文章编号　1004-924X(2023)19-2884-14

掩码生成动态调控弱监督视频实例分割

何自芬，徐　林，张印辉*，黄　滢
(昆明理工大学 机电工程学院，云南 昆明 650000)

摘要：针对全监督视频实例分割网络训练数据高度依赖精细掩码标注，时间和人工成本过高，导致智能机器无法快速适应新场景的问题，提出一种端到端的掩码生成动态调控弱监督视频实例分割(Weakly Supervised Video Instance Segmentation，WSVIS)网络。为克服初始掩码预测层通道维度突降导致的实例激活特征丢失问题，构建多级特征融合模块，利用特征复用策略预测初始实例特征并融合相对位置信息生成初始预测掩码。然后，提出动态调控机制在通道和空间维度上建立掩码特征依赖关系，强化初始预测掩码与实例感知信息之间的动态交互。最后，网络设计二元颜色相似性生成伪亲和标签取代精细掩码标注，联合边界框与掩码一致性损失实现仅边界框标注的弱监督视频实例分割。实验结果表明，在 BoxSet 和 YT-VIS 数据集上，WSVIS 网络能达到与全监督网络相近的分割精度和分割效果，同时能够满足实时推理要求，为智能机器快速适应新场景实现实时环境感知和理解提供了理论支撑和算法依据。

关　键　词：智能机器；弱监督视频实例分割；多级特征融合；动态调控；二元颜色相似性
中图分类号：TP394.1　　**文献标识码**：A　　**doi**：10.37188/OPE.20233119.2884

Mask generation dynamically regulates weakly supervised video instance segmentation

HE Zifen，XU Lin，ZHANG Yinhui*，HUANG Ying

(*Faculty of Mechanical and Electrical Engineering*，*Kunming University of Science and Technology*，*Kunming* 650500，*China*)
* *Corresponding author*，*E-mail*：*zyhhzf*1998@163.*com*

Abstract：The training data of fully supervised video instance segmentation networks are highly dependent on accurate mask annotations under high labor and time costs，owing to which intelligent machines are unable to quickly adapt to new scenes. Therefore，a mask generation，dynamically regulated weakly supervised video instance segmentation (WSVIS) network was proposed. First，to overcome the loss of instance activation features caused by the sudden dimension drop of the initial mask prediction layer channel，a multi-level feature fusion module was used to predict the initial instance features through a step-by-step feature reuse strategy and to generate the initial mask by fusing the relative position information. Second，a dynamic regulation mechanism was introduced to establish mask feature dependencies in the channel and spatial dimensions to strengthen the dynamic interaction between the initial predicted mask and instance-aware information. Finally，the network replaces fine mask labeling with the binary color similarity of images，and the bounding box consistency loss and supervised video instance segmentation mask were re-

收稿日期：2023-02-06；**修订日期**：2023-03-13.
基金项目：国家自然科学基金资助项目(No. 62171206，No. 62061022)

placed with bounding box labeling only. Experimental results reveal that on the BoxSet and YT-VIS datasets, the WSVIS network achieves similar segmentation accuracy and segmentation effect as the fully supervised network and can satisfy real-time reasoning, providing theoretical support and an algorithmic basis for intelligent machines to quickly adapt to new scenes to realize real-time environmental perception and understanding.

Key words: intelligent machine; weakly supervised video instance segmentation; multi-level feature fusion; dynamic regulation; binary color similarity

1　引　言

视频实例分割[1-3]任务旨在对时序变换场景中的多目标同时进行检测、分割和跟踪,是当前机器人视觉感知[4-5]、无人驾驶道路场景理解[6-7]、雷达识别与跟踪[8-9]等新一代智能机器前沿交叉领域的一项核心技术,广泛应用于交通、工业、医学和国防等重要领域。

根据分割网络是否需要提供训练集精细掩码标注信息,视频实例分割可分为全监督和弱监督两种训练类型。全监督网络的训练数据需要大量精细掩码标注,但单帧图像中各实例的精细掩码标注所需时间大约为 54～79 s[10],单个实例边界框的标注需要 7 s,单个实例图像级类别标注只需要 1 s[11]。因此,采用弱标注代替精细掩码标注的弱监督视频实例分割能大幅压缩标注成本,非常适用于需要快速备样以迅速适应新场景的智能机器视觉系统。

现有的弱监督视频实例分割方法分为图像级标注和边界框标注两种。Liu 等[12]提出第一个图像级标注弱监督视频实例分割网络,采用像元关系网络[13]结合目标实例的光流运动信息生成伪掩码标签,然后在掩码组成模块中利用视频帧的时间一致性进行实例匹配和分割。但该网络未实现端到端训练,且图像级标注无法使网络精确聚焦实例区域。边界框标注视频实例分割网络[14]沿用实例分割网络 BoxInst[15]的两种掩码监督方式和网络框架,并计算实例外观和运动信息的光流相似性,生成伪标签监督训练。但该网络初始掩码预测分支特征通道维度突降导致实例激活特征丢失,且初始预测掩码特征在通道和空间中信息缺乏关联,与实例感知信息的动态交互能力受到制约。另外,光流法受光线变化影响大,计算量大且耗时长,无法使智能机器分割算法适用实时性要求高的任务。

为了解决上述问题,本文构建了多级特征融合模块,利用特征复用策略相互学习各级特征以生成初始预测掩码,有效克服掩码分支预测一组原型实例时通道维度突降导致预测实例激活特征丢失的问题。设计了动态调控机制在初始预测掩码特征通道和空间中建立依赖关系,使分割网络初始预测掩码与实例感知信息动态交互,更多关注实例区域,进一步提升目标分割精度。提出了边界框与掩码一致性损失监督预测掩码仅在实例边界框范围内生成,并计算输入图像和预测掩码的二元颜色相似性,约束预测掩码更加接近实例区域。

2　相关工作

2.1　边框级弱监督实例分割

为了解决实例分割数据标注量过大的问题,早期方法[10,15,17]仅对训练数据进行边界框标注,并寻找实例区域的像元特征信息优化预测掩码,实现逐像元实例分割。第一个仅用边界框标注的实例分割框架[10](Simple Does It, SDI),将边界框以内的区域都标注为实例,以外的都标注为背景,以此作为训练的先验信息对分割结果进行迭代优化。然而,此方法先利用目标轮廓检测算法[16]生成伪标签,再以全监督的流程训练网络,并不能做到端到端训练。类不可知联合学习弱监督网络[17]认为在边界框的像元区域内,每一行或每一列像元至少有一个像元点属于实例,将这些行和列称作正包,边界框外的则称作负包,然后将正负包集合作为训练数据,设计多实例学习方案嵌入到 Mask R-CNN[18]全监督流程中完成

端到端训练。由于保留了两阶段实例分割先检测后分割的范式,依靠大量原图候选框映射到特征图相应区域的操作获得实例掩码,该方法的训练和推理速度较慢。BoxInst[15]在一阶段实例分割网络(Conditional Convolutions For Instance Seg-mentation,CondInst)[19]的基础上提出投影损失函数和成对相似性损失函数,利用实例边界框和掩码在X轴和Y轴上具有相同的投影、相邻像元大概率具有相同实例标签两个先验信息,约束网络生成最终的实例预测掩码。

