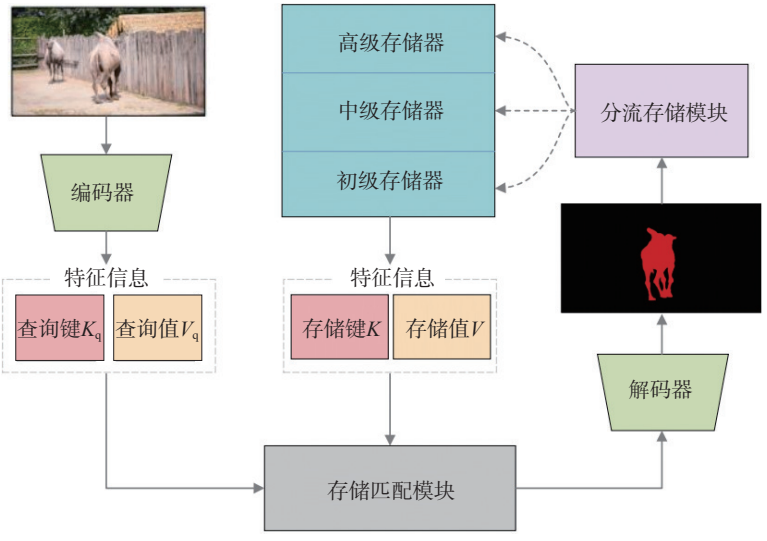

图1 网络框架图

Fig.1 Network framework diagram

2.3 存储匹配模块与解码器

存储匹配模块主要功能是将编码后的视频帧与存储库中的对象特征进行比较，这样可以从存储帧中提取有用特征，并以此推断出当前视频帧对象的具体外观、位置等语义信息，最后由解码器解码输出对象分割掩码。其中的存储匹配流程是首先计算查询键K_q与存储键K计算相似度得到相似权重，随后通过权重来检索存储值中V的重要值，最后再与查询值V_q相连接，补充背景信息得到匹配输出特征F_m，整个操作过程可以概括为

$$F_m = V * \mathcal{W}\left(K_q, K\right) \oplus V_q \tag{1}$$

式中：$K \in \mathbb{R}^{C^k \times N \times H \times W}$，$V \in \mathbb{R}^{C^v \times N \times H \times W}$；$N$为存储库内所有键的和；$\mathcal{W}(K_q, K)$是使用点积操作计算$K_q$和$K$之间的相似度，$\mathcal{W}(K_q, K) \in \mathbb{R}^{NHW \times HW}$；$\oplus$代表在通道维度上进行连接操作。

采用双线性上采样操作，将特征还原到原始视频帧的分辨率，最终输出得到当前视频帧的掩码分割图$M \in \mathbb{R}^{1 \times H \times W}$。

2.4 分流存储模块

分流存储模块如图2所示。首先，将当前视频帧与其分割掩码一起送入到分流编码器，为了降低计算复杂度，这里的编码器主干也采用的是预训练后的ResNet-50，同时也丢弃Res5层与后续全连接层，得到编码特征$f \in \mathbb{R}^{C \times H \times W}$。后续分为两条分支。一条分支是通过一层线性映射得到视频帧

与分割掩码融合后的特征键值对，即$K^* \in \mathbb{R}^{C^k \times H \times W}$和$V^* \in \mathbb{R}^{C^v \times H \times W}$，这个键值对就是后续存入到各自存储库当中的张量。另一条分支是经过多层卷积层与感知机后得到的分流系数$X \in (0,1)$，X决定了该视频帧分流到哪个存储库，有 3 种情况：

a. 当$X \leqslant X_{\min}$时，说明该视频帧特征与分割掩码内的有用特征信息不多，此时将该视频帧分流到初级存储库中。$X_{\min}$代表初级存储库与中级存储库的分界值。

b. 当$X_{\min} < X \leqslant X_{\max}$时，说明该视频帧特征与分割掩码内的有用特征信息较多，此时将该视频帧分流到中级存储库中。$X_{\max}$代表中级存储库与高级存储库的分界值。

