文章编号　1004-924X(2023)18-2736-16

尺度自适应生成调控的弱监督视频实例分割

张印辉，海维琪，何自芬*，黄　滢，陈东东
(昆明理工大学 机电工程学院，云南 昆明 650000)

摘要：视频实例分割是车辆辅助驾驶多目标感知和场景理解的一项关键技术。针对弱监督视频实例分割仅使用边界框对网络进行训练严重制约交通场景大尺度动态范围目标分割精度的问题，本文提出尺度自适应生成调控弱监督视频实例分割网络(Scale Adaptive Generation Regulation weakly supervised video instance segmentation network，SAGRNet)。首先，设计一种多尺度特征映射贡献度动态自适应调控模块，通过动态调整不同尺度特征映射信息贡献度取代原有的线性加权以强化对目标局部位置和整体轮廓的聚焦能力，解决了车辆、行人等目标由于成像距离远近造成的尺度动态范围过大问题；其次，构建目标实例多细粒度空间信息聚合生成调控模块，通过聚合基于不同空洞率提取的多细粒度空间信息生成权重参数以调控各尺度特征，实现了细化实例边界和增强跨通道信息交互掩码特征映射表征能力，有效弥补了实例边缘信息匮乏导致边缘轮廓分割 mask 连续性缺失问题。最后，为缓解边界框标签监督信息弱化，引入正交损失和颜色相似性损失缩小模型预测 mask 与真实边界框偏差并计算逐像素点对间标签属性归类模糊问题。Youtube-VIS2019 提取的交通场景数据集实验结果表明，SAGRNet 相较于弱监督基准网络平均分割精度提升 5.1% 达到 38.1%，为实现多目标感知和实例级场景理解提供了有效算法依据。

关　键　词：辅助驾驶；弱监督；视频实例分割；自适应生成调控；细粒度
中图分类号：TP391.4　　**文献标识码**：A　　**doi**：10.37188/OPE.20233118.2736

Weakly supervised video instance segmentation with scale adaptive generation regulation

ZHANG Yinhui, HAI Weiqi, HE Zifen*, HUANG Ying, CHEN Dongdong

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

Abstract: Video instance segmentation is critical in multi-target perception and scene understanding in assisted driving. However, as weakly supervised video instance segmentation is often applied to bounding box annotations for network training, the segmentation accuracies of targets with large-scale dynamic ranges in traffic scenes are severely restricted. To address this issue, we propose a scale adaptive generation regulation weakly supervised video instance segmentation network (SAGRNet). First, a multi-scale feature mapping contribution dynamic adaptive control module is proposed to replace the original linear weighting. This enables placing the focus on the local position and global contour of the target by dynamically adjusting the contribution of different scale feature mapping information, which solves the problem of

收稿日期：2022-12-14；**修订日期**：2023-01-18.
基金项目：国家自然科学基金资助项目(No. 62061022，No. 62171206，No. 61761024)

large-scale dynamic ranges caused by changes in the imaging distance between vehicles and pedestrians. Second, a target instance multi-fine-grained spatial information aggregation generation control module is constructed to regulate the feature maps of each scale using weight parameters, which are obtained by aggregating multi-fine-grained spatial information extracted based on different dilations. This module refines the instance boundary and improves the representation of cross-channel mask interaction information, effectively compensating for the lack of edge contour segmentation mask continuity caused by limited instance edge information. Finally, to alleviate the weak supervision derived from bounding box level annotations, orthogonal and color similarity losses are introduced to reduce the deviation between the model prediction mask and real bounding box and to address the pixel-wise label attribute classification ambiguity problem. Experimental results on a traffic scene dataset extracted from Youtube-VIS2019 indicate that the SAGRNet improves the mean accuracy by 5.1% to 38.1% compared with the weakly supervised baseline. These results prove that our method provides an effective theoretical basis for multi-target perception and instance level scene understanding.

Key words: assisted driving; weakly supervised; video instance segmentation; adaptive generation regulation; fine grain

1　引　言

近年来，辅助驾驶领域中对车辆周围复杂环境多目标感知和场景理解技术成为研究的重点方向。现阶段，针对辅助驾驶车辆环境感知技术包括激光雷达(Lidar)[1-4]、全球导航卫星系统(Global Navigation Satellite System, GNSS)[5-7]和惯性测量单元(Inertial Measurement Unit, IMU)[8]以及计算机视觉卷积神经网络(Convolutional Neural Networks, CNN)[9-11]等技术。其中全球导航卫星系统和惯性测量单元往往只用于定位，而激光雷达和计算机视觉既可用于定位也可用于识别，但激光雷达成本高，且无法给出跟踪目标的类别和尺寸信息。基于计算机视觉的实例分割技术具备对复杂环境进行实时感知的强大能力且成本较低，被广泛应用于辅助驾驶视觉识别系统，实现辅助驾驶车辆对周围障碍物的精准识别。

实例分割技术可分为图像实例分割和视频实例分割两个方向。其中图像实例分割[12-15]针对单帧图像进行实例分类、检测和分割；视频实例分割[16-20]以图像实例分割为基础，对不同帧中同一实例进行跨帧关联追踪，以时间序列形式获得各实例的分割掩膜及检测结果。目前基于深度学习的视频实例分割方法主要包括基于全监督和弱监督学习两类模型，在全监督学习中，MaskTrack R-CNN[16]在图像实例分割Mask R-CNN[12]头部基础上添加跟踪分支关联不同帧之间的目标实例，最先实现帧级实例同时检测、跟踪和分割，并在提出的Youtube-VIS2019[16]数据集验证模型有效性，但结合单帧图像分割和传统方法进行跟踪关联，忽略了关键的时间信息，导致网络分割精度低。Maskprop[21]在MaskTrack R-CNN基础上添加mask传播分支，将中间帧目标实例mask传播到视频其他帧以提升mask生成和关联质量，在使用较少标签数据进行预训练的情况下，在Youtube-VIS数据集上分割精度达到46.6%，比MaskTrack R-CNN高16.3%，但由于MaskProp采用离线学习方式导致模型占用内存大且分割时效性差。为克服检测到跟踪多阶段分割范式处理速度较慢且不利于发挥视频时序连续性的优势，STEm-Seg[22]采用三维卷积和高斯混合来改善时空嵌入特征表示，提升挖掘视频整体的空间和时序信息提取能力，且以较快的速度解决视频实例分割的问题。然而，该方法获得的实例嵌入特征仅包含像素级高斯后验概率估计，缺乏视频数据目标实例时变的高级上下文抽象和统计特征，极大限制了STEm-Seg算法的分割鲁棒性。CrossVIS[23]提出一种新的交叉学习方案，基于当前帧中的实例特征，以像素方式定位其他视频帧中相同实例，有效利用视频中固有的上下文信息来增强跨视频帧的实例表示，同时

削弱背景和无关实例信息，显著提高了网络分割精度。但上述基于全监督学习的视频实例分割技术对目标真实值像素级标注具有很强的依赖性，因此冗长的视频序列样本导致大量的人工精细化标注成本剧增。

目前基于边界框的弱监督实例分割方法仅将实例边界框坐标及类别信息作为真实值进行网络约束学习，极大节省人工标注成本[24]。Hsu等人[25]提出BBTP(Bounding Box Tightness Prior)方法将弱监督实例分割问题视为多示例学习任务，以真实边界框为界限区分前景与背景，结合MIL loss和DenseCRF对伪mask进一步优化，然而仅以边界框约束像素归类可能导致边界框内mask质量下降。Wang[26]等人基于BoxCaseg预训练模型生成伪标签，并通过边界框标签约束伪标签边界，最后用于代替Mask R-CNN实例分割模型训练过程中人工标注值，但受限于指定实例分割预训练模型难以适配现有视频实例分割网络。Tian等人[27]提出BoxInst方法，通过构建投影损失和颜色相似性损失函数替换CondInst[28]中mask分割损失，显著缩小了弱监督和全监督实例分割之间的性能差距。弱监督视频实例分割仅使用边界框对网络进行训练严重制约了交通场景大尺度动态范围目标分割精度的问题。

