陈萌放,徐迪红,李国亮,等.基于XMem-SimAM的半监督猪只视频分割方法[J].华中农业大学学报,2025,44(2):17-28.
DOI:10.13300/j.cnki.hnlkxb.2025.02.003

基于XMem-SimAM的半监督猪只视频分割方法

陈萌放[1,2],徐迪红[2,3],李国亮[1,2],刘小磊[4],周明彦[2,3],黎煊[2,3]

1.华中农业大学信息学院,武汉 430070; 2.农业农村部智慧养殖技术重点实验室,武汉 430070;
3.华中农业大学工学院,武汉 430070; 4.湖北洪山实验室,武汉 430070

摘要 为解决因猪场复杂环境、猪只动态生长及体型变化等因素导致的猪只精确分割难题,以种猪性能测定过程中动态采食和生长过程的猪只为研究对象,构建一个包括234个视频序列的猪只视频数据集,提出基于XMem-SimAM的半监督猪只视频分割方法。通过引入SimAM注意力进行多尺度特征融合,提升模型在不同尺度下对时序信息的提取能力,捕捉猪只动态移动的时序特征;利用空间-通道注意力模块,强化模型对时序语义特征的权重提取;优化多尺度特征融合策略和上采样模块,充分利用视频序列中的时序关联信息,从细粒度层面提高视频中猪只分割精度。经过测试对比,XMem-SimAM模型在猪只视频数据集上的区域相似度Jaccard、轮廓准确度F、平均度量$J\&F$和Dice系数分别达到96.9、95.8、98.0和98.0,优于MiVOS、STCN、DEVA、XMem++等视频对象分割方法,显示出卓越的分割性能;在推理阶段,处理速度达到58.5帧/s,内存消耗为795 MB,实现了处理效率与资源利用的良好平衡。结果表明,该方法可应用于猪场复杂环境下动态生长猪只的视频分割。

关键词 半监督; 视频分割; 猪只; SimAM注意力

中图分类号 TP391.4 **文献标识码** A **文章编号** 1000-2421(2025)02-0017-12

利用计算机视觉、机器学习等技术逐步实现生猪养殖数字化、管理精细化是现代生猪产业高质量发展的必然趋势[1-3]。近年来,基于深度学习的语义分割方法在猪只体尺测量、体温测量、体质量估计等猪只表型测试任务中得以广泛应用。耿艳利等[4]采用PointNet网络结合注意力模块,构建了猪只点云语义分割模型,实现了猪只体尺测量;Wang等[5]提出了一种基于RANSAC和欧几里得聚类的点云语义分割算法,用于自动提取猪只体型尺寸;Xie等[6]提出基于红外热像仪(infrared thermography,ITG)的猪只温度自动检测方法并利用改进的YOLOv5s-BiFPN模型实现感兴趣区域(region of interest,ROI)分割,从而实现猪只的自动温度检测;Zhang等[7]提出基于改进的凸性活动轮廓模型的方法,从红外图像中分割出猪只体温区域,进行体温测量与监控;He等[8]通过Swin Transformer从RGB和深度图像提取信息进行前景分割,并结合特征融合层对猪只体质量进行预测;Kwon等[9]基于快速重建网格模型,通过语义分割专注于猪只躯干部分,实现猪只体质量估计。然而,现代化养殖过程中,猪只存在动态移动、姿态变化等问题,上述语义分割方法依赖单帧图像,无法充分捕捉猪只视频序列的时序信息。这种局限性限制了语义分割技术在猪只养殖复杂环境中的应用。

视频对象分割通过利用视频序列中帧与帧之间的时序关联,能够更好地跟踪和分割动态场景中的猪只,增强模型适应能力。Cho等[10]提出了一种运动选项网络(treating motion-as-option network,TMO),在包含飞机、车辆等快速移动的DAVIS[11]数据集上测试,通过引入协作学习策略捕获动态目标的时序特征,分割精度达到85.7;Lee等[12]提出了一种原型记忆网络(prototype memory network,PMN),利用超像素算法增强帧序列的连通性,在包含光照变化室内场景的YouTube-Objects[13-14]数据

收稿日期:2024-12-16
基金项目:生猪现代育种技术研发及新品种选育(HBZY2023B006-03);武汉市生物育种重大专项(2022021302024853);华中农业大学-中国农业科学院深圳农业基因组研究所合作基金项目(SZYJY2022031)
陈萌放,E-mail:1334233852@qq.com
通信作者:黎煊,E-mail:lx@mail.hzau.edu.cn

集上分割精度达到85.9；Cho等[15]提出了一种Fake-Flow方法，在包含动物迁移、城市交通等复杂背景的YouTube-VOS[16]数据集上进行测试，通过合成的光流图生成图像-光流对以捕捉动态时序特征，分割精度达到86.1。上述研究表明，视频分割能够对动态对象有较好的处理能力。