c. 当$X_{\max} < X$时，说明该视频帧特征与分割掩码内的有用特征信息很多，此时将该视频帧分流到高级存储库中。

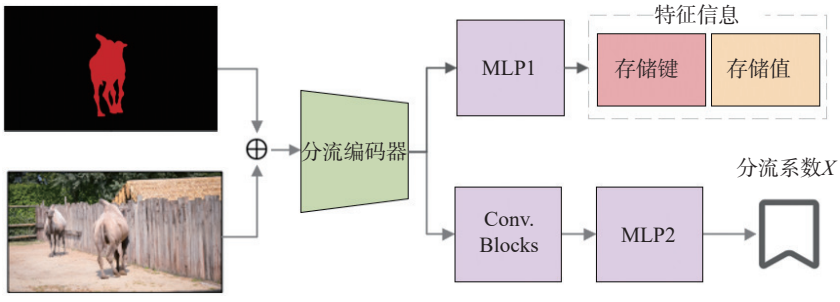

图 2　分流存储模块结构图

Fig.2　Structure diagram of split-flow memory module

当确定好分流路径之后，就会将另一条分支得到的特征键值对放入相应的存储库当中，执行存储更新操作。此外，第一帧视频与掩码将自动存入高级存储库中，最后一帧视频分割后，网络运行直接结束，不会再经过分流存储模块，同时所有存储库都将清零。

2.5　多级存储库

受 Ebbinghaus 遗忘曲线理论启发，多级存储库模拟人脑记忆遗忘规律分为三级存储库：初级、中级和高级存储库。存储库将对分流进来的分割掩码进行独特的存储和更新操作，由此来提升分割精确性并防止内存爆炸。图 3 为整个存储库的存储更新流程图，分割掩码经过分流存储模块进行分割质量评估后，会根据X的大小判定其分割质量。$X \leqslant X_{\min}$代表质量差；$X_{\min} < X \leqslant X_{\max}$代表质量良；$X_{\max} < X$代表质量优。随后，根据这个分割质量分别将分割掩码存储到相应存储库当中，各个存储库更新存储的具体操作有所不同。

2.5.1　初级存储库

初级存储库主要集中存储视频中每帧对象与背景的大致外观特征与位置信息。由于视频相邻帧之间存在时间空间连续性，因此分割帧与初级存储库进行匹配后可以快速锁定跟踪对象的大致外观形状与位置坐标，以便后续与另外两个存储库进一步确认对象语义信息，以此来节省存储空间。

具体来说，初级存储库内只存有一个初级特征键值对$K_p \in \mathbb{R}^{C^k \times H \times W}$和$V_p \in \mathbb{R}^{C^v \times H \times W}$，其包含着视频帧中对象粗略的外观以及位置信息，在初始化时其值为零，随后根据存入该库的视频帧

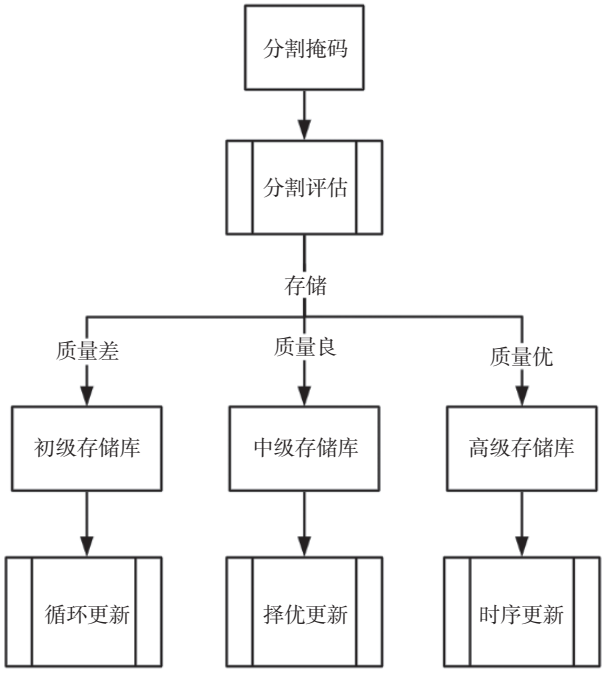

图 3　存储库存储更新流程

Fig.3　Memory and update process of the memory bank

与分割掩码融合后的特征键值对$K^* \in \mathbb{R}^{C^k \times H \times W}$和$V^* \in \mathbb{R}^{C^v \times H \times W}$共同进行更新。总的来说，初级存储库没有存储操作，只有更新操作。具体更新方式是采用循环神经网络中经典的门控循环更新方式[18]，将其用在此处就是使得初级向量F_p在当前帧特征的引导下，持续掌握对象的大致外观以及位置信息。整个更新过程可以概括为

$$K_p' = \mathrm{GRU}\left(K_p, K^*\right) \tag{2}$$

$$V_p' = \mathrm{GRU}\left(V_p, V^*\right) \tag{3}$$

式中：GRU代表门控循环更新操作；K_p和V_p为初级存储库内更新前的特征键值对；K^*和V^*为分流到初级存储库的特征键值对；K_p'和V_p'为初级存储库内更新后的特征键值对。