2.2 视频实例分割

MaskTrack R-CNN[1]是第一个提出视频实例分割任务的两阶段网络,它创建了第一个大规模视频实例分割数据集YT-VIS,并在Mask R-CNN[18]基础上增加新的跟踪分支,利用外部记忆追踪不同帧之间的实例,为边框回归头生成候选框并分配实例标签。但该网络需利用RoIAlign操作生成大量候选建议框,网络训练速度和推理速度缓慢,难以应用在实时性要求较高的任务中。受到一阶段实时实例分割网络YOLACT[20]的启发,Sipmask[21]引入轻量级空间保存模块,在每个边界框中生成单独空间系数,解决了预测边界框中空间信息不足的问题,更好地描绘空间中相邻实例;另外,利用掩码对齐权重损失和利用可变形卷积进行特征与回归框位置对齐,将掩码预测和实例检测相关联。Li等[22]发现单阶段实例分割网络存在卷积特征既不与预测框对齐也不与真实边界框对齐的问题,这降低了掩码预测的空间感知能力,因此,为锚框和真实边界框设计了特征校准策略以获得更精确的空间特征;同时,利用视频固有特性构建了一个时间融合模块聚合当前帧与参考帧的掩码特征,提高了相邻帧之间的时间相关性。然而,基于视频帧的时间建模方法[21-22]仅停留在相邻帧之间。由于时间维度承载的丰富场景信息对网络分割、定位和类别预测有重要作用,在线交叉学习网络[23](Crossover learning for fast online video instance segmentation,CrossVIS)在实例分割网络CondInst[19]的基础上建立交叉学习模块,用视频序列中丰富的上下文信息,将同一目标在非相邻两帧中的位置信息跨帧融合到每帧的预测掩码,引入位置信息增强实例表示,同时削弱背景和与实例无关的信息。CrossVIS[23]利用全卷积网络作为初始掩码预测头,摆脱了两阶段实例分割方法训练和推理速度慢的缺点,但存在初始预测掩码实例激活特征丢失严重,无法有效融合实例感知信息的问题。

3 WSVIS网络结构

3.1 生成调控部分

本文提出了弱监督视频实例分割网络(Weak Supervised Video Instance Segmentation,WSVIS),其整体结构如图1所示,分为生成调控部分和交叉预测部分。

WSVIS网络在同一视频序列中随机抽取两帧($F_t\in\boldsymbol{R}^{B\times3\times H\times W}$,$F_t'\in\boldsymbol{R}^{B\times3\times H\times W}$,$B$表示Batch size)图像作为网络生成调控部分输入。生成调控部分由初始特征提取网络、检测头、生成融合模块和动态调控机制组成。F_t,F_t'经过参数共享的ResNet50骨干网络[24]和特征金字塔网络(Feature Pyramid Network,FPN)[25])提取图像初始特征。检测头包含分类头(Class Head)、边框回归头(Box Head)和实例感知头(Inst-Aware Head)。FPN输出特征分别经过检测头和生成融合模块。

检测头根据FPN各层$[\boldsymbol{P}_3,\boldsymbol{P}_4,\boldsymbol{P}_5,\boldsymbol{P}_6,\boldsymbol{P}_7]$输出特征的每个像元预测与之相关联的类别,并直接回归锚点到边框的距离完成实例边界框预测,同时实例感知头与分类头结构相同用于预测掩码的实例感知信息$\boldsymbol{A}_t$,$\boldsymbol{A}_t'$。生成融合模块的输入是FPN高分辨率层的输出特征$\boldsymbol{P}_3\in\boldsymbol{R}^{B\times128\times H\times W}$,$\boldsymbol{P}_3'\in\boldsymbol{R}^{B\times128\times H\times W}$,由于实例预测所需的相对位置信息来自FPN,为了构建两者的联系,将初始预测特征$\boldsymbol{F}_{\text{mask}}$,$\boldsymbol{F}'_{\text{mask}}$尺寸重置为$N\times8\times(H\times W)$($N$取决于边框回归头保留的预测实例数量),并与它到FPN各特征层之间的相对位置信息$\boldsymbol{L}_t\in\boldsymbol{R}^{B\times2\times(H\times W)}$,$\boldsymbol{L}_t'\in\boldsymbol{R}^{B\times2\times(H\times W)}$在通道维度上进行拼接,然后重置为原始特征图尺寸,生成初始预测掩码$\boldsymbol{I}_t=\in\boldsymbol{R}^{B\times2\times(H\times W)}$,$\boldsymbol{I}_t'=\in\boldsymbol{R}^{B\times2\times(H\times W)}$。动态调控机制生成通道和空间权重加权,强化初始预测掩码在交叉预测部分中与实例感知信息的动态交互能力。