为实现辅助驾驶车辆对周围复杂环境的多尺度动态目标精准感知和场景理解，以及节省训练所需人工精细化标注成本，本文设计了一种基于尺度自适应生成调控弱监督视频实例分割算法(Scale Adaptive Generation Regulation，SAGRNet)。首先针对全监督网络对目标真实值像素级标注具有很强的依赖性，冗长的视频序列样本导致大量的人工精细化标注成本剧增的问题，本文引入正交损失函数和颜色相似性损失函数代替全监督CrossVIS网络实例mask分割损失，仅利用边界框标签对初始预测mask进行联合训练，实现了基于边界框的弱监督视频实例分割算法Box-CrossVIS，并以此作为本文的基准网络。其次，在特征金字塔(Feature Pyramid Networks，FPN)[29]自上而下融合路径嵌入多尺度特征映射贡献度动态自适应调控模块，通过动态调整不同尺度特征映射信息贡献度以强化对目标局部位置和整体轮廓的聚焦能力，增强网络对前景目标多尺度变化情况下的识别感知能力。最后，在mask预测分支前添加目标实例多细粒度空间信息聚合生成调控模块，采用通道注意力机制[30]聚合基于不同空洞率提取的多细粒度空间信息生成权重参数以调控各尺度特征，有效细化实例边缘轮廓并增强跨通道信息交互掩码特征映射表征能力。SAGRNet算法在Youtube-VIS2019提取的交通场景数据集上进行综合实验，平均分割精度达到38.1%，且在2080Ti上最高分割速度可达36 FPS，为车辆辅助驾驶实现实时多目标感知和实例级场景理解提供了有效算法依据。

2　本文算法

2.1　SAGRNet网络结构

SAGRNet网络包含特征提取和后处理两个阶段，其网络结构如图1所示。在特征提取阶段，首先将视频t和$t+\delta$帧图像输入到ResNet50提取语义特征$\{C1,C2,C3,C4,C5\}$。其次，利用FPN增强上下文多尺度目标特征信息提取能力，通过自顶向下和横向连接融合方式为低层特征引入丰富的高层语义信息，得到特征图$\{P3,P4,P5\}$。为解决由于距离变化造成交通场景中车辆和行人等障碍物目标尺度动态范围扩大问题，本文在FPN融合路径嵌入自适应调控模块，动态调整不同尺度特征映射信息贡献度以强化对不同尺度目标的感知识别能力。最后，将FPN输出特征经3×3卷积操作，并将得到的最高层特征进行2倍下采样操作，最终得到特征图$\{F3,F4,F5,F6,F7\}$。在特征后处理阶段，首先将特征提取阶段输出的各尺度特征分别输入Mask Branch和Controller Head分支。其中Mask Branch用于生成实例mask预测的F_{mask}特征，并结合相对位置信息*Coord*拼接生成实例mask特征图$\tilde{F}_{x,y}(t)$，$\tilde{F}_{x',y'}(t+\delta)$；Controller Head用于生成实例特定动态滤波器$\theta_{x,y}(t)$，$\theta_{x',y'}(t+\delta)$，并预测该位置实例动态条件卷积MaskHead的参数。针对实例边缘轮廓分割不完整、质量粗糙的问题，本文在mask预测分支前添加生成调控模块以细化实例边界并实现特征跨通道信息交互增强掩码特征映射表征能力，提高网络对实例边缘轮廓的分割

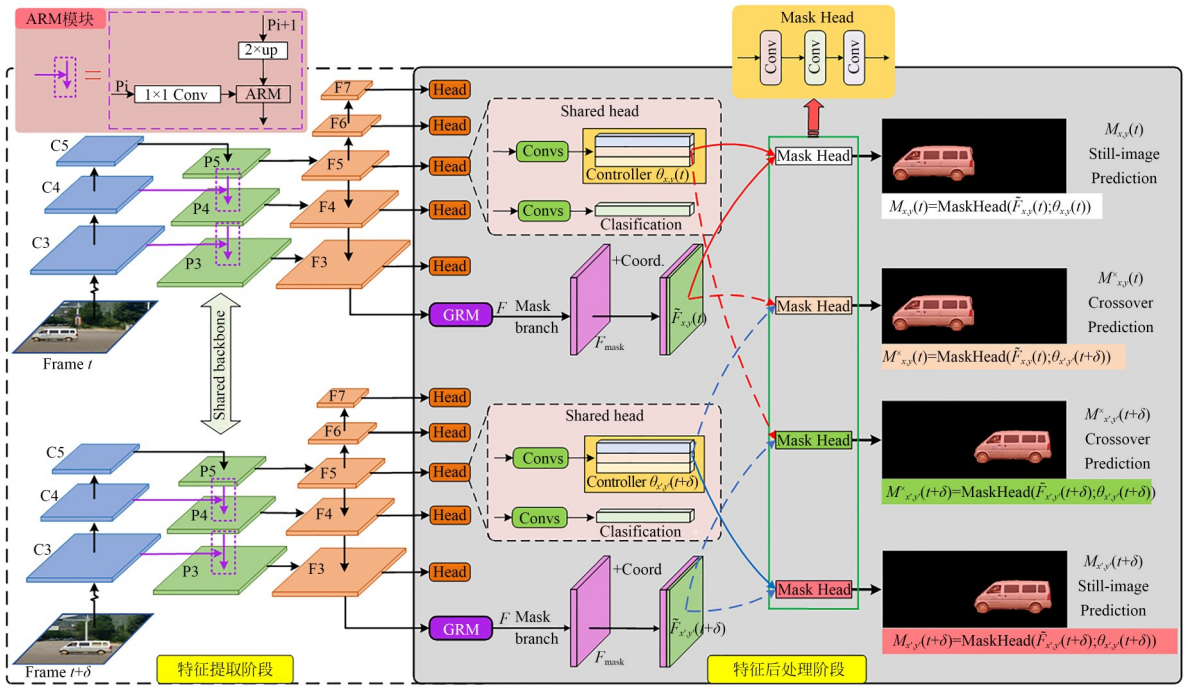


图 1　SAGRNet 网络结构

Fig. 1　SAGRNet Network Structure

能力。然后，将生成的 mask 特征图和动态滤波器交叉输入条件卷积 MaskHead 生成静态和交叉动态 mask。t 帧和 $t+\delta$ 帧图像生成的静态实例 mask 可表示为：

$$M_{x,y}(t)=MaskHead\left(\tilde{F}_{x,y}(t);\ \theta_{x,y}(t)\right),\quad(1)$$

$$M_{x',y'}(t+\delta)=MaskHead\left(\tilde{F}_{x',y'}(t+\delta);\ \theta_{x',y'}(t+\delta)\right),\quad(2)$$

t 帧和 $t+\delta$ 帧图像生成的交叉动态实例 mask 可表示为：

$$M^{\times}_{x,y}(t)=MaskHead\left(\tilde{F}_{x,y}(t);\ \theta_{x',y'}(t+\delta)\right),\quad(3)$$

$$M^{\times}_{x',y'}(t+\delta)=MaskHead\left(\tilde{F}_{x',y'}(t+\delta);\ \theta_{x,y}(t)\right),\quad(4)$$

其中，Maskhead 由三个卷积层组成，以实例为条件动态生成卷积参数。最后，引入正交损失和颜色相似性损失函数代替全监督 CrossVIS 网络实例 mask 分割损失，利用边界框标签对初始预测 mask 进行约束输出预测分割结果。

2.2　自适应调控模块

在特征提取网络中，FPN 主要对骨干网络提取的各层级特征图进行高层语义信息和低层细节信息的融合，用于增强特征图表达能力并提高网络对不同尺度目标的感知能力[31]。低层特征具有优秀的细粒度空间分辨率，包含丰富的细节信息特征，但语义信息表征能力弱，更适合检测小尺度目标；而高层特征拥有较大感受野，能提取到丰富的语义信息，但特征分辨率低，几何信息表征能力弱，更适合检测大尺度目标。在本文交通场景数据集中，由于车辆、行人等障碍物目标距离远近容易造成目标尺度动态范围过大，原有 FPN 将高低层特征进行简单的线性加权融合不仅会削弱高层特征图对大尺度目标局部位置信息的感知，还会降低低层特征对小尺度目标细节信息的提取能力，导致在目标大尺度范围变化下网络对前景目标的识别能力降低。因此，本文提出自适应调控模块（Adaptive Regulation Module，ARM），通过动态调整 FPN 不同层级信息贡献度以强化对目标局部位置和整体轮廓的聚焦能力，提高网络分割精度，具体结构如图 2 所示。

首先，ARM 模块将高层特征 X_H 和低层特征 X_L 经 1×1 卷积操作捕捉空间特征信息并压缩通道为 c，在本文实验中，c 设置为 256。然后，将通道压缩后的高层特征通过双线性插值上采样，使