2.5.2 中级存储库

中级存储库模仿人的常规记忆方式，越正确，记忆越深刻。具体存储过程为：将接收到的$K^* \in \mathbb{R}^{C^k \times H \times W}$和$V^* \in \mathbb{R}^{C^v \times H \times W}$的空间维度上新增一个时间维度$T$，来代表存储的是哪一时刻下的特征，后续存入的特征都会在时间维度上进行连接，各帧之间互不影响。最终总体存储内容则是多帧特征键值对的集合$K_m \in \mathbb{R}^{C^k \times T_m \times H \times W}$和$V_m \in \mathbb{R}^{C^v \times T_m \times H \times W}$，其中$T_m$代表存储库中存储帧数。

当存储库内帧数不断增加到最大存储帧数后，将会触发更新操作。这里的更新标准原则是：保留库中在分流存储模块中分流系数高的视频帧，删除分流系数最低的视频帧。这样做的目的是为了防止低质量的视频帧错误地影响后续视频帧的分割，在整体上尽量降低存储信息中的扰动和噪声对分割结果的影响，具体操作可概括为

$$(K_m', V_m') = (K_m, V_m) \ominus (K_m^x, V_m^x) \tag{4}$$

式中：K_m和V_m为中级存储库内更新前的存储键值对；K_m^x和V_m^x为X最小时的存储特征键值对；$\ominus$代表删去操作；K_m'和V_m'为中级存储库内更新后的存储键值对。

2.5.3 高级存储库

高级存储库存储着所有存储库中质量最高的对象特征信息，其存储方式与中级存储库相同，也是按时间存储由分流存储模块分流来的多个视频帧编码键值对，总体存储内容为$K_s \in \mathbb{R}^{C^k \times T_s \times H \times W}$和$V_s \in \mathbb{R}^{C^v \times T_s \times H \times W}$，$T_s$为存储帧数。与中级存储库更新方式不同，当高级存储库容量达到最大存储值时，会按照遗忘曲线那样，按照存储的时间顺序删去那些存入时间最早的视频帧，整个操作可以概括为

$$(K_s', V_s') = (K_s, V_s) \ominus (K_s^t, V_s^t) \tag{5}$$

式中：K_s、V_s为高级存储库内更新前的特征键值对集合；K_s^t、V_s^t为存入时间最早的存储特征键值对；$\ominus$代表删去操作；K_s'、V_s'为高级存储库内更新后的特征键值对集合。特别注意的是，高级存储库内最先存入的第一帧视频与掩码会被一直保留，因为其代表了整个视频的分割目的与标准，需要以它作为主要指引。

2.6 损失函数

本网络在训练过程中使用交叉熵损失$\mathcal{L}_{ce}$和均方误差损失$\mathcal{L}_{mse}$之和[19]。其中：$\mathcal{L}_{ce}$用来衡量网络分割结果与标签之间像素级的误差；$\mathcal{L}_{mse}$来衡量分割结果中含有的语义信息与标签所含有语义信息之间的差距。完整的损失函数表达式为

$$\mathcal{L}_{ce} = -\frac{1}{P}\sum_{i}^{P} y_i \log p_i \tag{6}$$

$$\mathcal{L}_{mse} = \frac{1}{Q}\sum_{i}^{Q} \left(X_i - \mathrm{maskiou}\left(GT_i, S_i\right)\right)^2 \tag{7}$$

$$\mathcal{L} = \mathcal{L}_{ce} + \mathcal{L}_{mse} \tag{8}$$

式中：P代表视频帧中所有像素点数量；y_i代表i像素点下标签的类别；p_i代表i像素点下分割预测为y_i类别的概率；Q代表视频帧中含有类别的数量；X_i代表由分割存储模块输出针对i这一类别的分流系数；GT_i代表i这一类别所对应的标签图；S_i代表i这一类别所对应的分割图；maskiou代表标签图像与分割图像做交并比操作[20]。

3 实验设置

3.1 数据集

为了验证网络的有效性，本文采用了两种数据集来进行实验探究，这两个数据集都是来自于视频分割领域中使用范围最广、影响最大的密集标注数据集DAVIS。第一个数据集是单个对象标注的视频数据集DAVIS2016[13]，每个视频当中只标记一个对象作为前景，剩下的都属于背景。该数据集共有50个视频序列，每个视频类别都不同，视频内每一帧的尺寸大小为854×480像素，所