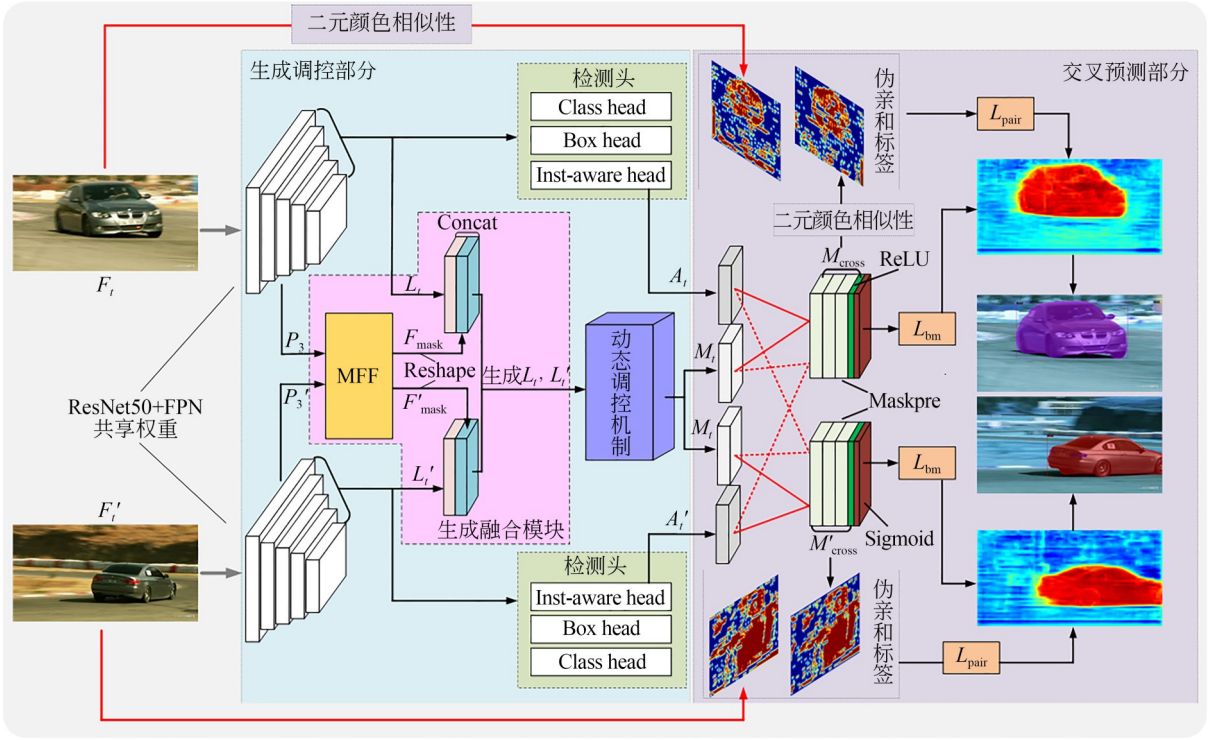


图 1　WSVIS 网络结构

Fig. 1　Structure of WSVIS network

3.1.1　多级特征融合模块

CrossVIS[23]原始掩码预测层如图 2(左下角)所示，包含 4 个卷积核大小为 3×3，通道数为 128 的串行卷积层，最后一层通道数从 128 减少到 8，代表生成 8 组实例掩码。但实例通道突降会导致初始预测掩码激活特征丢失，因此受人体姿态估计网络[26]启发，本文提出多级特征融合(Multi-level Feature Fusion，MFF)模块。为了增强高分辨率信息，通过将高分辨率和低分辨率卷积并行连接，达到多尺度特征交互的目的。不同的是，MFF 模块包含 5 级通道数不同的串行卷积层，各级卷积层之间采用特征复用机制进行特征融合。如图 2 所示，用 L_i^j 代表 MFF 的 5 级通道变换层，$i\in[1,2,3,4,5]$ 代表层级代号，$j\in[1,2,3]$ 代表 3 个串行卷积从左到右的顺序代号。MFF 模块输入为 FPN 输出特征 $P_3\in\mathbf{R}^{B\times 128\times 48\times 80}$，$P_3'\in\mathbf{R}^{B\times 128\times 48\times 80}$。特征经过 L_1^1 层后并行输入到 L_1^2 和 L_2^1 层，L_2^1 层输出通道为 64，然后经过二倍通道上采样与 L_1^2 层在通道维度对齐并相加作为 L_1^3 层的输入；L_1^2 层输出通过 1/2 倍通道下采样与 L_2^2 层在通道维度对齐并作为 L_2^2 层的输入；然后 L_2^1 层输出作为 L_3^1 层输入，L_1^3 层输出、L_2^2 层输出和 L_3^1 层输出之和作为 L_2^3 层的输入。同理，其他相邻级通道特征也通过这种高级特征与低级特征交互融合的方式建立联系，能有效缓解激活特征丢失的问题。MFF 模块各级特征融合计算流程可表示为：

$$out_{L_1^3}=L_1^3\left(out_{L_1^2}+out_{L_2^1}\right),\tag{1}$$

$$out_{L_2^3}=L_2^3\left(out_{L_1^3}+out_{L_2^2}+out_{L_3^1}\right),\tag{2}$$

$$out_{L_3^3}=L_3^3\left(out_{L_2^3}+out_{L_3^2}+out_{L_4^1}\right),\tag{3}$$

$$out_{L_4^3}=L_4^3\left(out_{L_3^3}+out_{L_4^2}+out_{L_5^1}\right),\tag{4}$$

$$out_{L_5^3}=L_5^3\left(out_{L_4^3}+out_{L_5^2}\right),\tag{5}$$

式中：$out_{L_i^j}$ 表示各卷积层输出，$L_i^j\left(out_{L_i^j}\right)$ 表示输出经过 L_i^j 层。

经过 MFF 模块后，输入特征尺寸并不会改变，因此最终输出特征与原始掩码预测层的输出维度一致，可表示为 $F_{\text{mask}}=\text{ReLU}\left(out_{L_5^3}\right)\in\mathbf{R}^{B\times 8\times H\times W}$，$F'_{\text{mask}}=\text{ReLU}\left(out_{L_5^3}\right)\in\mathbf{R}^{B\times 8\times H\times W}$。为了使网络从初始预测到交叉预测保持动态联系，其中 L_1^1 层是为每个样本学习一个卷积核参数

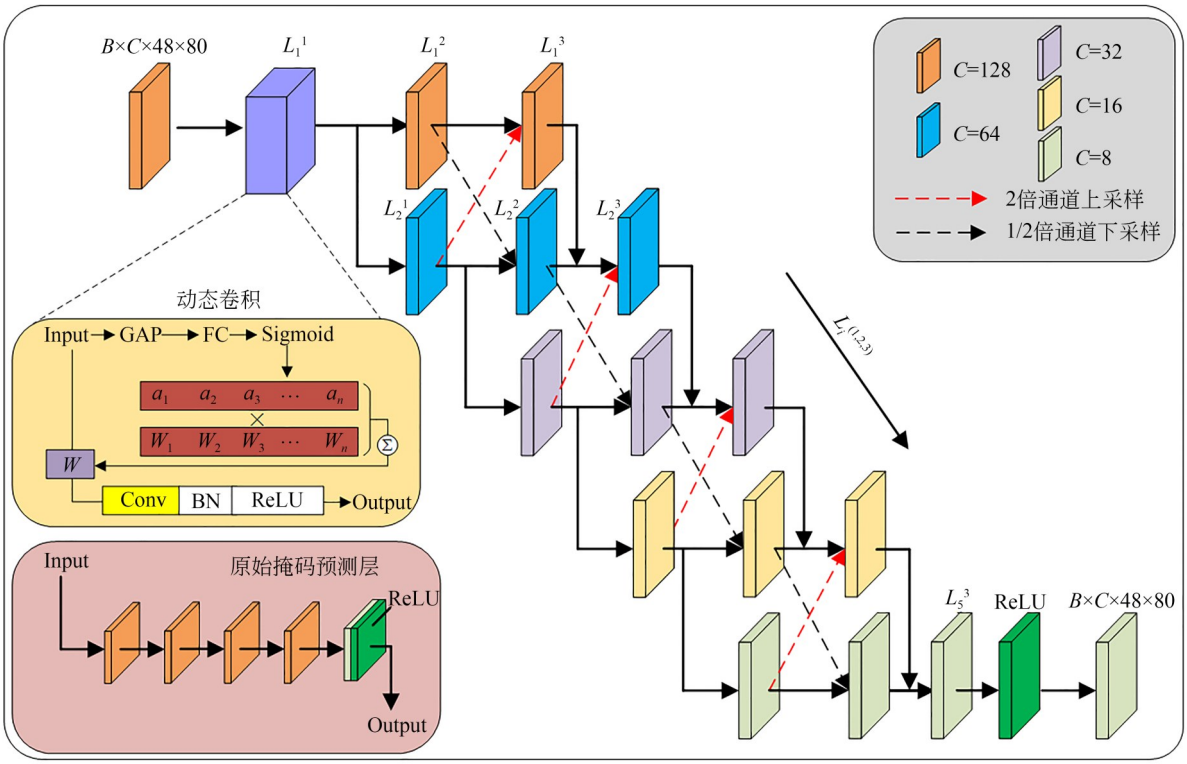


图 2　MFF 模块结构

Fig. 2　Structure of MFF module

的动态卷积 CondConv[27]，动态卷积等同于多个标准卷积的线性组合，每个卷积中根据特征输入决定卷积核权重，再对这些卷积核加权求和得到一个适合该输入的动态卷积核权重因子。动态卷积计算过程可表示为：

$$\begin{cases} output = \mathrm{ReLU}\left(\mathbf{Input}\sum_{1}^{n} a_i w_i\right) \\ a_i = \mathrm{Sigmoid}\left(FC\left(GAP\left(\mathbf{Input}\right)\right)\right) \end{cases}. \tag{6}$$

输入特征 $\mathbf{Input}(F_{\mathrm{mask}}, F'_{\mathrm{mask}})$ 经过全局平均池化(Global Average Pooling，GAP)层得到由所有通道的特征图像的像素平均值组成的池化特征向量，然后特征向量经过全连接层(Fully Connected，FC)后通过 Sigmoid 函数生成一系列与初始卷积权重 ω_i 一一对应的权重 a_i，训练过程中 ω_i 同时学习和训练，因此对于每个预测实例都会生成特定权重。将 ω_i 和 a_i 分别相乘再相加后与输入 **Input** 做卷积运算，最后使用 ReLU 函数增加各层特征之间的非线性关系。

3.1.2　动态调控机制

为加强初始预测掩码与动态实例感知信息之间的联系，本文提出动态调控机制寻找空间和通道上对于区分前背景更有效的信息，提高网络对实例区域的关注。如图 3 所示，动态调控机制分为通道调控机制和空间调控机制两个部分，通道调控可以区分网络预测特征不同通道的重要程度，采取全局最大池化和全局平均池化对动态生成的 N 维实例特征图进行压缩，以此建立各通道特征之间的关系。空间调控在 N 维实例特征图上寻找对区分前景和背景起重要作用的像元区域。