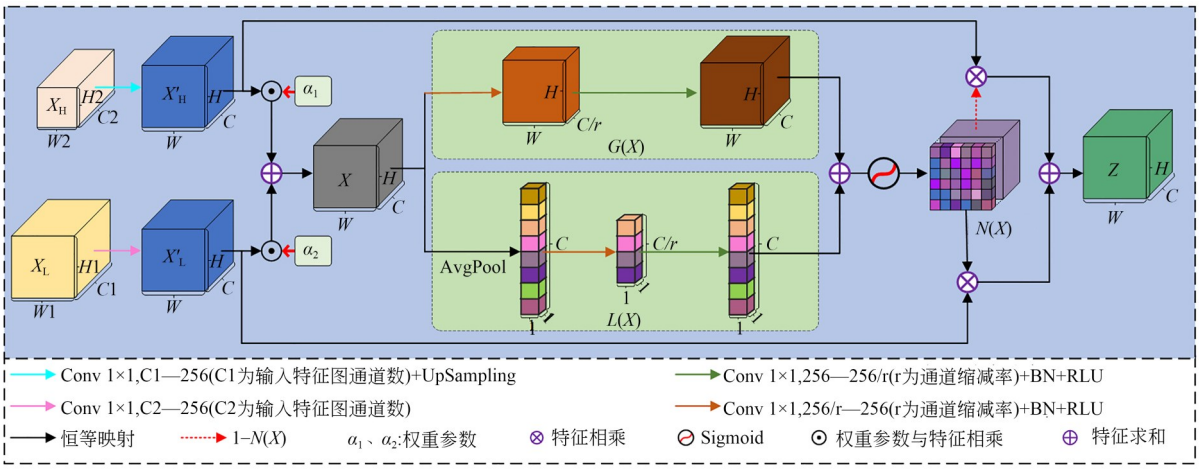


图 2　自适应调控模块

Fig. 2　Adaptive regulation module

其与 X_L 保持相同分辨率。最后，根据高低层特征贡献程度自适应赋予权重参数 α_1，α_2 后进行特征融合，其中 α_1，α_2 在模型训练过程中，通过网络梯度反向传播不断学习，自适应调整和更新以适应当前的分割任务，最终得到融合特征图 X：

$$\begin{cases} X=\alpha_1\cdot U\left(\mathrm{Conv}^1_{1\times1}(X_H)\right)\oplus\alpha_2\cdot\mathrm{Conv}^1_{1\times1}(X_L)= \\ \qquad \alpha_1\cdot X'_H\oplus\alpha_2\cdot X'_L \\ \alpha_i=\dfrac{\mathrm{e}^{w_i}}{\sum_j \mathrm{e}^{w_j}}\left(i=1,2;j=1,2\right) \end{cases}, \tag{5}$$

其中：α_i 为归一化权重，$\sum\alpha_i=1$，w_i 为初始化指数权重，w_j 为特征权重；$Conv^m_{k\times k}(\cdot)$ 表示卷积核为 $k\times k$，步长为 m 的卷积操作；U 表示上采样操作；$\oplus$ 表示逐元素相加；$\cdot$ 表示权重系数与特征图相乘。

由于高低层特征图之间的细粒度和语义信息不同，融合后的特征会造成信息冲突和冗余，削弱特征图的表达能力。因此，ARM 模块通过两个分支 $L(X)$ 和 $G(X)$ 来提取通道注意力，增强相关信息的关注，减弱不相关信息的干扰。$L(X)$ 采用全局平均池化操作提取全局上下文信息，然后采用 1×1 卷积将得到的注意力特征进行通道缩减，再使用 1×1 卷积操作进行通道还原，实现特征跨通道信息交互和信息整合，并降低网络计算量。$G(X)$ 采用两个 1×1 卷积进行通道信息交互，在不降低特征通道维度情况下建立权重映射关系，从而避免特征信息损耗。最终将两个分支输出的特征进行特征融合，对融合后的特征使用 Sigmoid 激活函数进行权重归一化以滤掉冗余信息，实现从不同尺度特征中自适应选择分割任务所需特征信息，生成注意力权重 $N(X)$：

$$\begin{cases} L(X)=BN\left(Conv^{c/r\to c}_{1\times1}\left(\delta\left(BN\left(Conv^{c\to c/r}_{1\times1}\left(GAP(X)\right)\right)\right)\right)\right) \\ G(X)=BN\left(Conv^{c/r\to c}_{1\times1}\left(\delta\left(BN\left(Conv^{c\to c/r}_{1\times1}(X)\right)\right)\right)\right) \\ N(X)=Sigmoid\left(L(X)\oplus G(X)\right) \end{cases}, \tag{6}$$

其中：$GAP(\cdot)$ 表示全局平均池化操作；$Conv^{n\to m}_{k\times k}$ 表示卷积核为 $k\times k$，输入通道数为 n，输出通道数为 m 的卷积操作；δ 表示 ReLU 激活函数；BN 表示批量归一化操作；$\oplus$ 表示逐元素相加。

为了抑制无关背景噪声的干扰和防止网络性能退化，将 Sigmoid 函数生成的注意力权重 $N(X)$ 分别与 X'_H 和 X'_L 相乘进行加权融合，最终生成自适应融合特征图 Z：

$$Z=\left(N(X'_L)\otimes X_L\right)\oplus\left(\left(1-N(X'_H)\right)\otimes X_H\right), \tag{7}$$

其中：$\otimes$ 表示逐元素相乘。自适应调控融合后的特征图 Z 能够有效强化对目标局部位置和整体轮廓的聚焦能力，克服了车辆、行人等目标由于距离远近造成的尺度动态范围过大问题。

2.3　生成调控模块

CrossVIS将特征提取阶段输出的F4，F5上采样与F3融合输入Mask Branch，通过一系列卷积操作生成原型掩膜。然而受限于卷积核尺寸，原有特征提取阶段只能有效表征局部信息，导致部分实例边缘纹理信息丢失。因此，本文在mask预测分支基础上设计生成调控模块（Generating Regulatory Module，GRM），其包含多细粒度提取模块和空间信息聚合模块，采用通道注意力机制聚合基于不同空洞率提取的多细粒度空间信息生成权重参数以调控各尺度特征，有效细化了实例边缘轮廓并增强了跨通道信息交互掩码特征映射表征能力，提高了模型对目标的定位精度和边缘轮廓分割精度，网络结构如图3所示。

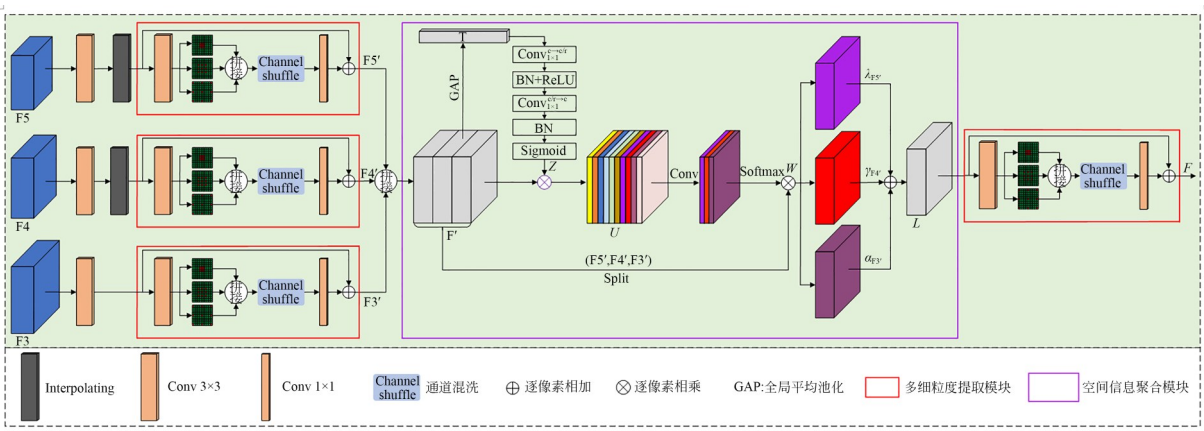


图3　生成调控模块

Fig. 3　Generating regulatory module

为有效表征不同层级特征实例边缘轮廓细节信息，本文基于ResNet50与FPN框架提取的$F3\in\boldsymbol{R}^{256\times48\times80}$，$F4\in\boldsymbol{R}^{256\times24\times40}$，$F5\in\boldsymbol{R}^{256\times12\times20}$作为生成调控模块的输入。首先，对输入各层级特征进行3×3卷积将其通道压缩为128，并对卷积后的高层特征分别进行双线性插值上采样与低层特征F5保持相同分辨率，然后分别送入多细粒度提取模块。具体地，使用3×3卷积进行初步特征提取，为了获取不同细粒度特征信息减少信息丢失，使用由三个空洞率$r=[1,2,5]$的空洞卷积并联组成的混合空洞卷积（Hybrid Dilation Convolution，HDC）[32]对特征进行实例边缘轮廓细节信息提取；将提取的不同细粒度特征在通道维度进行拼接和混洗，提高通道间信息的流动以增强通道间的关联性，然后使用1×1卷积进行通道降维；为防止网络性能退化，最后采用残差结构将输入特征与提取的轮廓细节信息进行跳跃连接得到新的特征$Fi'\in\boldsymbol{R}^{128\times48\times80}$，在增强特征提取能力的同时，丰富目标实例边缘轮廓细节信息，多细粒度提取模块计算公式为：

$$\begin{cases} HDC=SN\left(Cat\begin{pmatrix} AConv(r_1,1),\\ AConv(r_2,2),\\ AConv(r_3,5)\end{pmatrix}\right)\\ Fi'=Conv^1_{1\times1}\left(HDC\left(Conv^1_{3\times3}(Fi)\right)\right)\oplus Fi\end{cases},\quad(8)$$