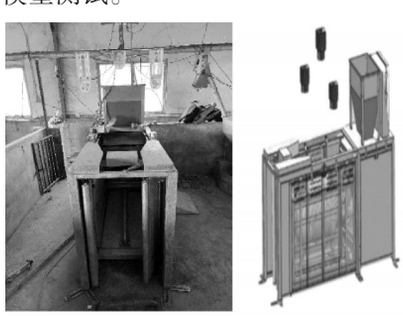
半监督视频分割方法仅需视频第一帧像素级标注，后续帧通过学习到的时序特征自动完成分割，通过充分利用视频序列中的时序关联信息，具有对复杂场景中动态移动目标的高效识别和精准分割的优点。近年来，视频对象分割的半监督方法取得了显著进展。Yang等[17]将目标转换器关联（associating objects with transformers，AOT）方法引入Transformer架构，通过单帧标注信息，利用强大的序列建模能力捕捉视频帧之间的时序关联信息。这种动态适应机制能提高对快速移动目标的跟踪效果，能够动态调整复杂背景下的特征提取能力。Yang等[18]将目标可扩展转换器关联（associating objects with scalable transformers，AOST）方法引入注意力机制，通过对单帧标注进行时序特征的加权处理，有效建模帧与帧之间的时序动态关系，使得在复杂场景和快速运动的目标中表现更加稳健，展示了其显著的精度提升和适应能力。Cheng等[19]提出了XMem方法，通过高效利用内存机制和长距离建模从首帧标注信息提取时序特征，动态调整对视频帧的关注点，确保对目标对象的持续追踪和精确分割，其优势在于能够有效处理长视频序列，并在长时间的动态变化中保持较高的精确度。然而，针对猪只的视频对象分割方法研究相对不足，限制了其在猪只表型测试领域的发展。

本研究以种猪性能测定过程中动态采食和生长过程的猪只为对象，构建猪只视频数据集，引入无参注意力机制[20]和多尺度特征融合，优化上采样和特征融合策略，增强对动态特征的捕捉能力，提升对猪只关键区域的聚焦；设计基于XMem-SimAM的半监督猪只视频分割模型，并与STCN[21]、DEVA[22]、Mi-VOS[23]、XMem++[24]等视频对象分割方法进行对比，提高模型分割精度，以期为猪场复杂环境下动态生长猪只视频分割提供新方法。

1 材料与方法

1.1 猪只视频数据获取

研究猪只为测定站饲养，数据采集于猪只在测定站区域采食阶段，采集过程为全天候自动进行，为了从多视角增强模型对猪只全局和局部特征的学习能力，本研究将3台REB-USB1080P01-LS43相机架设在猪只养殖测定站上方，从左视角、俯视角和右视角3个角度帧率同步采集猪只视频数据（视频分辨率为1 920像素×1 080像素），如图1所示，数据采集过程中，所有相机均采用15帧/s的设置进行稳定采集，以确保猪只动态移动的情况下，能够捕捉到清晰的视频帧，确保模型在处理动态特征时的准确性和有效性，为后续的时序特征提取和模型训练提供数据支持。数据分别于2021年7月31日—2021年8月20日、2022年11月19日—2023年1月6日和2023年6月15日—2023年7月27日在广西某猪场、湖北某猪场进行采集。3个批次共采集61头猪只的数据，其中第一批次48头猪只，第二批次6头猪只，前2个批次数据用于模型的训练和验证，第三批次7头猪只用于本研究的模型测试。



A B

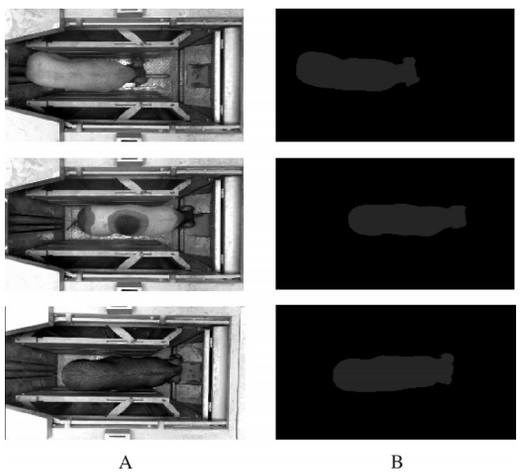
A: 采集装置图 Acquisition device diagram；B: 相机部署图 Camera deployment diagram.