通道调控机制的输入为 $F_{\mathrm{in}} \in \mathbf{R}^{1\times N\times 48\times 80}$。首先输入特征按通道进行全局最大池化(Global Max Pooling，GMP)和全局平均池化(Global Average Pooling，GAP)得到两组尺寸为 $N\times 1\times 1$ 的池化特征，然后将两组特征向量分别送入 FC 后对应元素相加，最后利用 Sigmoid 函数计算得到通道权重系数 $\boldsymbol{\omega}_{\mathrm{c}} \in \mathbf{R}^{1\times N\times 1\times 1}$。

$$\boldsymbol{\omega}_{\mathrm{c}} = S\left(\theta_1\left(GAP\left(F_{\mathrm{in}}\right)\right) + \theta_2\left(GMP\left(F_{\mathrm{in}}\right)\right)\right), \tag{7}$$

其中：θ_1，θ_2 表示全连接层权重系数，S 表示 Sigmoid 函数。

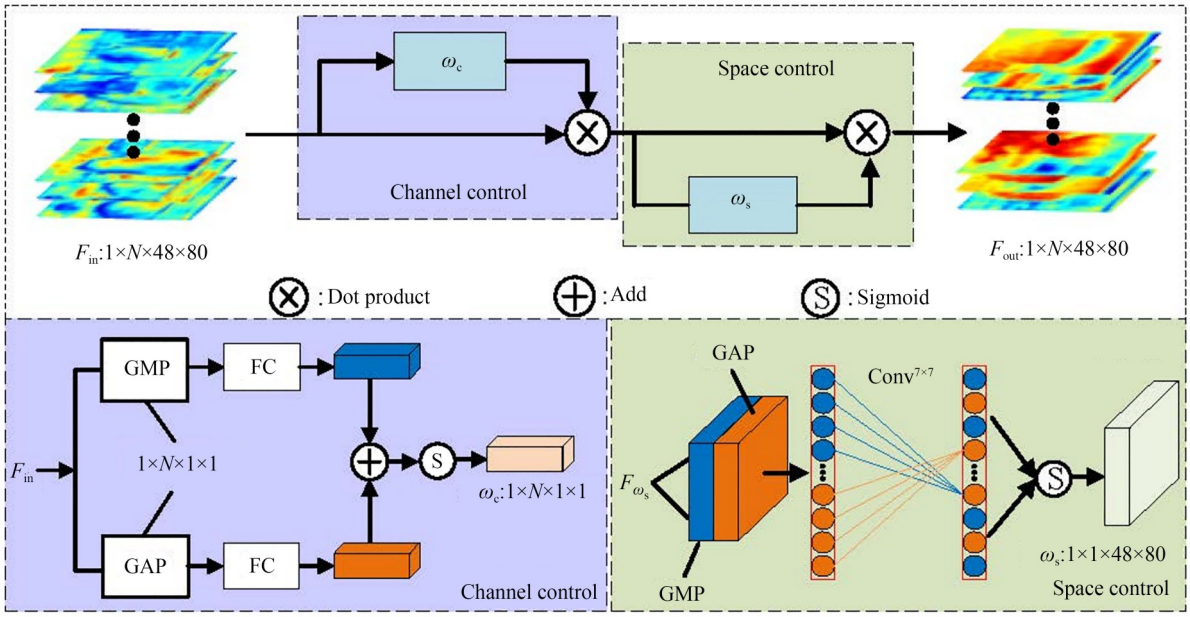


图 3　动态调控机制

Fig. 3　Dynamic control mechanism

$\boldsymbol{\omega}_c$ 为输入特征 F_{in} 每一个实例通道赋予不同的权重，得到 N 维通道调控加权的特征输出 $F_{\omega_c}\in \mathrm{R}^{1\times N\times 48\times 80}$。通道调控加权后的特征 F_{ω_c} 输入空间调控机制，首先将 F_{ω_c} 在空间上进行 GMP 和 GAP 并拼接得到尺寸为 2×48×80 的池化特征，然后使用一维卷积将两个池化特征图在通道维度进行信息交互，最终经过 Sigmoid 生成空间注意力权重系数 $\boldsymbol{\omega}_s\in \mathrm{R}^{1\times N\times 48\times 80}$。

$$\boldsymbol{\omega}_s = S\left(Conv^{7\times 7}\left(Cat\left(GAP(F_c), GMP(F_c)\right)\right)\right), \tag{8}$$

其中：$Conv^{7\times 7}$ 表示尺寸为 7×7 的卷积核，Cat 表示拼接。然后，$\boldsymbol{\omega}_s$ 与 F_{ω_c} 进行逐像素点积运算得到最终的动态调控机制输出 $\boldsymbol{M}_t\in \mathrm{R}^{1\times N\times 48\times 80}$，$\boldsymbol{M}_t'\in \mathrm{R}^{1\times N\times 48\times 80}$。

3.2　交叉预测部分

WSVIS 网络的交叉预测部分包含交叉学习方案和损失计算。F_t, F_t' 两帧经过生成调控部分后得到动态预测掩码 $\boldsymbol{M}_t, \boldsymbol{M}_t'$，以及来自实例感知头的实例感知信息 $\boldsymbol{A}_t, \boldsymbol{A}_t'$。对于视频实例分割，同一实例可能出现在视频两帧中的不同位置，因此两帧中同一实例的外观信息和位置信息可以相互引导。交叉学习方案可表示为：

$$\boldsymbol{M}_{\text{cross}} = Maskpre(\boldsymbol{M}_t, \boldsymbol{A}_t) + Maskpre(\boldsymbol{M}_t, \boldsymbol{A}_t'), \tag{9}$$

$$\boldsymbol{M}_{\text{cross}}' = Maskpre(\boldsymbol{M}_t', \boldsymbol{A}_t') + Maskpre(\boldsymbol{M}_t', \boldsymbol{A}_t), \tag{10}$$

其中：$Maskpre$ 表示最终掩码预测层，$\boldsymbol{M}_t, \boldsymbol{M}_t'$ 是 $Maskpre$ 的输入特征，$\boldsymbol{A}_t, \boldsymbol{A}_t'$ 提供 $Maskpre$ 所需的卷积核权重和偏置，$\boldsymbol{M}_{\text{cross}}$ 和 $\boldsymbol{M}_{\text{cross}}'$ 表示最终实例预测掩码。$Maskpre$ 第四层使用 ReLU 作为激活函数，最后一层使用 Sigmoid 得到最终实例预测概率。

3.2.1　边界框与掩码一致性损失

本文网络训练时不具备精细掩码标注信息，而已知的先验信息是：在平面上，实例边界框与实例掩码水平方向的长和垂直方向的宽具有一致性。因此，本文利用真实标注实例边界框约束预测掩码在边界框内生成。边界框与掩码一致性如图 4 所示。

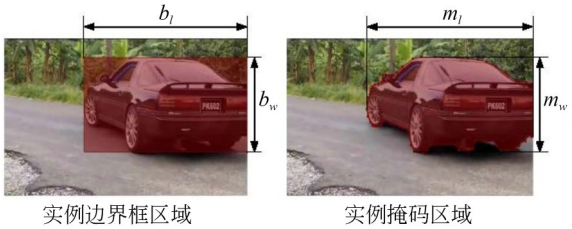


图 4　实例边界框与掩码区域一致性示意图

Fig. 4　Consistency images of instance bounding box and mask area

假设真实标注边界框水平方向长为b_l、垂直方向宽为b_w，预测掩码水平方向长为m_l、垂直方向宽为m_w，利用相似度度量函数 Dice Loss[28]衡量二者偏差，则损失函数表示为：

$$L_{\mathrm{bm}}=Dice(b_l,m_l)+Dice(b_w,m_w)=2-\sum_{z\in[l,w]}\frac{2|b_z\cap m_z|}{|b_z|+|m_z|},\tag{11}$$