其中：$AConv(r,n)$表示空洞率$r=n$的空洞卷积；$Cat(\cdot)$表示在通道维度进行拼接；$SN(\cdot)$表示通道混洗；Fi表示输入特征；Fi'表示输出特征；$\oplus$表示逐元素相加。

为解决卷积神经网络中各层级和通道间信息价值不等的问题，将由不同空洞率提取的多细粒度特征输入空间信息聚合模块生成权重参数以调控各尺度特征。具体地，首先将各层级特征图$F3'$，$F4'$，$F5'$在通道维度进行拼接生成新的特征$F'\in\boldsymbol{R}^{384\times48\times80}$。然后，采用挤压与激励（Squeeze and Excitation）[30]操作计算特征F'的通道注意力，增强关键通道信息并抑制无关冗余信息，提高网络对特征可分辨性。具体地，首先将拼接后的特征F'进行挤压操作，即对F'进行全局平均池化将全局信息压缩，建立不同通道间的相互依存关系，得到特征向量$T\in\boldsymbol{R}^{384\times1\times1}$。然后，

为自动获取每个特征通道重要程度，并抑制对当前任务用处不大的低效或无效的通道信息，将池化后的特征经两个1×1卷积操作完成特征激励，为有效降低计算量，通道压缩比例设置为$r=4$。最后使用Sigmoid函数生成各通道权重$Z\in\boldsymbol{R}^{384\times1\times1}$，并与原特征$Fi'$相乘得到特征图$U\in\boldsymbol{R}^{384\times48\times80}$：

$$\begin{cases}F'=Cat\left(\left[F5',F4',F3'\right]\right)\\T=GAP\left(F'\right)\\Z=\sigma\left(BN\left(Conv_{1\times1}^{c/r\to c}\left(\delta\left(BN\left(Conv_{1\times1}^{c\to c/r}\left(T\right)\right)\right)\right)\right)\right)\\U=Z\otimes F'\end{cases}\tag{9}$$

为进一步捕捉不同尺度特征信息对分割任务的重要性，通过1×1卷积操作压缩各尺度特征信息U通道数为3，利用Softmax函数进行空间信息权重归一化，得到权重矩阵$W\in\boldsymbol{R}^{3\times48\times80}$，然后权重矩阵在通道方向进行分割，以此生成特征重要性权重参数$\lambda,\gamma,\alpha\in\boldsymbol{R}^{1\times48\times80}$以调控各尺度特征，权重参数与各尺度特征相乘后得到新的特征$L\in\boldsymbol{R}^{128\times48\times80}$：

$$\begin{cases}W=\left[\lambda,\gamma,\alpha\right]=Softmax\left(\mathrm{Conv}_{1\times1}^{384\to3}\left(U\right)\right)\\L=\lambda F5'+\gamma F4'+\alpha F3'\end{cases}.\tag{10}$$

最后，将整合后的特征L再次送入多细粒度提取模块，通过扩大感受野增强特征全局信息和细粒度信息提取能力，进一步提升网络模型分割精度。

2.4　弱监督损失构建

2.4.1　正交损失约束

基于边界框的弱监督实例分割模型，用于网络监督学习的真实值仅为边界框标注信息。BoxInst[27]为确保覆盖生成预测mask最小外接框与真实边界框相匹配而提出的正交损失函数，通过边界框标注信息监督预测mask水平和垂直投影，缩小模型预测mask与真实边界框的偏差，具体操作如下：

首先，假设训练图像尺寸为$W\times H$，用于网络监督学习的真实边界框左上角和右下角坐标分别为(x_1,y_1)和(x_2,y_2)。然后，对训练图像建立横向真实行矩阵$X_{gt}\in\boldsymbol{R}^{1\times W}$和纵向真实值列矩阵$Y_{gt}\in\boldsymbol{R}^{H\times1}$，令行矩阵$\boldsymbol{X}_{gt}$的$x_1$至$x_2$位置所对应元素全为1，其余位置元素均为0，列矩阵$\boldsymbol{Y}_{gt}$的$y_1$

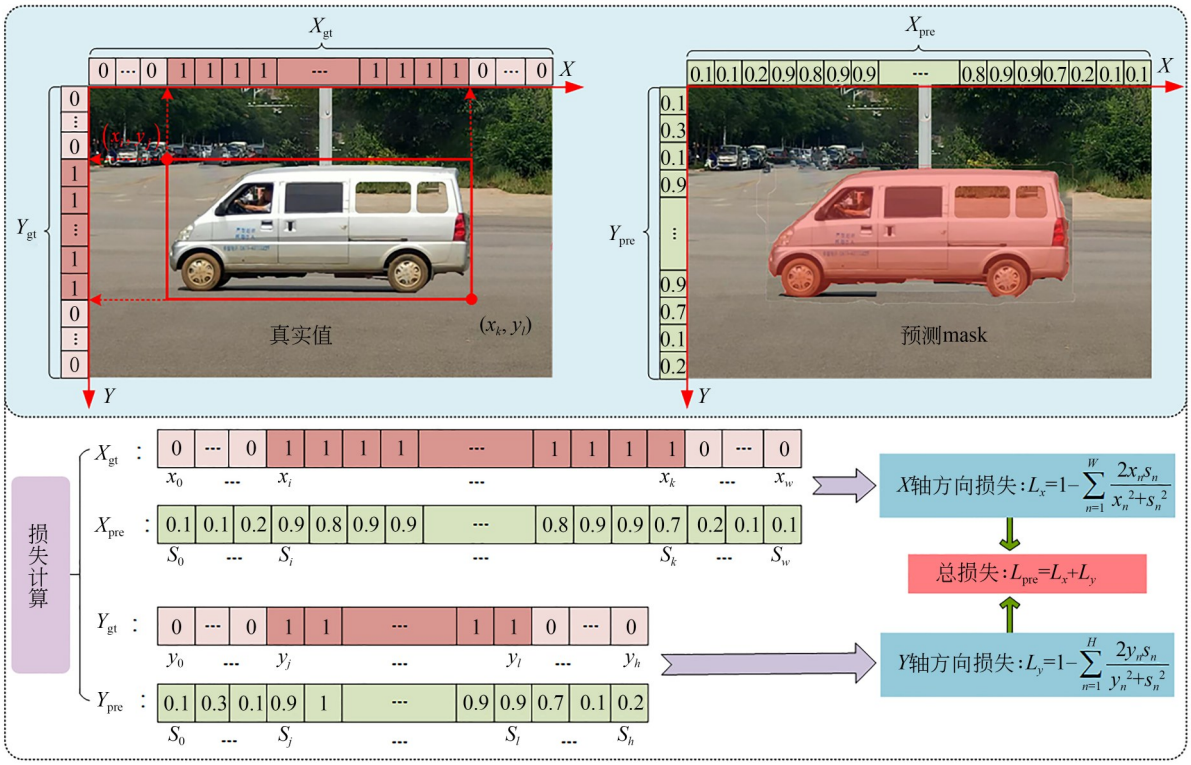


图4　正交损失构建

Fig. 4　Construction of orthogonal loss

至 y_2 位置对应元素为 1，其余位置元素均为 0。最后，假设 $S\in(0,1)^{H\times W}$ 为网络预测 mask，S 表示该像素点是前景目标的概率。计算预测 mask 分数每行和每列最大值 $S_n\in(0,1)$，组成列矩阵 $Y_{\text{pre}}\in\boldsymbol{R}^{H\times 1}$ 和行矩阵 $X_{\text{pre}}\in\boldsymbol{R}^{W\times 1}$。为了使预测的 mask 趋近真实 mask，应使网络预测 mask 与真实边界框在轴 X，Y 轴投影尽可能接近。因此，定义正交损失 L_x，L_y 如下：