有视频按照 3∶2 划分为训练集与测试集。第二个数据集是多个对象标注的视频数据集 DAVIS2017[14]，每个视频当中标记了两个以上的对象作为目标，总共有 150 类的对象。该数据集共有 150 段视频，每帧视频尺寸大小也为854×480像素，视频序列按照 2∶1 划分为训练集与测试集。

3.2　环境设置

本文网络采用 24 GB 的 NVIDIA GeForce RTX 4090 显卡，在 DAVIS2016 和 DAVIS2017 数据集上分别进行训练与测试，在训练 DAVIS2017 数据集时额外采用 Youtube-VOS 数据集[20] 进行拓展训练。本文参照 MiVOS[10]的训练策略，在训练时从视频中按时间顺序随机选择 3 帧作为一组训练样本。随机选择的范围根据训练过程的不同而不同：在训练的中后期阶段，采取更大的选择范围，用于提高网络的鲁棒性；在训练的开始与结束阶段，则设置较小的选择范围，以加速网络收敛及缩小训练推理之间的差距。网络训练使用初始学习率为 2×10^{-5}、权重衰减为 0.1 的 Adam 优化器来优化网络参数[21]。超参数部分设定特征键值对通道空间维度 C^{k} 为 128、C^{v} 为 512，训练轮次为 66，批大小为 8。训练时存储库没有差别，统一按照 STM 网络训练时设置存储库内存储为 3 帧，在测试时初级存储库存储帧数为 1，中级存储库存储帧数为 6，高级存储库存储帧数为 3。

3.3　评价指标

为了衡量网络针对视频的分割性能，本文采用了区域相似度 $\mathcal{J}$ (Jaccard index)，轮廓精确度 $\mathcal{F}$ (contour accuracy)，以及两者的平均衡量标准 $\mathcal{J}\&\mathcal{F}$ $\mathcal{J}\&\mathcal{F}$[15] 来评估网络的分割效果。其中，$\mathcal{J}$ 衡量分割结果与标签整体之间的重合程度，$\mathcal{F}$ 衡量分割结果与标签轮廓之间的吻合程度，$\mathcal{J}\&\mathcal{F}$ 则是兼顾两者的综合评价指标。3 个指标值越接近 1，代表分割效果越好。具体计算式如下：

$$\mathcal{J}=\frac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{P}}+F_{\mathrm{N}}} \tag{9}$$

$$\mathcal{P}=\frac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{P}}} \tag{10}$$

$$\mathcal{R}=\frac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{N}}} \tag{11}$$

$$\mathcal{F}=\frac{2\mathcal{P}\mathcal{R}}{\mathcal{P}+\mathcal{R}} \tag{12}$$

$$\mathcal{J}\&\mathcal{F}=\frac{(\mathcal{J}+\mathcal{F})}{2} \tag{13}$$

式中：T_{P}、F_{N} 分别为预测分割结果中预测正确和错误的对象像素集合；F_{P} 为预测分割结果中预测错误的背景像素集合；$\mathcal{P}$为查准率；$\mathcal{R}$为查全率。

为了衡量网络针对视频的分割速度，本文采用每秒帧数（FPS）来评估网络分割速度。FPS 越大，意味着每秒显示的帧数越多，从而视频中对象的动作就会越流畅，由此也可证明网络分割显示结果的速度更快。

4　实　验

4.1　DAVIS2016 数据集

表 1 展示了本文网络与近年提出的 10 种不同网络在数据集 DAVIS2016 下的分割结果。由于本文整体框架基于时空存储网络的主干，所以基于控制变量的原则，选取的所有网络都采用了时空存储架构。整体实验流程为网络先在 DAVIS2016 的训练集上进行训练，再在其测试集上进行测试。对比网络的分割实验结果均来源于原文，其中分割速度指标都是在同一设备环境下重新复现得到的结果。

由表 1 可以得出，本文网络在 DAVIS2016 数据集上的 4 个评价指标均处于最优：$\mathcal{J}$ 为 89.7%，$\mathcal{F}$ 为 90.7%，$\mathcal{J}\&\mathcal{F}$ 为 90.2%，FPS 为 22.3。且与 10 个网络中最优的指标相比，本文网络在 $\mathcal{J}$、$\mathcal{F}$、$\mathcal{J}\&\mathcal{F}$ 和 FPS 指标上分别提高 1.4%、0.2%、0.8%、5.4%。由此可以看出，本文网络在单对象数据集上有较好的性能，分割速度更快，并且在 $\mathcal{J}$ 上有明显的优势，说明分割的结果与标签更加贴近。图 4 是选取本文网络与次优网络 RMNet[24] 在 DAVIS2016 数据集上同一视频的可视化结果对比。