其中：L_{bm}应用于所有n个预测实例概率图和真实边界框的偏差计算，最终损失值为是L_{bm}/n。

3.2.2 伪亲和标签生成与二元颜色相似性损失

这里以输入图像和预测掩码二元像元颜色相似性生成的伪亲和标签计算损失优化预测掩码。首先，将图片下采样到与交叉预测部分预测掩码特征图的相同尺寸，为了更加直观地判断颜色的相似性，将下采样后的图像数据从 RGB 空间转换到更加接近人类视觉感知的 Lab 空间，然后计算输入图像和预测掩码的二元颜色相似性，以此生成伪亲和标签，最后计算二者损失优化预测掩码。计算流程如图 5 所示。

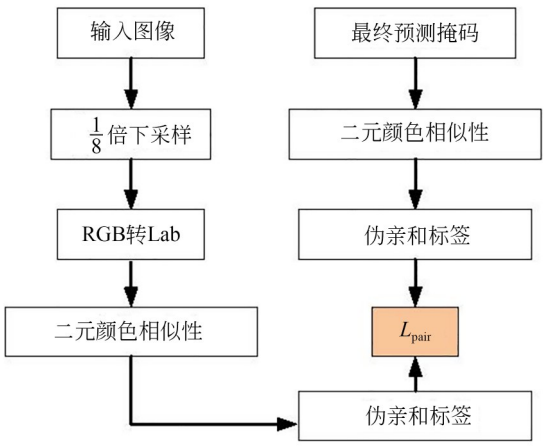


图 5　二元颜色相似性损失计算流程

Fig. 5　Flow chart for binary color similarity loss calculation

针对下采样后图像分辨率降低导致原有像素特征丢失的问题，本文受到空洞卷积[29]的启发，将邻域像元点采样间隔设置为 2，在保证像元采样感受野的同时能获得最优分割性能。

像元采样如图 6 所示，以图像上颜色相近的两像元点大概率具有相同实例标签作为先验，中心像元与领域 8 个像元分别计算相似度，然后利用颜色相似度阈值过滤相似度低的一对像元，计算二元颜色相似度生成交叉预测部分所示的伪亲和标签。

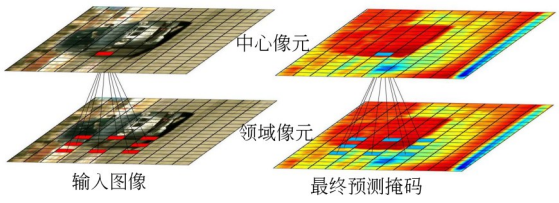


图 6　像元采样示意图

Fig. 6　Pixel sampling diagram

假设p_1表示中心像元，p_2表示中心像元邻域任一像元，那么在 Lab 空间中p_1和p_2的颜色相似度可表示为：

$$S(p_1,p_2)=\exp\left(-\frac{\left\|Lab_{p_1}-Lab_{p_2}\right\|}{2}\right),\tag{12}$$

其中$\left\|Lab_{p_1}-Lab_{p_2}\right\|$表示$p_1$和$p_2$在 Lab 空间中的感知相似度。

本文相似度阈值λ是预先设置的超参数，延续弱监督实例分割网络 BoxInst[15]的设置，λ设置为 0.1。当$S\geqslant\lambda$时，两像元点相似且具有相同的实例标签；当$S<\lambda$时，两像元点所属标签不相同。对于预测掩码，上述标签学习问题转化为像元对是否同属于实例和背景的二分类概率问题，在某个实例的预测掩码$\boldsymbol{M}_{\mathrm{cross}}$，$\boldsymbol{M}'_{\mathrm{cross}}$中某对像元$g$同为实例的概率表示为$P(g=1)$，其余为背景的概率表示为$P(g=0)$。二分类概率计算像元对采样方式与伪标签相似度计算方式一致。

边界框与掩码一致性损失函数约束掩码仅在预测框内生成，最后采用二元交叉熵（BCE）损失函数优化预测掩码与输入图像伪亲和标签之间的差距。二元颜色相似性损失函数可写为：

$$L_{\mathrm{pair}}=-\frac{1}{N}\sum_{g\in G_{\mathrm{b}}}\tau\log P\left(g=1\right),\tag{13}$$

其中：N表示经颜色相似度阈值过滤后保留的相似像元对数量，G_{b}表示边界框内的像元对的集合。当$S\geqslant\lambda$时，$\tau=1$；当$S<\lambda$时，$\tau=0$。

3.2.3 总损失函数

本文联合边界框与掩码一致性损失和二元颜色相似性损失，实现仅边界框标注的 WSVIS 网络，端到端进行检测、分割和跟踪，每个训练样

本的多任务损失函数表示为:

$$L_{\text{total}}=L_{\text{cls}}+L_{\text{box}}+L_{\text{bm}}+L_{\text{pair}}+L_{\text{id}},\quad(14)$$

其中:L_{cls},L_{box}和L_{id}与CrossVIS[23]一致,分别表示分类损失、边界框损失和跟踪损失,L_{bm}和L_{pair}为本文边界框与掩码一致性损失和二元颜色相似性损失。

4　实验结果及分析

4.1　实验数据

本文数据集BoxSet由382段视频数据组成,其中329段视频用于训练,53段视频用于测试。训练集仅使用边界框和类别标注,测试集的标注方法与全监督数据集一致并用于网络模型分析和调优。

为了验证WSVIS网络的稳定性,在大型视频实例分割数据集YT-VIS[1] 2019上与现有弱监督视频实例分割网络FlowIRN[12]和FlowSimi[14]进行对比。YT-VIS[1]数据集共包含2 883段视频,其中训练集视频2 238段,测试集338段,训练时不使用精细掩码标注信息。

4.2　训练细节

实验硬件环境如下:基于Ubuntu18.04操作系统,CPU型号为Intel(R)Core(TM)i5-11400,GPU型号为NVIDIA GeForce RTX2080Ti。软件环境如下:深度学习框架为PyTorch1.8.1,python版本为3.7,采用CUDA10.2运算平台和CUDNN8.3.2工具包加速模型训练。超参数设置如表1所示。

表1　训练参数设置

Tab.1　Training parameter settings

超参数及学习策略	值
图像输入尺寸	360×640
Batch Size	4
动量	0.9
迭代步数	Boxset:23 000,YT-VIS:190 000
学习率	0.005

4.3　视频实例分割评价指标

本文的网络评估指标与MaskTrack R-CNN[1]一致,其中平均精度(Average Precision,AP)是交并比(Intersection Over Union,IoU)从50%到95%之间步长为5%的10个阈值的总平均精度,AP_{50},AP_{75}分别代表IoU阈值为50%和75%时的平均精度;AR_1,AR_{10}则表示限定条件下的平均召回率。与静止图像实例分割不同的是,视频中每一个实例在整个视频序列中都包含了一系列掩码,因此视频实例分割评估时IoU的计算不仅需要在空间域上进行,也要在时间域上进行,则IoU定义为每一个实例的预测掩码与真实标签的交集与并集的商,即[1]:

$$IoU(i,j)=\frac{\sum_{t=1}^{T}\left|m_t^i\cap\tilde{m}_t^i\right|}{\sum_{t=1}^{T}\left|m_t^i\cup\tilde{m}_t^i\right|}.\quad(15)$$