$$L_x=1-\sum_{n=1}^{W}\frac{2x_n s_n}{x_n^2+s_n^2},\tag{11}$$

$$L_y=1-\sum_{n=1}^{H}\frac{2y_n s_n}{y_n^2+s_n^2}.\tag{12}$$

最后，将 X 轴和 Y 轴方向的损失相加得到正交损失 $L_{\text{pre}}=L_x+L_y$。

2.4.2　颜色相似性损失

仅通过正交损失对初始预测 mask 的约束，可以初步提高预测 mask 质量，但存在多个 mask 投影到同一真实边界框，导致预测 mask 质量和精细化程度下降。受启发于颜色相似的相邻像素间属于同类别概率较高，引入颜色相似性损失对预测 mask 进一步约束。在实例分割领域中，利用像素颜色差异可以对复杂背景中的目标对象进行有效区分，若像素间颜色相似性较高，则这些像素较大概率具有相同实例标签。因此，通过确定颜色相似性阈值 τ，当某两个像素点颜色相似性高于 τ，则它们标签相同的概率较高，由此引入颜色相似性[27]：

$$C_e=C(s_{i,j},s_{l,k})=\exp\left(-\frac{\|s_{i,j}-s_{l,k}\|}{\theta}\right),\tag{13}$$

其中：$s_{i,j}$，$s_{l,k}$ 为像素点 (i,j) 和 (l,k) 的颜色信息，e 表示像素点 (i,j) 和 (l,k) 之间的连线，C_e 表示 (i,j) 和 (l,k) 的颜色相似度，θ 是一个超参数，本文设置为 2，τ 值本文设置为 0.3。

为构建颜色相似性损失函数，在图像上建立一个无向图 $G=(V,E)$，其中 V 表示图像中所有像素点的集合，E 表示代表图像中两像素点连线的集合。在计算像素间两两相似性时，采取隔像素采样的方法以增大感受野，将每个像素同时与周围 8 个相邻点计算颜色相似性，示意图如图 5 所示。

定义 $y_e\in(0,1)$ 为边 e 的标签，当 $C_e>\tau$ 时，

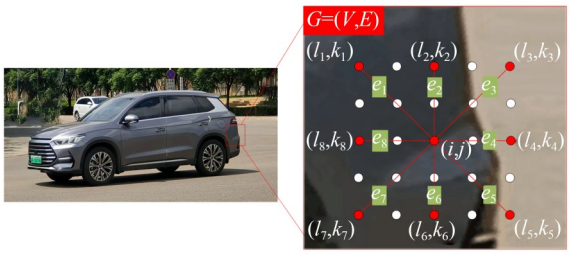


图 5　颜色相似性损失构建

Fig. 5　Construction of color similarity loss

$y_e=1$，表示边缘连接的两像素具有相同的标签，否则 $y_e=0$，表示两个像素点标签不同。设像素 (i,j) 和 (l,k) 为边缘的两个端点，网络预测 $p_{i,j}$ 可以看作像素 (i,j) 为前景的概率，则 $y_e=1$ 和 $y_e=0$ 的概率分别表示为：

$$P(y_e=1)=p_{i,j}\cdot p_{l,k}+(1-p_{i,j})\cdot(1-p_{i,j}),\tag{14}$$

$$P(y_e=0)=1-P(y_e=1).\tag{15}$$

因此，颜色相似性损失函数为：

$$L_c=-\frac{1}{N}\sum_{e\in E_{\text{in}}}\eta y_e\log P(y_e=1)+(1-y_e)\log P(y_e=0),\tag{16}$$

其中：E_{in} 表示框中至少包含一个像素的边的集合，使用 E_{in} 代替 E 可以防止损失被框外无效像素控制，N 是 E_{in} 的边数。当 $y_e=0$ 时，该边的标签未知，容易对网络造成错误的监督，所以在损失中丢弃 $(1-y_e)\log P(y_e=0)$ 一项，结合颜色相似性约束，最终损失函数为：

$$L_c=-\frac{1}{N}\sum_{e\in E_{\text{in}}}\eta y_e\log P(y_e=1),\tag{17}$$

其中：$\eta=\begin{cases}1, C_e\geqslant\tau\\ 0, C_e<\tau\end{cases}$。

3　实验结果与分析

3.1　实验数据集的建立

本文在 Youtube-VIS2019 数据集基础上抽取了交通场景中常见的人、摩托车、滑板、轿车、卡车、火车、狗七个类别目标作为本文数据集，其中训练集 329 个视频片段，总帧数 7 212 帧包含 603 个实例，验证集 53 个视频片段，总帧数 1 097 帧包含 88 个实例。本文在训练过程中仅使用训练集中的边界框标签和类别标签对网络进行监

督训练，而测试集使用与全监督数据集一致像素级标签、边界框标签和类别标签对网络模型定量分析。

3.2 实验配置

本文实验平台为Ubuntu18.04操作系统，CPU为Intel(R) Core(TM) i9-10400F处理器，GPU为NVIDIA GTX 3060显卡，显存为12GB的计算机。深度学习框架为pytorch1.8.0，python版本为3.7、采用CUDA11.1和cuDNN8.0.5加速网络模型训练。

实验过程中，将输入图像尺寸统一为360×640并将批处理尺寸(Batch Size)设置为4。在训练阶段初始学习率设置为0.000 5、迭代次数为12Epoch，每个Epoch输出一个模型权重，对最后的训练模型的精度和推理速度综合比较后选出最优模型。

3.3 评价指标

本文使用平均精度(Average Precision，AP)和平均召回率(Average Recall，AR)作为网络模型的评价指标，而在实例分割任务中，常以预测值与真实值的交并比IoU(Intersection over Union，IoU)来确定算法的评价指标AP和AR值。视频例分割中IoU的定义与图像实例分割有所不同，较为注重相同实例在时序上空间位置的关联情况。给定一个视频序列的真实掩膜$m_{i\cdots j}$和预测掩膜$\tilde{m}_{\tilde{i}\cdots\tilde{j}}$，其中$i,j$代表时序信息。假如在$t$帧静态图像中没有出现目标实例，那么利用空白掩膜对该帧信息进行填补，具体可表示为$m_t=0$或$\tilde{m}_{\tilde{t}}=0$，即把IoU从图像扩展到视频序列，视频实例分割IoU计算公式如式(18)所示：

$$IoU(a,b)=\frac{\sum_{t=1}^{\mathrm{T}}\left|m_t^a\cap\tilde{m}_t^b\right|}{\sum_{t=1}^{\mathrm{T}}\left|m_t^a\cup\tilde{m}_t^b\right|},\tag{18}$$

其中：a和b分别表示为某个实例的真实值和预测值。

求得IoU之后，按照0.05的增量在0.50至0.95区间取值10个IoU作为阈值，AP为这10个阈值下对应的平均精度的均值，AP(50)和AP(75)分别表示IoU阈值为50%和75%时的平均精度，AP值越大表示视频实例分割效果越好；召回率AR表示真实分割结果的所有目标像素中被分割出来的目标像素所占的比例，主要衡量模型预测正样本的能力，其中AR1表示每帧图像按照IoU由高到低选取1个结果计算平均召回率，AR10表示每帧图像按照IoU由高到低选取10个结果计算平均召回率。相关计算公式如下：

$$AP=\frac{TP}{TP+FP}\times 100\%,\tag{19}$$

$$AR=\frac{TP}{TP+FN}\times 100\%,\tag{20}$$

其中：TP表示正确检测为正样本的个数，FP表示误检为正样本的个数，FN表示漏检为正样本的个数。

3.4 实验结果与定量分析

3.4.1 自适应调控实验分析

本节根据是否自适应更新权重参数以及权重是否归一化将自适应调控模块设计为3类，分别为对FPN中高低层特征赋予常量α_1和α_2的权重平衡模型(Weight Balance Model，WBM)、对高低层特征赋予初始化为1的自适应权重未归一化模型(Weight Unnormalization Model，WUM)和初始化为1的自适应权重归一化模型(Weight Normalization Madel，WNM)，并对三类模型进行实验，其实验结果如表1所示。