图1 数据采集装置

Fig. 1 Data acquisition device

1.2 猪只视频数据集构建

本研究构建了一个猪只视频数据集，该数据集从3个视角捕捉猪只的动态生长过程，并依据猪只耳标信息分布进行组织，共包含234个视频文件7 103帧，分辨率为1 920像素×1 080像素，为了优化计算资源和提高视频对象分割模型的处理速度，每个视频样本分辨率被预处理至854像素×480像素，像素级别标注如图2所示。数据集被划分为三部分：训练集包含第一和第二批次51头猪只的148个视频文件，验证集包含第二批次3头猪只的30个视频文件，测试集包含第三批次7头猪只的56个视频文件，测试集视频文件以每3 d一个间隔，反映猪只动态生长过程。数据集结构如表1所示，其中204～514代表猪只耳标范围，表1中数据代表视频文件数量。



A: 猪只序列帧 Pig sequence frames；B: 像素级标注 Pixel level annotation.

图 2　像素级别标注样式

Fig. 2　Pixel level annotation style

表 1　猪只视频数据集组织形式

Table 1　Organizational form of pig video dataset

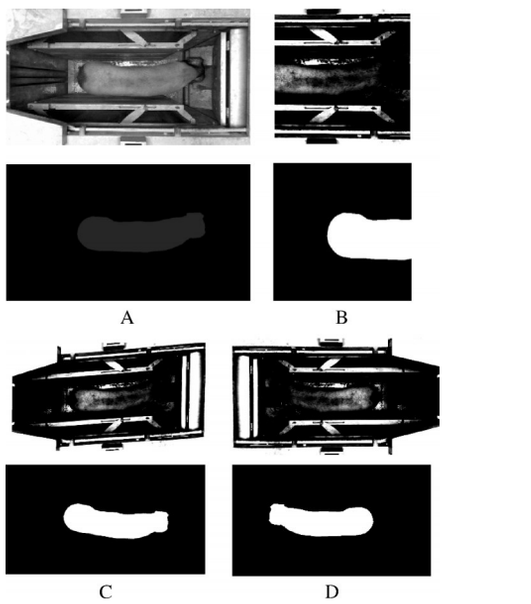
数据集 Dataset	耳标范围 Ear label range	不同相机视角 Different camera perspectives			共计 Total
		左视角 Left perspective	俯视角 Top perspective	右视角 Right perspective	
训练集 Train set	204～514	65	19	64	148
验证集 Valid set	515～530	0	30	0	30
测试集 Test set	535～934	21	35	0	56

1.3　数据增强

为提高模型对数据不同变化的适应性和泛化能力，本研究对序列帧和掩码图像进行数据增强操作，进一步增强模型的鲁棒性，包括对视频序列帧进行归一化操作，将数据的值范围调整到标准化范围内，加快模型的收敛速度；对数据的亮度、对比度和饱和度进行轻微的随机抖动；进行随机仿射变换，包括旋转、平移和缩放等；进行概率转换灰度图；进行随机大小裁剪，裁剪尺寸为 384 像素×384 像素的切片（patches）。数据增强如图 3 所示。

1.4　半监督视频对象分割网络设计及其改进

1）半监督视频对象分割模型设计。本研究引入猪只视频数据集到半监督视频对象分割领域，构建 XMem-SimAM 网络进行实验分析。该网络包括 3 个深度集成的卷积神经网络模块和 1 个多尺度计算相似度的模块：查询编码器、值编码器、解码器和记忆读取模块。查询编码器从输入的序列帧中提取多尺度时序语义特征，将特征输出到解码器，同时通过卷积映射的操作输出键值 k（key）、收缩项 s（shrink）和选择项 e（selection）到记忆读取模块，以便进行像素相似度计算；值编码器从预测分割的历史帧序列及其掩码图像序列中提取并融合多尺度时序特征，将这些特征传递到记忆读取模块，以进行更深层次的时序语义特征提取；记忆读取模块利用时空序列信息，并行接收查询编码器和值编码器的输出，计算视频序列帧与历史帧之间的相似度权重，为解码器生成高质量的掩码图像提供支持；解码器用于将查询编码器和值编码器的特征输出和记忆读取模块的特征输出并行融合，通过上采样转换成掩码图像。模型总体框架如图 4 所示。



A：序列帧和掩码图像 Sequence frames and mask images；B：随机裁剪 Random cropping；C：随机抖动 Random jitter；D：随机仿射变换 Random affine transformation.