在测试阶段,计算预测值与真实值的IoU是否大于给定阈值,若IoU大于阈值表示为TP,小于阈值则表示为FP。则总平均分割精度定义为:

$$AP=\frac{1}{|C|}\sum_c\left(\frac{1}{|thr|}\sum\frac{TP(t)}{TP(t)+FP(t)}\right),\quad(16)$$

其中:C代表数据集中的类别数量,c为当前类别,thr阈值数,t为当前阈值。

4.4　实验结果分析

实验对比了WSVIS与其他视频实例分割网络在BoxSet和YT-VIS[1]上的各项平均精度,结果如表2所示。3种全监督视频实例分割网络均为单阶段视频实例分割网络,在BoxSet上AP分别达到了36.2%,37.5%,39.9%,WSVIS网络的AP超过了Sipmask[21],与STMask[22]相近,仅比CrossVIS[23]低2.4%,AP_{50}优于其他3个全监督网络,达到64.7%。由于FlowIRN[12]和FlowSimi[14]网络并未公开,因此在YT-VIS[1]上对3种弱监督网络进行对比。图像级弱监督视频实例分割网络FlowIRN[12]的AP为10.5%,边界框弱监督视频实例分割网络FlowSimi[14]的AP为29.0%,WSVIS网络的AP为30.1%,相比FlowIRN[12]提升19.6%,比最先进的FlowSimi[14]网络高1.1%。

在YT-VIS[1]上,3种弱监督视频实例分割网络精度分割均低于全监督视频实例分割网络。由于全监督视频实例分割网络采用精细掩码标注,网络利用优质的监督信息在迭代训练中能学习到更优的参数,因此在测试中通常能

表 2　不同视频实例分割网络的实验结果对比

Tab. 2　Comparison of experiment results for different video instance segmentation networks　(%)

Methods	BoxSet					YT-VIS[1]				
	AP	AP_{50}	AP_{75}	AR_1	AR_{10}	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
Sipmask[21]	36.2	61.8	34.2	38.5	41.4	32.5	53.0	33.3	33.5	38.9
STMask[22]	37.5	61.0	38.6	39.3	42.9	33.5	52.1	36.9	31.1	39.2
CrossVIS[23]	39.9	64.1	41.5	41.0	46.7	34.8	54.6	37.9	34.0	39.0
FlowIRN[12]	—	—	—	—	—	10.5	27.2	6.2	12.3	13.6
FlowSimi[14]	—	—	—	—	—	29.0	50.2	29.4	—	—
WSVIS	37.5	64.7	40.0	37.8	42.5	30.1	50.5	31.2	31.1	37.0

获得更好的分割精度和效果。而不论是图像级标注或是边界框标注，弱监督视频实例分割网络都缺乏精细的真实掩码标注，仅有接近真实标注的伪亲和标签监督网络训练，网络不可避免地会引入错误监督信息，导致分割精度降低。WSVIS网络在弱监督学习范式下超过了目前最先进的FlowSimi[14]，分割精度和分割效果接近于全监督视频实例分割网络，同时由于不需要精细掩码标注，与全监督网络相比，数据标注量大大减少，加快了各类场景弱监督网络的开发进程。

图7展示了WSVIS网络与CrossVIS[23]网络的3段视频分割结果。从分割效果可以看到，WSVIS网络具有与全监督网络相近的分割性能。在3段视频中，均可以看到CrossVIS[23]网络将一些稀疏背景区域或实例部分区域错误分割，而WSVIS却不存在此类问题，这得益于WSVIS网络边界框与掩码一致性损失的良好性能，使实例预测掩码总是在边界框中生成。总体来看，WSVIS网络在训练数据仅包含边界框标注的情况下获得了准确的分割结果，具备良好的视频场景感知和分析能力。

图 7　视频分割结果

Fig. 7　Video segmented results

4.5　消融实验

为证明 MFF 模块、动态调控机制、边界框与掩码一致性损失 L_{bm} 和二元颜色相似性损失 L_{pair} 对 WSVIS 网络的有效性,将各模块和损失在 BoxSet 数据集上进行消融实验。

4.5.1　L_{bm} 和 L_{pair} 实验

L_{bm} 和 L_{pair} 消融实验结果如表 3 所示。可以看出,当仅使用边界框与掩码一致性损失 L_{bm} 监督网络掩码分支训练时,总平均分割精度为 23.7%;仅使用二元颜色相似性损失 L_{pair} 监督网络掩码分支训练时,总平均分割精度为 22.5%;而当 WSVIS 网络同时使用二者监督训练,总平均分割精度提升 10% 左右。由此可见,边界框与掩码一致性损失 L_{bm} 和二元颜色相似性损失 L_{pair} 共同使用能显著提升 WSVIS 网络的弱监督分割性能。

表 3　L_{bm} 和 L_{pair} 的有效性验证

Tab. 3　Effectiveness verification for L_{bm} and L_{pair} (%)

L_{bm}	L_{pair}	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
√	—	23.7	55.4	20.7	25.6	27.9
—	√	22.5	54.1	18.7	24.3	26.7
√	√	32.9	60.3	32.8	35.6	39.2

4.5.2　MFF 模块实验

MFF 模块的消融实验结果如表 4 所示。当 WSVIS 网络仅利用 L_{bm} 和 L_{pair} 两项损失进行监督时,在本文数据集上的总平均分割精度为 32.9%。加入 MFF 模块后,网络总平均分割精度从 32.9% 提高到 36.3%。

表 4　MFF 模块的有效性验证

Tab. 4　Effectiveness verification of MFF module (%)

MFF	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
—	32.9	60.3	32.8	35.6	39.2
√	36.3	64.7	35.1	37.7	42.4

为了验证 MFF 模块 5 级特征通道逐级融合对提升网络分割精度的有效性,对 MFF 模块采用不同层级通道融合与使用标准卷积和动态卷积作为 L_1^1 层进行对比实验,结果如表 5 所示。当 L_1^1 层使用标准卷积时,MFF 模块层级从 3 级到 5 级的总平均分割精度增加 2.2%;而使用动态卷积 CondConv[27] 时,MFF 模块层级从 3 级到 5 级精度增加 2.0%。CondConv 为 5 级时,MFF 模块提升网络的总平均分割精度最大。

表 5　MFF 模块不同级、L_1^1 层不同的卷积实验结果

Tab. 5　Convolution experiment results of different MFF module levels and different L_1^1 layers (%)

卷积层级	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
Conv 3	32.6	59.8	32.5	34.4	39.4
Conv 4	33.7	62.5	36.2	35.6	39.2
Conv 5	34.8	63.4	36.8	36.0	40.1
CondConv[27] 3	34.3	62.5	33.8	37.5	40.7
CondConv[27] 4	35.6	69.5	37.1	37.0	41.0
CondConv[27] 5	36.3	64.7	35.1	37.7	42.4