从表1可知，WBM将高低层特征赋予常量α_1和α_2的权重，由于高低层特征对网络分割任务的贡献度不等，人为赋值权重α_1和α_2需要大量调参实验才能取得最优解，当人为赋值$\alpha_1=1$，$\alpha_2=0.25$时，网络平均分割精度为34.8%，相较于Box-CrossVIS基准提升了1.8%。WUM由于自适应生成的权重参数未进行归一化处理，会导致权重参数过大引起网络梯度爆炸，平均分割精度为33.4%，较Box-CrossVIS提升了0.4%。WNM将自适应权重进行归一化处理，使模型根据数据特征分布来自行决定特征权重，强化了对目标局部位置和整体轮廓的聚焦能力，在自适应调控模块中平均分割精度达到最高35.6%，较Box-CrossVIS提升了2.6%。因此本文选择初始化为1的自适应权重归一化模型WNM作为自适应调控模块的最终模型。

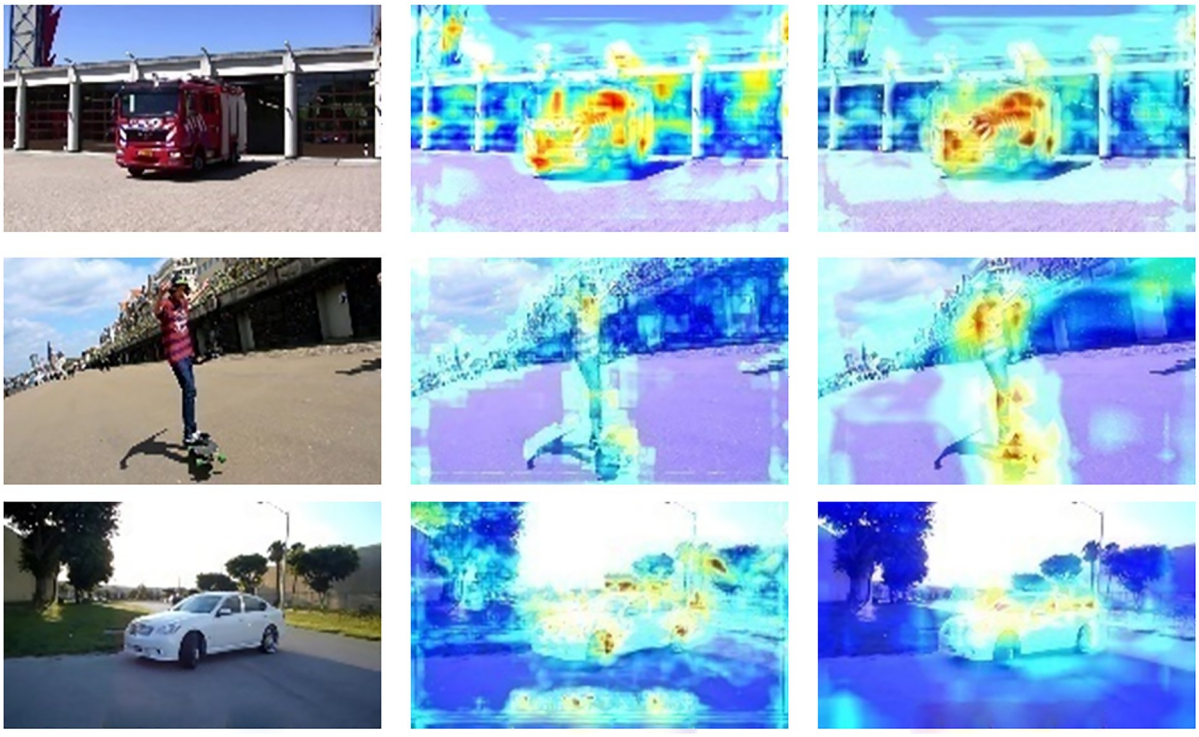
为解释自适应调控模块的工作机理以及对最终分割结果的有效性，本文对特征金字塔网络的最低层特征F3进行热力图可视化分析。热力图可以直观反映模型在图像上的关注区域，热力

表 1　不同权重实验结果对比

Tab. 1　Comparison of experimental results with different weights

模型	方法	AP/%	AP50/%	AP75/%	AR1	AR10	FPS
Baseline	Box-CrossVIS	33.0	61.5	32.3	35.4	39.0	18
Constant α_1, α_2	Box-CrossVIS+ARM(WBM:$\alpha_1=0.25, \alpha_2=1$)	34.4	66.7	32.6	34.9	40.4	18
	Box-CrossVIS+ARM(WBM:$\alpha_1=1, \alpha_2=0.25$)	34.8	64.3	35.4	34.9	39.9	17
	Box-CrossVIS+ARM(WBM:$\alpha_1=\alpha_2=0.25$)	34.6	64.5	36.7	35.1	40.0	18
	Box-CrossVIS+ARM(WBM:$\alpha_1=0.50, \alpha_2=0.75$)	34.5	61.8	37.6	34.9	39.8	18
	Box-CrossVIS+ARM(WBM:$\alpha_1=0.75, \alpha_2=0.50$)	33.7	0.1	31.7	35.6	39.8	18
	Box-CrossVIS+ARM(WBM:$\alpha_1=\alpha_2=0.50$)	34.1	62.2	34.8	35.7	39.9	17
	Box-CrossVIS+ARM(WBM:$\alpha_1=\alpha_2=0.75$)	32.7	59.7	33.4	34.4	38.6	18
	Box-CrossVIS+ARM(WBM:$\alpha_1=\alpha_2=1$)	34.1	64.8	34.9	35.3	40.5	18
Adaptive α_1, α_2	Box-CrossVIS+ARM(WUM)	33.4	60.2	33.6	35.4	39.6	17
	Box-CrossVIS+ARM(WNM)	**35.6**	**64.3**	**36.8**	**36.7**	**41.1**	**18**

图内颜色越深,表明模型对该区域的关注程度越高。如图 6 所示(彩图见期刊电子版),(a)为模型输入的原图像、(b)和(c)分别为 Box-CrossVIS 基准网络和嵌入 ARM 模块后的热力映射图。图(b)中红色高亮感兴趣区域除了集中在前景目标上之外,还扩散到背景目标上,对分割任务存在一定的干扰。图(c)中红色高亮区域明显集中于需要精确分割的前景目标上,并对背景进行抑



(a) Original　(b) Box-CrossVIS　(c) Box-CrossVIS+ARM

图 6　热力图可视化对比

Fig. 6　Visual Comparison of heat map

制，引导网络在特征提取过程中重点关注目标实例特征信息，明显地减少了无关区域的特征权重占比。说明嵌入自适应调控模块后，网络能有效学到感兴趣区域信息，并对前景重要目标特征予以关注，增强了网络分割效果。

3.4.2　生成调控模块实验分析

为获取最佳的多层级特征细粒度空间信息聚合效果，实验中对特征提取阶段输出的5个特征层按逐层递增顺序进行了不同尺度的融合实验，实验结果如表2所示。

表2　多尺度融合策略对分割结果的影响对比

Tab. 2　Comparison of effects of multi-scale fusion strategies on segmentation results

层数	输入层	AP/%	AP50/%	AP75/%	AR1	AR10	FPS
基准	Box-CrossVIS	33.0	61.5	32.3	35.4	39.0	18
2	F3+F4	34.9	64.8	33.6	36.7	39.5	18
2	F6+F7	17.6	45.0	10.3	20.4	23.8	17
3	**F3+F4+F5**	**35.6**	**64.3**	**36.3**	**36.5**	**40.5**	**18**
4	F3+F4+F5+F6	34.6	62.2	32.5	36.2	39.8	17
5	F3+F4+F5+F6+F7	34.2	62.9	34.2	35.3	40.0	17