4.2　DAVIS2017 数据集

表 2 展示了本文网络与 10 种不同网络在数据集 DAVIS2017 下的分割结果。与上述采取的对比原则一致，这里比较的所有网络也都是采用时空存储架构，分割精确性（$\mathcal{J}$、$\mathcal{F}$ 和 $\mathcal{J}\&\mathcal{F}$）指标的具体数值来源于各自网络原文中的数据，分割速度指标（FPS）也都是在同一设备环境下重新复现得到的结果。为了提升最终的测试表现，保证对比实验结果的公平性，在训练时除了使用 DAVIS2017 的训练集以外，所有网络还额外使用 Youtube-VOS[20] 进行补充训练，再在 DAVIS2017 测试集上做最终测试。

由表 2 可以得出，本文网络在 DAVIS2017 数

表 1 不同网络在 DAVIS2016 数据集上的分割实验结果

Tab.1 Segmentation experiment results of different networks on the DAVIS2016 dataset

不同网络	$\mathcal{J}$/%	$\mathcal{F}$/%	$\mathcal{J}\&\mathcal{F}$/%	FPS
SSM[22]	86.2	85.6	85.9	13.5
STM[7]	88.7	89.9	89.3	6.3
FEELVOS[23]	81.1	82.2	81.7	2.2
RMNet[24]	88.9	88.7	88.8	11.9
MiVOS[9]	87.8	90.0	88.9	16.9
KAMNet[25]	87.1	88.1	87.6	8.3
SEVOS-X[26]	87.7	86.7	87.2	8.2
OSOSM[27]	80.0	80.8	80.4	14.8
SKVOS[28]	80.9	83.7	82.3	10.1
TrickVOS[29]	88.7	89.9	89.3	18.7
本文	**89.7**	**90.7**	**90.2**	**22.3**

表 2 不同网络在 DAVIS2017 数据集上的分割实验结果

Tab.2 Segmentation experiment results of different networks on the DAVIS2017 dataset

不同网络	额外训练数据	$\mathcal{J}$/%	$\mathcal{F}$/%	$\mathcal{J}\&\mathcal{F}$/%	FPS
AOST[30]	Youtube-VOS	**81.2**	86.1	83.7	12.5
STM[7]	Youtube-VOS	79.2	84.3	81.8	4.8
FEELVOS[23]	Youtube-VOS	69.1	74.0	71.5	3.9
RMNet[24]	Youtube-VOS	81.0	86.0	83.5	4.4
MiVOS[9]	Youtube-VOS	80.5	85.8	83.1	11.2
KAMNet[25]	Youtube-VOS	80.0	85.2	82.6	8.3
SEVOS-X[26]	Youtube-VOS	76.5	80.8	78.7	13.0
PRM[31]	Youtube-VOS	75.4	79.7	77.6	6.8
MoBox[32]	Youtube-VOS	75.4	81.8	78.6	17.8
MAMP[33]	Youtube-VOS	69.7	68.3	71.2	15.8
本文	Youtube-VOS	80.9	**86.9**	**83.9**	**18.5**

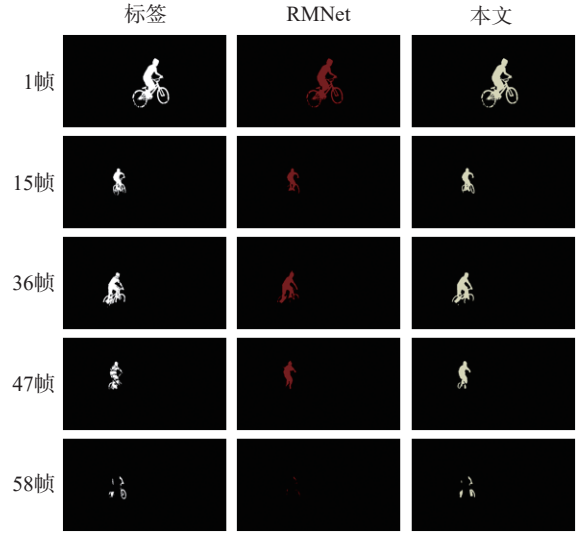

图 4 本文网络与 RMNet 在 DAVIS2016 数据集的可视化结果

Fig.4 Visualization results of the proposed network and RMNet on the DAVIS2016 dataset