4.5.3　动态调控机制实验

WSVIS 网络在加入动态调控机制后的总平均精度从 36.3% 提高到 37.5%,如表 6 所示。如

图 8 所示，特征经过动态调控机制后，实例区域得到了更多关注，从而更有利于初始预测掩码与动态感知权重和偏置之间进行交互。

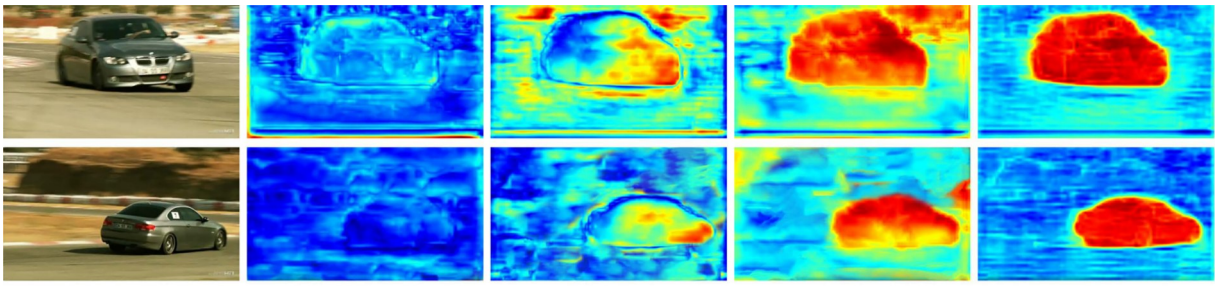
仅边界框与掩码一致性损失、二元颜色相似性损失共同监督时，WSVIS 网络的总平均分割精度为 32.9%，通过构建 MFF 模块和动态调控机制，WSVIS 网络的分割精度提升至 37.5%。

由于 MFF 模块采用多级特征复用策略进行通道特征融合，在一定程度上解决了由于掩码预测分支特征通道突然下降导致实例激活特征严重丢失的问题。原始预测掩码经过特征激活后前景和背景区域的差别并不明显，实例激活特征丢失严重，而 MFF 模块增强了图像中属于实例的目标区域特征信息。动态调控机制在空间和通道维度上进一步增强了网络对于实例和背景的感知能力。

表 6　动态调控机制的有效性验证

Tab. 6　Effectiveness verification of dynamic regulation mechanism　(%)

动态调控机制	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
—	36.3	64.7	35.1	37.7	42.4
√	37.5	64.7	40.0	37.8	42.5



(a) 输入图像 (a) Input images　(b) 原始预测掩码 (b) Original predicted mask　(c) MFF模块输出 (c) Output of MFF module　(d) 动态调控机制输出 (d) Output of dynamic regulation machenism　(e) 最终预测掩码 (e) Final predicted mask

图 8　热力图对比

Fig. 8　Heat map comparison

对比图 8(d)和 8(b)可看出，MFF 模块和动态调控机制不仅有效增强了网络对实例区域的关注，一些稀疏的背景区域也得到了增强。然而，这些稀疏背景区域会在交叉预测部分受到交叉学习方案、边界框与掩码一致性损失和二元颜色相似性损失的共同抑制，使预测掩码更接近真实实例区域，如图 8(e)所示。

4.5.4　交叉学习和帧间隔实验

WSVIS 网络设计了交叉预测部分，将同一实例在不同帧的外观和实例信息关联起来实现帧交叉学习，在加入 MFF 模块和动态调控机制的 WSVIS 网络基础上验证两帧交叉学习的有效性以及不同采样帧间隔对网络的影响，实验结果如表 7 所示。采样帧间隔最小为 1，最大为 35。

由表 7 可知，当采样帧间隔趋向于 1 时，网络总平均分割精度逐渐下降，这是因为当采样间隔较小时，采样两帧图像信息非常相似，这会使得网络退化为静止图像分割的形式，交叉学习不能获得最大的增益。随着采样帧间隔变大，采样的两帧中实例位置和场景上下文信息有明显的位置和外观差异，交叉学习方式能更好地利用实例外观信息和位置信息，将两个采样帧中的同一实例进行关联；当采样帧间隔为 20 时，WSVIS 网络获得了最好的总平均分割精度，为 37.5%；而当采样帧间隔大于 20 时，同一实例在两帧中的位置和外观差异较大，网络的总平均分割精度逐渐下降，因此两帧交叉学习获得了相反的引导。此

表 7　交叉帧间隔对比实验结果

Tab. 7　Results of cross frame interval comparison experiment　(%)

Cross	1	5	10	15	20	25	30	35
—	34.3	34.6	34.9	34.3	34.6	34.6	34.0	34.1
√	35.1	35.4	36.2	36.3	37.5	36.2	36.2	35.8

外,当选择最优采样帧间隔20时,不使用交叉学习方案的网络总平均分割精度仅为34.6%。由此可知,交叉学习方案可以有效利用不同采样帧的实例信息,当采样帧间隔为20时,交叉学习的作用最大。

4.6　计算复杂度分析

WSVIS网络除去在数据集上有良好表现外,在模型复杂度和推理速度上也具有一定优势。由于弱监督视频实例分割网络FlowIRN[12]和FlowSimi[14]未公开其原始算法,因此对比了WSVIS网络和全监督视频实例分割网络Sipmask[21],STMask[22]和CrossVIS[23]的模型参数量、模型计算量和推理速度,结果如表8所示。分析实验在YT-VIS[1]测试集上进行,推理速度不计算数据加载和测试后处理,仅统计所有视频推理完成后的平均推理速度。

实验表明,相比Sipmask[21]网络,WSVIS网络模型的参数量降低74%、计算量降低29.4%;相比STMask[22]网络,WSVIS网络模型的参数量降低76.7%、计算量降低69.3%。相比CrossVIS[23]网络,MFF模块、动态调控机制、边界框与掩码一致性损失和二元颜色相似性损失使WSVIS网络仅有0.22M参数量、6.61G计算量的增长;另外,WSVIS网络在弱监督学习模式下仍然具有实时(FPS>30)推理的表现,推理速度降低了5.4 frame/s,仍高于Sipmask[21]网络和STMask[22]网络4~6 frame/s,能够满足智能机器快速适应新场景实现实时环境感知和理解的需求。

表8　不同网络的模型复杂度和推理速度对比

Tab.8　Comparison of model complexity and inference speed of different networks

Network	Size	Param(M)	FLOPs(G)	FPS
Sipmask[21]	384×640	32.75	226.62	30.0
STMask[22]	384×640	36.79	521.46	28.6
CrossVIS[23]	360×640	8.37	153.33	39.8
WSVIS	360×640	8.59	159.94	34.4

5　结　论

本文提出了WSVIS网络,首先构建多级特征融合模块MFF生成初始预测特征,有效缓解了网络通道下降带来的激活特征丢失;其次,通过动态调控机制在初始预测掩码特征通道和空间中建立依赖关系,有效加强网络对实例目标区域的敏感度;最后,利用边界框与掩码一致性损失约束预测掩码在边界框内生成,同时计算输入图像和预测掩码的二元颜色相似性,约束预测掩码更加接近实例区域。实验结果及热力图可视化表明,MFF模块解决了初始预测特征丢失问题;动态调控机制能有效增强网络对实例区域的关注;视频帧采样间隔为20时,交叉学习方式能够最大程度提升网络的分割精度。WSVIS网络在BoxSet数据集上的总平均分割精度为37.5%,达到与全监督网络相近的分割精度和分割效果;在YT-VIS[1]测试集上各项分割精度均优于先进的FlowSimi[14]网络,分割速度为34.4 frame/s,具备视频场景实时环境感知和分析理解的能力。

参考文献:

[1] YANG L J, FAN Y C, XU N. Video instance segmentation[C]. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 27 - *November* 2, 2019, *Seoul*, *Korea* (*South*). IEEE, 2020: 5187-5196.