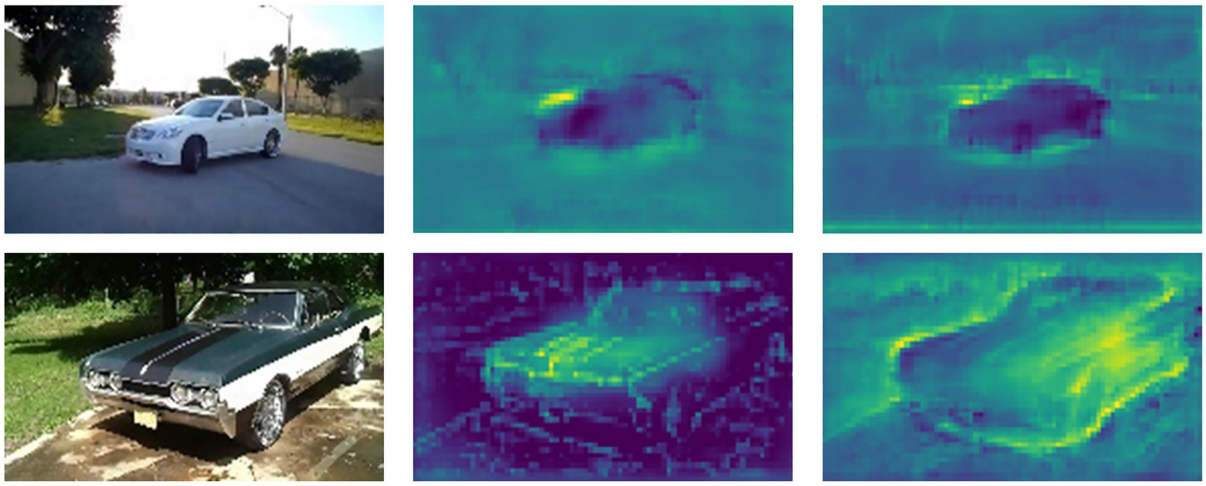
结果表明，当生成调控模块对输入特征F3和F4进行聚焦融合时，模型平均分割精度为34.9%，较基准网络Box-CrossVIS提升1.9%。当生成调控模块对输入特征F6和F7聚焦融合时，模型平均分割精度仅为17.6%，较基准网络Box-CrossVIS降低15.4%。当输入的特征图层数为3时，F3，F4和F5三个特征层聚合多细粒度空间信息生成权重参数以调控各尺度特征(GRM)取得了最好的分割效果，模型平均分割精度达到35.6%，较基准网络Box-CrossVIS提升2.6%。当输入特征图层数为3和4时，模型平均分割精度分别为34.6%和34.2%，相较于基准Box-CrossVIS分割精度分别提升1.6%和1.2%。综上所述，当生成调控模块对输入特征F3，F4或F3，F4和F5进行聚焦融合时，模型平均分割精度均有不同程度的提升，但随着高层特征F6和F7的加入，模型平均分割精度随层数增加有所降低，且当生成调控模块仅聚焦融合高层特征F6和F7时，模型平均分割精度有较大降低。充分说明低层特征F3，F4，F5对边界轮廓特征信息贡献度较大，而高层F6，F7对边界轮廓特征信息贡献度偏低。

为了更直观突显生成调控模块在交通场景视频序列中对障碍物实例边缘细节信息提取的有效性，本文选择对mask预测分支的输入特征图进行可视化分析。考虑到该特征包含128个通道维度，分别提取Box-CrossVIS基准模型与嵌入GRM模块后模型的第一层通道特征进行可视化分析以保证对比条件的一致性，引入GRM模块前后特征图可视化对比结果如图7所示。从图(b)可视化结果可以看出，Box-CrossVIS提取的特征图实例边缘轮廓粗糙，前景与背景对比度不明显，而从图(c)可以看出，嵌入生成调控模块后，模型对前景目标实例的边缘轮廓特征提取能力明显高于基准模型，增大了背景与前景的反差对比度，优化了模型对目标边缘轮廓的定位准确性，有效减少图像边缘信息丢失。结果表明生成调控模块可以通过注意力机制聚合基于不同空洞卷积率提取的多细粒度空间信息，细化了实例边缘轮廓，有效弥补边缘轮廓分割mask连续性缺失，提高了本文算法的分割精度。

3.4.3　不同网络实验结果对比

考虑到弱监督视频实例分割相关工作较少，本文选择全监督网络YolactEdge，STMask，CrossVIS与本文模型做对比，以客观评价SAGRNet模型对交通场景障碍物识别分割任务的优越性。为保证验证结果有效性和公平性，对比实验均在同一设备上开展且使用同一数据集，算法性能对比如表3所示。

结果表明，本文模型SAGRNet平均分割精度最高达到38.1%，较弱监督Box-CrossVIS基准网络分割精度提升5.1%，较全监督网络



(a) Original　(b) Box-CrossVIS　(c) Box-CrossVIS+ARM

图 7　特征图可视化结果

Fig. 7　Visualization of characteristic image

表 3　不同模型的实验结果对比

Tab. 3　Comparison of experimental results of different models

网络	弱监督	AP/%	AP50/%	AP75/%	AR1	AR10	FPS
Yolactedge[20]		36.0	60.6	37.28	—	—	29
STMask[18]		40.0	62.7	42.9	0.421	0.455	22
CrossVIS[23]		40.6	65.7	42.2	43.2	46.4	16
Box-CrossVIS	√	33.0	61.5	32.3	35.4	39.0	18
SAGRNet	√	**38.1**	**69.2**	**38.3**	**37.8**	**42.7**	**17**

YolactEdge 分割精度提升 2.1%,但平均分割速度降低了 11FPS;与全监督网络 CrossVIS 和 STMask 网络相比,分割精度分别降低仅 2.5% 和 1.9%,但数据集的标注成本却有大幅度降低。综上所述,本文算法能在弱监督条件下取得与部分全监督网络相当的分割效果,验证了本文改进算法 SAGRNet 的优越性。

3.4.4　SAGRNet 消融实验

为验证各改进模块在交通场景数据集中针对目标障碍物的有效分割,本文以弱监督 Box-CrossVIS 算法为基准,分别与添加了本文 ARM 和 GRM 模块的不同网络进行实验对比,实验结果如表 4 所示。

由表 4 可知,弱监督 Box-CrossVIS 算法平均分割基准精度为 33.0%。首先,在 FPN 融合路径嵌入 ARM 模块,改进了网络高低层特征融合方式,解决了多尺度信息直接融合效率低下的问题。结果表明,嵌入 ARM 模块后在平均分割速度保持不变的情况下,平均分割精度达到 35.6%;其次,在 mask 预测分支前添加 GRM 模块,通过注意力机制聚合基于不同空洞率提取的多细粒度空间信息进行多尺度特征调控,以有效

表 4　消融实验结果

Tab. 4　Result of ablation experiments

方法	ARM	GRM	AP/%	AP50/%	AP75/%	AR1	AR10	FPS
Box-CrossVIS	×	×	33.0	61.5	32.3	35.4	39.0	18
Ours	√	×	35.6	64.3	36.8	36.7	41.1	18
Ours	×	√	35.6	64.3	36.3	36.5	40.5	18
Ours	√	√	**38.1**	**69.2**	**38.3**	**37.8**	**42.7**	**17**

弥补实例边缘信息匮乏导致的边缘轮廓分割mask连续性缺失问题，分割精度达到35.6%；最后，将两种方法组合使用，以38.1%的平均分割精度达到最优结果，以上模块比Box-CrossVIS基准分别提升了2.6%，2.6%和5.1%，验证了本文算法有效性。

3.4.5　实验平台搭建与验证

由上述对比实验可知，在交通场景数据集上，本文提出的算法能够有效提高模型对障碍物目标的识别和分割精度。为有效验证SAGRNet算法在交通场景中应用的可行性，本文基于辅助驾驶小车获取复杂交通场景下的视频数据，并与CrossVIS，Box-CrossVIS算法进行对比验证。该辅助驾驶小车搭载激光雷达、毫米波、GPS和摄像机等设备，其中获取视频数据所使用的摄像头型号为1080P(SP5268)，最大分辨率为1 920×1 080，网络分割可视化结果如图8所示。对比(a1)，(a2)分割结果，CrossVIS网络生成的mask质量优秀，边缘轮廓清晰，而对于仅使用边界框进行训练的弱监督Box-CrossVIS算法，由于监

图8　分割结果可视化

Fig. 8　Visualization of segmentation results

督信息减弱导致网络难以准确地挖掘和定位目标实例，造成分割实例存在边缘轮廓粗糙、不连续等问题。在(a3)中本文算法SAGRNet通过聚合基于不同空洞率提取的多细粒度空间信息改善了实例边缘纹理信息丢失的问题，实现了对目标实例的准确定位与分割。在(b1)，(b2)视频序列中，均存在行人不完全分割、过分割的问题，在(b3)中由于自适应调控模块强化了对目标局部位置和整体轮廓的聚焦能力，提高了网络对目标实例的捕捉能力。综上所述，本文模型SAGRNet相比于Box-CrossVIS而言能更好适应交通场景大尺度动态范围目标的分割问题，有效降低模型的误检率和漏检率，有更高的检测分割精度以及更好的鲁棒性。