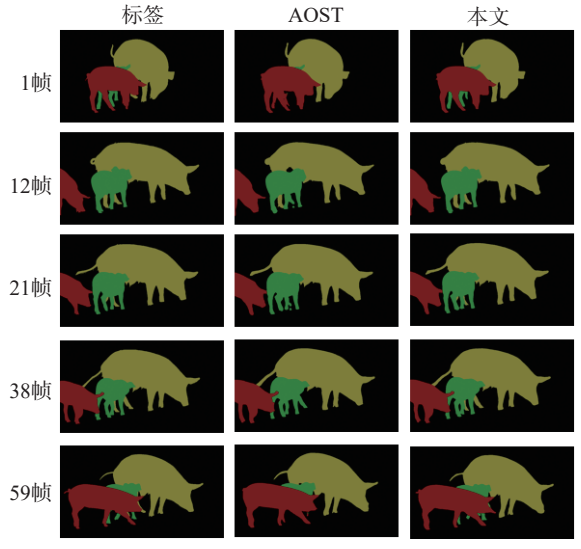

图 5 本文网络与 AOST 在 DAVIS2017 数据集的可视化结果

Fig.5 Visualization results of the proposed network and AOST in the DAVIS2017 dataset

据集上的4项测试结果为：$\mathcal{J}$为80.9%，$\mathcal{F}$为86.9%，$\mathcal{J}\&\mathcal{F}$为83.9%，FPS为18.5。其中$\mathcal{F}$、$\mathcal{J}\&\mathcal{F}$、FPS均为最优。本文与AOST[30]网络相比，虽然$\mathcal{J}$降低0.3%，但$\mathcal{F}$和$\mathcal{J}\&\mathcal{F}$却分别提高0.8%和0.2%。综合几个评价指标，本文网络的分割性能与分割速度处于第一梯队，尤其是通过$\mathcal{F}$可以看出，本网络在处理多对象分割时，对对象细节的处理更为精确。图5是选取的本文网络与次优网络AOST在DAVIS2017数据集上同一视频的可视化结果。

4.3 消融实验

表3与表4分别为本文网络在DAVIS2016数据集上的消融实验结果。表3展示的是原始网络在去掉任何一种存储库后其它部分保持不变的分割结果。其中基于控制变量，整体存储库的存储帧数为10帧，去掉初级存储库后，将初级存储库中原本存储的一帧放入一般存储库中(即传统先入先出的存储库)，去掉中级存储库后将中级存储库中原本存储的6帧放入一般存储库中，去掉高级存储库后将中级存储库中原本存储的3帧放入一般存储库中。当整体网络去掉初级存储库后，$\mathcal{J}$、$\mathcal{F}$和$\mathcal{J}\&\mathcal{F}$分别降低1.1%、0.9%、1.0%。由网络设计可以看出，初级存储库内存放的特征信息

表 3　网络在 DAVIS2016 数据集上的消融实验结果

Tab.3　Ablation results of network on the DAVIS2016 dataset

网络	$\mathcal{J}$/%	$\mathcal{F}$/%	$\mathcal{J}\&\mathcal{F}$/%
本文去掉初级存储库	88.6	89.8	89.2
本文去掉中级存储库	77.4	76.4	76.9
本文去掉高级存储库	89.2	90.4	89.8
本文	**89.7**	**90.7**	**90.2**

表 4　不同分流边界值在 DAVIS2016 数据集上的消融实验结果

Tab.4　Ablation with different split-flow boundary values on the DAVIS2016 dataset

$X_{\min}$、$X_{\max}$	$\mathcal{J}$/%	$\mathcal{F}$/%	$\mathcal{J}\&\mathcal{F}$/%
0.6、0.8	88.7	89.9	89.3
0.7、0.8	88.9	90.1	89.5
0.8、0.9	89.6	90.5	90.1
0.7、0.9	**89.7**	**90.7**	**90.2**

绝大部分都是短时间内的对象背景信息，如果将其删除则就忽略了视频帧信息在时间上的连续性，必然会使分割稳定性降低。当整体网络去掉中级存储库后，可以观察到 3 项指标大幅降低，这说明了中级存储库是网络分割的核心，因为其保留了大量的对象高分辨率特征信息。若没有这一部分信息，在存储匹配时就无法匹配到正确的对象信息。当整体网络去掉高级存储库后，3 项指标小幅下降。高级存储库中存放了高质量的对象特征信息，缺少了这类信息虽然不会明显影响网络的分割性能，但是仍会影响网络在细节上的表现。可见，当同时拥有初级、中级和高级存储库时，本文网络会综合发挥 3 种存储库的优点，达到最佳分割性能，$\mathcal{J}$、$\mathcal{F}$和$\mathcal{J}\&\mathcal{F}$分别为 89.7%、90.7%、90.2%。同时，为了增加可解释性，图 6 给出了同一视频序列中各级存储库中存储的视频帧特征图片。可以看出：初级存储库中存储的是对象的位置大小信息；中级存储库中则存储的是对象的主要信息，具有车大致的外观语义信息；高级存储库中存储的是对象的细节特征，即车边界的纹理与细节。