[2] 毛琳, 任凤至, 杨大伟, 等. 实例特征深度链式学习全景分割网络[J]. 光学 精密工程, 2020, 28(12):2665-2673.
MAO L, REN F ZH, YANG D W, *et al*. INF-Net: Deep instance feature chain learning network for panoptic segmentation [J]. *Opt. Precision Eng.*, 2020, 28(12):2665-2673. (in Chinese)

[3] 梁新宇, 林洗坤, 权冀川, 等. 基于深度学习的图像实例分割技术研究进展[J]. 电子学报, 2020, 48(12): 2476-2486.
LIANG X Y, LIN X K, QUAN J CH, *et al*. Research on the progress of image instance segmentation based on deep learning[J]. *Acta Electronica Sinica*, 2020, 48(12): 2476-2486. (in Chinese)

[4] 曹天扬,蔡浩原,方东明,等. 结合图像内容匹配的机器人视觉导航定位与全局地图构建系统[J]. 光学 精密工程,2017,25(8):2221-2232.
CAO T Y, CAI H Y, FANG D M, *et al*. Robot vision system for keyframe global map establishment

and robot localization based on graphic content matching [J]. *Opt. Precision Eng.*, 2017, 25(8): 2221-2232. (in Chinese)

[5] 钱夔, 宋爱国. 一种改进型机器人仿生认知神经网络[J]. 电子学报, 2015, 43(6):1084-1089.
QIAN K, SONG A G. An improved bionic cognitive neural network for robot[J]. *Acta Electronica Sinica*, 2015, 43(6):1084-1089. (in Chinese)

[6] 伍锡如, 薛其威. 基于激光雷达的无人驾驶系统三维车辆检测[J]. 光学 精密工程, 2022, 30(4): 489-497.
WU X R, XUE Q W. 3D vehicle detection for unmanned driving systerm based on lidar[J]. *Opt. Precision Eng.*, 2022, 30(4): 489-497. (in Chinese)

[7] 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.
QIN F W, SHEN X Y, PENG Y, *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)

[8] 李淑慧, 邓志红, 冯肖雪, 等. 强杂波背景下基于变分贝叶斯推理的机载雷达目标跟踪算法[J]. 电子学报, 2022, 50(5): 1089-1097.
LI SH H, DENG ZH H, FENG X X, *et al*. Variational Bayesian Inference? Based airborne radar target tracking algorithm in strong clutter[J]. *Acta Electronica Sinica*, 2022, 50(5): 1089-1097. (in Chinese)

[9] 王树亮, 毕大平, 阮怀林, 等. 基于信息熵准则的认知雷达机动目标跟踪算法[J]. 电子学报, 2019, 47(6): 1277-1284.
WANG SH L, BI D P, RUAN H L, *et al*. Cognitive radar maneuvering target tracking algorithm based on information entropy criterion[J]. *Acta Electronica Sinica*, 2019, 47(6): 1277-1284. (in Chinese)

[10] KHOREVA A, BENENSON R, HOSANG J, *et al*. Simple does it: weakly supervised instance and semantic segmentation[C]. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). *July* 21-26, 2017, *Honolulu*, *HI*, *USA*. IEEE, 2017: 1665-1674.

[11] 任冬伟, 王旗龙, 魏云超, 等. 视觉弱监督学习研究进展[J]. 中国图象图形学报, 2022, 27(6): 1768-1798.
REN D W, WANG Q L, WEI Y CH, *et al*. Progress in weakly supervised learning for visual understanding[J]. *Journal of Image and Graphics*, 2022, 27(6): 1768-1798. (in Chinese)

[12] LIU Q, RAMANATHAN V, MAHAJAN D, *et al*. Weakly supervised instance segmentation for videos with temporal mask consistency[C]. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 20-25, 2021, *Nashville*, *TN*, *USA*. IEEE, 2021: 13963-13973.

[13] AHN J, CHO S, KWAK S. Weakly supervised learning of instance segmentation with inter-pixel relations [C]. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 15-20, 2019, *Long Beach*, *CA*, *USA*. IEEE, 2020: 2204-2213.

[14] IKEDA J, MORI J. Weakly supervised instance segmentation using motion information via optical flow[J/OL]. *arXiv preprint* arXiv:2202.13006.

[15] TIAN Z, SHEN C H, WANG X L, *et al*. BoxInst: high-performance instance segmentation with box annotations[C]. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 20-25, 2021, *Nashville*, *TN*, *USA*. IEEE, 2021: 5439-5448.

[16] MANINIS K K, PONT-TUSET J, ARBELÁEZ P, *et al*. Convolutional oriented boundaries: from image segmentation to high-level tasks[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, 40(4): 819-833.

[17] HSU CC, HSU KJ, TSAI CC, *et al*. Weakly supervised instance segmentation using the bounding box tightness prior[J]. *Advances in Neural Information Processing Systems*, 2019, 32: 6586-6597.

[18] HE K M, GKIOXARI G, DOLLÁR P, *et al*. Mask R-CNN[C]. 2017 *IEEE International Conference on Computer Vision* (*ICCV*). *October* 22-29, 2017, *Venice*, *Italy*. IEEE, 2017: 2980-2988.

[19] TIAN Z, SHEN C H, CHEN H. *Conditional Convolutions for Instance Segmentation*[M]. Computer Vision - ECCV 2020. Cham: Springer International Publishing, 2020: 282-298.

[20] BOLYA D, ZHOU C, XIAO F Y, *et al*. YOLACT: real-time instance segmentation[C]. 2019 *IEEE/CVF International Conference on*

Computer Vision (*ICCV*). *October* 27 - *November* 2, 2019, *Seoul*, *Korea* (*South*). IEEE, 2020: 9156-9165.

[21] CAO J L, ANWER R M, CHOLAKKAL H, *et al*. *SipMask*: *Spatial Information Preservation for Fast Image and Video Instance Segmentation*[M]. Computer Vision - ECCV 2020. Cham: Springer International Publishing, 2020: 1-18.

[22] LI M H, LI S, LI L D, *et al*. Spatial feature calibration and temporal fusion for effective one-stage video instance segmentation [C]. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 20-25, 2021, *Nashville*, *TN*, *USA*. IEEE, 2021: 11210-11219.

[23] YANG S S, FANG Y X, WANG X G, *et al*. Crossover learning for fast online video instance segmentation[C]. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 10-17, 2021, *Montreal*, *QC*, *Canada*. IEEE, 2022: 8023-8032.

[24] HE K M, ZHANG X Y, REN S Q, *et al*. Deep residual learning for image recognition[C]. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 27-30, 2016, *Las Vegas*, *NV*, *USA*. IEEE, 2016: 770-778.

[25] LIN T Y, DOLLÁR P, GIRSHICK R, *et al*. Feature pyramid networks for object detection[C]. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). *July* 21-26, 2017, *Honolulu*, *HI*, *USA*. IEEE, 2017: 936-944.

[26] SUN K, XIAO B, LIU D, *et al*. Deep high-resolution representation learning for human pose estimation[C]. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 15-20, 2019, *Long Beach*, *CA*, *USA*. IEEE, 2020: 5686-5696.

[27] YANG B, BENDER G, LE Q V, *et al*. *CondConv*: *Conditionally Parameterized Convolutions for Efficient Inference* [EB/OL]. 2019: *arXiv*: 1904.04971. https://arxiv.org/abs/1904.04971.

[28] MILLETARI F, NAVAB N, AHMADI S A. V-net: fully convolutional neural networks for volumetric medical image segmentation [C]. 2016 *Fourth International Conference on 3D Vision* (*3DV*). *October* 25-28, 2016, *Stanford*, *CA*, *USA*. IEEE, 2016: 565-571.

[29] YU F, KOLTUN V. *Multi-scale Context Aggregation by Dilated Convolutions* [EB/OL]. 2015: *arXiv*: 1511.07122. https://arxiv.org/abs/1511.07122.

作者简介:

何自芬(1976—),女,山西阳泉人,博士,副教授,硕士生导师,2000年、2005年于西安理工大学分别获得学士和硕士学位,2013年于昆明理工大学获得博士学位,主要从事图像处理和机器视觉等方面的研究。E-mail: zyhhzf1998@163.com