4　结　论

本文针对辅助驾驶车辆对复杂交通场景下多目标感知和场景理解的需求，提出一种自适应生成调控弱监督视频实例分割算法SAGRNet。首先，引入正交损失和颜色相似性损失代替CrossVIS实例mask分割损失，利用边界框信息监督网络训练，实现基于边界框的弱监督视频实例分割Box-CrossVIS；其次，引入自适应调控模块强化对目标局部位置和整体轮廓的聚焦能力，增强网络对不同尺度变化情况下前景目标的感知能力；最后，设计生成调控模块聚合多细粒度空间信息，弥补边缘轮廓分割mask连续性缺失问题。经实验验证，本文算法能有效提高辅助驾驶车辆对复杂交通场景下多目标障碍物的检测和分割精度，平均分割精度达到38.1%，较Box-CrossVIS模型提高5.1%，且在2080Ti上最高分割速度可达36 FPS，能够满足实时检测分割需求。此外，本文还搭建了辅助驾驶小车实验平台验证本文算法的可行性。

尽管已经取得了显著的进展，本文算法仍存在进步空间。一方面，辅助驾驶车辆在交通场景下自主行驶时捕获的障碍物目标普遍存在相互遮挡的情况，严重的遮挡会带来易混淆的遮挡边界及非连续自然的物体形状，影响网络对物体整体结构的判断，出现欠分割或错分现象，网络抗干扰能力有待提高；另一方面，本文算法仅依靠边界框信息对网络进行训练，由于监督信息的减弱会面临局部聚焦，难以准确地挖掘和定位所有目标实例等问题。因此，在正交损失和颜色相似性损失的基础上，通过引入光流等相关技术获取视频序列中的外观和运动信息对初始预测mask进一步约束，缩小视觉弱监督学习与全监督学习的性能差异，并将其应用于实际的视觉理解应用，仍然是未来视觉弱监督视频实例分割研究的重点。

参考文献：

[1] BAYU A, WIBISONO A, WISESA H A, *et al*. Semantic segmentation of lidar point cloud in rural area[C]. 2019 *IEEE International Conference on Communication, Networks and Satellite* (*Comnetsat*). 1-3, 2019, *Makassar, Indonesia*. IEEE, 2019: 73-78.

[2] MIHAI S, SHAH P, MAPP G, *et al*. Towards autonomous driving: a machine learning-based Pedestrian *Detection System Using* 16-*Layer LiDAR*[C]. 2020 13*th International Conference on Communications* (*COMM*). 18-20, 2020, *Bucharest, Romania*. IEEE, 2020: 271-276.

[3] 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)

[4] RAGURAMAN S J, PARK J. Intelligent drivable area detection system using camera and lidar sensor for autonomous vehicle[C]. 2020 *IEEE International Conference on Electro Information Technology* (*EIT*). IEEE, 2020: 429-436.

[5] DOVIS F, IMAM R, QIN W J, *et al*. Opportunistic use of gnss signals to characterize the environment by means of machine learning based processing [C]. *ICASSP* 2020 - 2020 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). 4-8, 2020, *Barcelona, Spain*. IEEE, 2020: 9190-9194.

[6] JOUBERT N, REID T G R, NOBLE F. Developments in modern gnss and its impact on autonomous vehicle architectures[C]. 2020 *IEEE Intelligent Ve-*

hicles Symposium (*IV*). *October* 19 - *November* 13, 2020, *Las Vegas*, *NV*, *USA*. IEEE, 2021: 2029-2036.

[7] NAVARRO M, ARRIBAS J, VILÀ-VALLS J, *et al*. Hybrid GNSS/INS/UWB positioning for live demonstration assisted driving[C]. 2019 *IEEE Intelligent Transportation Systems Conference* (*ITSC*). 27-30, 2019, *Auckland*, *New Zealand*. IEEE, 2019: 3294-3301.

[8] CHEN C, XIONG G M, ZHANG Z H, *et al*. 3D LiDAR-GPS/IMU calibration based on hand-eye calibration model for unmanned vehicle[C]. 2020 3*rd International Conference on Unmanned Systems* (*ICUS*). 27-28, 2020, *Harbin*, *China*. IEEE, 2020: 337-341.

[9] 王中宇, 倪显扬, 尚振东. 利用卷积神经网络的自动驾驶场景语义分割[J]. 光学 精密工程, 2019, 27(11): 2429-2438.
WANG Z Y, NI X Y, SHANG Z D. Autonomous driving semantic segmentation with convolution neural networks[J]. *Opt. Precision Eng.*, 2019, 27(11): 2429-2438. (in Chinese)

[10] LEE K F, CHEN X Z, YU C W, *et al*. An intelligent driving assistance system based on lightweight deep learning models[J]. *IEEE Access*, 2022, 10: 111888-111900.

[11] 孙建波, 张叶, 常旭岭. 基于改进Mask R-CNN+LaneNet的车载图像车辆压线检测[J]. 光学 精密工程, 2022, 30(7):854-868.
SUN J B, ZHANG Y, CHANG X L. Vehicle pressure line detection based on improved Mask R-CNN+LaneNet[J]. *Opt. Precision Eng.*, 2022, 30(7):854-868. (in Chinese)

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

[13] WANG Y Q, XU Z L, SHEN H, *et al*. Centermask: single shot instance segmentation with point representation[C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 13-19, 2020, *Seattle*, *WA*, *USA*. IEEE, 2020: 9310-9318.

[14] 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.

[15] CHEN H, SUN K Y, TIAN Z, *et al*. Blendmask: top-down meets bottom-up for instance segmentation[C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 13-19, 2020, *Seattle*, *WA*, *USA*. IEEE, 2020: 8570-8578.

[16] 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.

[17] 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.

[18] 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*). 20-25, 2021, *Nashville*, *TN*, *USA*. IEEE, 2021: 11210-11219.

[19] WANG Y Q, XU Z L, WANG X L, *et al*. End-to-end video instance segmentation with transformers[C]. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 20-25, 2021, *Nashville*, *TN*, *USA*. IEEE, 2021: 8737-8746.

[20] LIU H T, RIVERA SOTO R A, XIAO F Y, *et al*. Yolactedge: real-time instance segmentation on the edge[C]. 2021 *IEEE International Conference on Robotics and Automation* (*ICRA*). *May* 30-*June* 5, 2021, *Xi'an*, *China*. IEEE, 2021: 9579-9585.

[21] BERTASIUS G, TORRESANI L. Classifying, segmenting, and tracking object instances in video with mask propagation[C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 13-19, 2020, *Seattle*, *WA*, *USA*. IEEE, 2020: 9736-9745.

[22] ATHAR A, MAHADEVAN S, OŠEP A, *et al*. *STEm-Seg: Spatio-Temporal Embeddings for Instance Segmentation in Videos*[M]. Computer Vision - ECCV 2020. Cham: Springer International Publishing, 2020: 158-177.

[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*). 10-17, 2021, *Montreal*, *QC*, *Canada*. IEEE, 2022: 8023-8032.

[24] SONG C F, HUANG Y, OUYANG W L, *et al*. Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation[C]. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 15-20, 2019, *Long Beach*, *CA*, *USA*. IEEE, 2020: 3131-3140.

[25] 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: 6582-6593.

[26] WANG X G, FENG J P, HU B, *et al*. Weakly-supervised instance segmentation via class-agnostic learning with salient images[C]. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 20-25, 2021, *Nashville*, *TN*, *USA*. IEEE, 2021: 10220-10230.

[27] 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*). 20-25, 2021, *Nashville*, *TN*, *USA*. IEEE, 2021: 5439-5448.

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

[29] 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*). 21-26, 2017, *Honolulu*, *HI*, *USA*. IEEE, 2017: 936-944.

[30] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18-23, 2018, *Salt Lake City*, *UT*, *USA*. IEEE, 2018: 7132-7141.

[31] 黄滢, 何自芬, 杨宏宽, 等. 极化自注意力调控的情景式视频实例多尺度分割[J]. 计算机学报, 2022, 45(12): 2605-2618.
HUANG Y, HE Z F, YANG H K, *et al*. Multi-scale segmentation of episodic video instance through polarized self-attention manipulation[J]. *Chinese Journal of Computers*, 2022, 45(12): 2605-2618. (in Chinese)

[32] WANG P Q, CHEN P F, YUAN Y, *et al*. Understanding convolution for semantic segmentation[C]. 2018 *IEEE Winter Conference on Applications of Computer Vision* (*WACV*). 12-15, 2018, *Lake Tahoe*, *NV*, *USA*. IEEE, 2018: 1451-1460.

作者简介:

张印辉(1977—),男,河北故城人,博士,教授、博士生导师,2000年和2005年于西安理工大学分别获得学士和硕士学位,2010年于昆明理工大学获得博士学位,主要从事图像处理、机器视觉及机器智能等方面的研究。E-mail:yinhui_z@163.com

通讯作者:

何自芬(1976—),女,山西阳泉人,博士,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