图 6　在同一视频序列下 3 种存储库内存储特征的可视化结果

Fig.6　Visualization results of features stored in three repositories within the same video sequence

表 4 是网络在不同分流边界值下的分割实验结果。可以看出：当$X_{\min}$取 0.6、$X_{\max}$取 0.8 时，网络的 3 项评价指标为最低；当将$X_{\min}$提升至 0.7、$X_{\min}$保持在 0.8 时，3 项评价指标有了提升，这说明分流系数在 0.6～0.7 之间的特征存在噪声，这部分的特征存入中级存储库后必然影响分割精度；当$X_{\min}$保持在 0.7、$X_{\max}$提升至 0.9 时，3 项指标继续提升，这个操作是扩大了中级存储库的存储范围，提高了高级存储库的存储门槛，这样使得中级存储库内的特征更多，高级存储库内的特征更精确，因此分割精度会提高；继续尝试提升$X_{\min}$至 0.8、$X_{\max}$保持在 0.9 时，分割精度出现下降。综上可以得出，当分流分界值$X_{\min}$取 0.7、$X_{\max}$取 0.9 时，网络性能最好。

5　总　结

针对于时空存储架构网络，为了提高视频分割精度的同时还能合理控制内存的消耗，本文借助心理学的人类的记忆曲线，提出了分流存储网络用于视频分割，设计了初级、中级和高级 3 种类型的存储库，利用分流存储模块来合理分流存储特征信息。在单对象与多对象视频分割数据集上进行测试，并与同类型的网络比较，证明了本网络具有一定优势。未来会继续尝试将网络轻量化，提升其现实实用性。

参考文献：

[1] FAN Z, WANG C L, SUN Z Q. RETRACTED CHAPTER: fast video object segmentation network based on multi-scale attention feature fusion[C]//Proceedings of 2023 Chinese Intelligent Systems Conference. Ningbo: Springer, 2023: 175–176.

[2] LI X X, LOY C C. Video object segmentation with joint re-identification and attention-aware mask propagation[C]// Proceedings of the 15th European Conference on Computer

Vision. Munich: Springer, 2018: 93–110.

[3] PERAZZI F, KHOREVA A, BENENSON R, et al. Learning video object segmentation from static images[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition . Honolulu: IEEE, 2017: 3491–3500.

[4] CHEN Y H, PONT-TUSET J, MONTES A, et al. Blazingly fast video object segmentation with pixel-wise metric learning[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 1189–1198.

[5] HU Y T, HUANG J B, SCHWING A. VideoMatch: Matching based video object segmentation[C]//Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018: 56–73.

[6] YANG Z X, WEI Y C, YANG Y. Associating objects with transformers for video object segmentation[C]// Proceedings of the 35th International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc., 2021: 2491-2502.

[7] OH S W, LEE J Y, XU N, et al. Video object segmentation using space-time memory networks[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE, 2019: 9225–9234.

[8] CHENG H K, TAI Y W, TANG C K. Rethinking space-time networks with improved memory coverage for efficient video object segmentation[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc. , 2021: 11781-11794.

[9] CHENG H K, TAI Y W, TANG C K. Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 5555–5564.

[10] FU L H, ZHAO Y, SUN X W, et al. Video object segmentation based on motion-aware ROI prediction and adaptive reference updating[J]. Expert Systems with Applications, 2021, 167: 114153.

[11] LIU Y, YU Y, YIN F, et al. Learning quality-aware dynamic memory for video object segmentation[C]// Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022: 468–486.

[12] MIAO B, BENNAMOUN M, GAO Y S, et al. Region aware video object segmentation with deep motion modeling[J]. IEEE Transactions on Image Processing, 2024, 33: 2639−2651.

[13] PERAZZI F, PONT-TUSET J, MCWILLIAMS B, et al. A benchmark dataset and evaluation methodology for video object segmentation[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 724–732.

[14] PONT-TUSET J, PERAZZI F, CAELLES S, et al. The 2017 DAVIS challenge on video object segmentation[DB/OL]. [2025-01-12]. https://arxiv.org/abs/1704.00675.

[15] EBBINGHAUS H. Memory: a contribution to experimental psychology[J]. Annals of Neurosciences, 2013, 20(4): 155−156.

[16] HUANG Z J, HUANG L C, GONG Y C, et al. Mask scoring R-CNN[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 6402–6411.

[17] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770–778.

[18] CHO K, VAN MERRIËNBOER B, BAHDANAU D, et al. On the properties of neural machine translation: Encoder-decoder approaches[C]//Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation. Doha: Association for Computational Linguistics, 2014: 103–111.

[19] YANG Z X, WEI Y C, YANG Y. Associating objects with transformers for video object segmentation[C]// Proceedings of the 35th International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc. , 2022.

[20] XU N, YANG L J, FAN Y C, et al. YouTube-VOS: a large-scale video object segmentation benchmark[DB/OL]. [2025-01-12]. https://arxiv.org/abs/1809.03327.

[21] KINGMA D P, BA J. Adam: a method for stochastic optimization[DB/OL]. [2025-01-12]. https://arxiv.org/abs/1412.6980.

[22] ZHU W C, LI J H, LU J W, et al. Separable structure modeling for semi-supervised video object segmentation[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(1): 330−344.

[23] VOIGTLAENDER P, CHAI Y N, SCHROFF F, et al. FEELVOS: fast end-to-end embedding learning for video object segmentation[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 9473–9482.

（下转第 208 页）

[21] 王莹, 高岩. 计及需求响应的随机综合能源系统优化配置 [J]. 上海理工大学学报, 2022, 44(2): 157−165.
[22] 王文烨, 姜飞, 张新鹤, 等. 含规模氢能综合利用的高比例风光多能源系统低碳灵活调度 [J]. 电网技术, 2024, 48(1): 197−206.
[23] ANWAR S, KHAN F, ZHANG Y H, et al. Recent development in electrocatalysts for hydrogen production through water electrolysis[J]. International Journal of Hydrogen Energy, 2021, 46(63): 32284−32317.
[24] GÖTZ M, LEFEBVRE J, MÖRS F, et al. Renewable power-to-gas: a technological and economic review[J]. Renewable Energy, 2016, 85: 1371−1390.
[25] 邓杰, 姜飞, 王文烨, 等. 考虑电热柔性负荷与氢能精细化建模的综合能源系统低碳运行 [J]. 电网技术, 2022, 46(5): 1692−1702.
[26] 张兴平, 张又中. 计及 P2G 和 CCS 的园区级电−热−气综合能源系统多目标优化 [J]. 电力建设, 2020, 41(12): 92−101.
[27] KARIMI M, SHIRZAD M, SILVA J A C, et al. Carbon dioxide separation and capture by adsorption: a review[J]. Environmental Chemistry Letters, 2023, 21(4): 2041−2084.

（编辑：何代华）

（上接第 158 页）

[24] XIE H Z, YAO H X, ZHOU S C, et al. Efficient regional memory network for video object segmentation[C]// Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021: 1286–1295.
[25] WANG G Q, LI L, ZHU M, et al. Kernel based local matching network for video object segmentation[J]. Machine Vision and Applications, 2024, 35(3): 01524.
[26] GAO X Y, LI Z L, SHI H L, et al. Scribble-supervised video object segmentation via scribble enhancement[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2025, 35(4): 2999−3012.
[27] ELAYAPERUMAL D, SAKTHI K S S, JEONG J H, et al. Semi-supervised one-shot learning for video object segmentation in dynamic environments[J]. Multimedia Tools and Applications, 2025, 84(6): 3095–3115.
[28] YANG R L, LI D, HU C H, et al. SKVOS: sketch-based video object segmentation with a large-scale benchmark[J]. Applied Sciences, 2025, 15(4): 1751.
[29] SKARTADOS E, GEORGIADIS K, YUCEL M K, et al. TRICKVOS: a bag of tricks for video object segmentation[C]//Proceedings of 2023 IEEE International Conference on Image Processing. Kuala Lumpur: IEEE, 2023: 2430–2434.
[30] YANG Z X, MIAO J X, WEI Y C, et al. Scalable video object segmentation with identification mechanism[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(9): 6247−6262.
[31] WANG Z, ZHANG W B, LI J H, et al. PRM: a pixel-region-matching approach for fast video object segmentation[C]//Proceedings of the 7th Chinese Conference on Pattern Recognition and Computer Vision. Urumqi: Springer, 2024: 76–91.
[32] LI X M, WANG Q H, LI D Z, et al. MoBox: enhancing video object segmentation with motion-augmented box supervision[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2025, 35(1): 405−417.
[33] MIAO B, BENNAMOUN M, GAO Y S, et al. Self-supervised video object segmentation by motion-aware mask propagation[C]//Proceedings of 2022 IEEE International Conference on Multimedia and Expo. Taipei, Taiwan: IEEE, 2022: 1–6.

（编辑：董 伟）