图 3　数据增强示例

Fig. 3　Data augmentation example

2）查询编码器。查询编码器以 ResNet-50[25] 为骨干网络，首先将输入的查询帧（待分割的猪只视频序列帧）切分成 384 像素×384 像素的 patches，有助于在复杂环境中更好捕捉局部细节特征。为了捕捉视频序列中猪只动态变化和生长过程中的时序特征，XMem-SimAM 基于 4 层残差组（layer 1、layer 2、layer 3、layer 4）结构进行了优化调整，重构前 3 组的残差组（layer 1、layer 2、layer 3），同时为减少计算量、提高运算速度，丢弃了最后一组（layer 4）的残差块、平均池化（average pooling）和全连接层（fully-connection layer）。

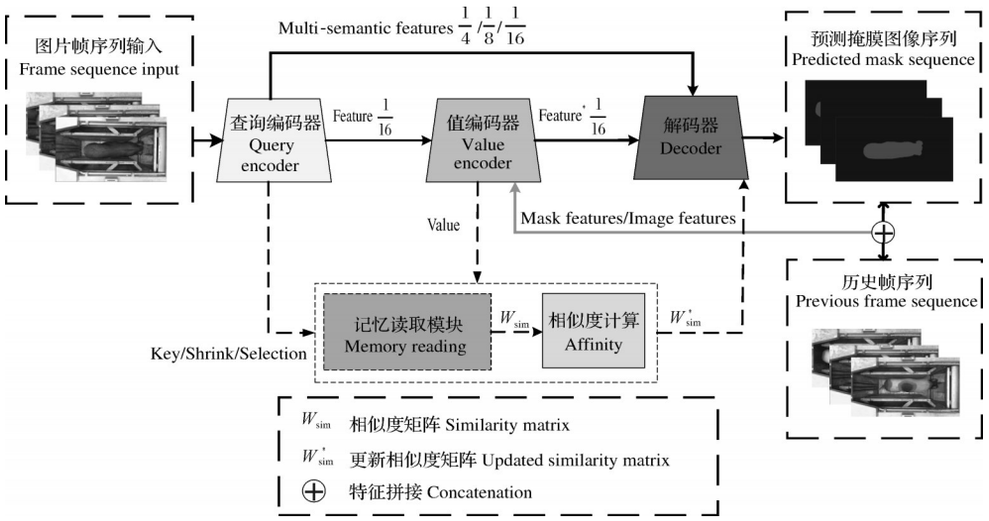


图4 网络模型总体框架图

Fig. 4 Overall framework diagram of network mode

SimAM注意力模块通过对每个层级的特征图进行空间重加权，使编码器更关注猪只序列帧像素间的差异性，更精细捕捉猪只在光照变化、复杂环境中的周期性生长特征。田甜等[26]通过在骨干网络引入SimAM模块显著增强复杂环境下特征的提取能力，并验证了该注意力处理机制的有效性。基于此，本研究在每个残差组(layer 1、layer 2、layer 3)后集成SimAM无参注意力机制(图5)，旨在优化猪只特征的空间分布并增强模型对猪只关键区域(耳朵、鼻子)的关注。通过精确定位识别耳朵和鼻子的位置，可以利用包络算法有效定位到颈部中点等关键体尺测点进行下游任务。SimAM注意力计算公式如下所示：

$$y_i=\frac{(x_i-\mu)^2}{4\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2+\lambda\right)}+0.5 \tag{1}$$

$$x_{\text{att}}=x\odot\sigma(y_i) \tag{2}$$

式(1)～(2)中，x_i、n、μ、λ分别表示输入特征图、特征图像素数、特征图均值、稳定归一化的常数，y_i、x_{att}表示特征图归一化、注意力加权后的输出特征图，$\odot$、$\sigma()$分别表示逐元素乘法、Sigmoid激活函数。

前3组残差块输出多尺度时序语义特征(1/4、1/8、1/16)输出到解码器用于猪只序列帧掩码图像预测，第三组残差块提取的1/16尺度特征送入到值编码器，进行深层次特征提取，查询编码器通过一个3×3卷积的特征映射(key projection)模块，保证提取的猪只局部特征分布不变，提取视频序列帧的收缩项s、选择项e和键值k，为计算视频序列帧与历史帧特征的相似度提供特征时序信息，具体公式如下：

$$k=W_k\cdot X+b_k \tag{3}$$

$$s=(W_d\cdot X+b_d)^2+1 \tag{4}$$

$$e=\sigma(W_e\cdot X+b_e) \tag{5}$$

式(3)～(5)中，X为视频序列帧时序特征；W_k、W_d、W_e为卷积层权重值；b_k、b_d、b_e为偏置项，$\sigma()$为Sigmoid激活函数。

3)值编码器。值编码器采用ResNet-18作为骨干网络处理已经分割的视频序列帧及其掩码图像。值编码器通过将视频序列帧和掩码融合特征进行拼接输出值v(value)，用于在记忆读取模块进行相似度计算。值编码器采用3层的残差组，为更精确融合猪只动态变化的时序特征，值编码器在每个残差组(layer 1、layer 2、layer 3)后集成了SimAM无参注意力机制。

此外，值编码器包含特征融合模块，将来自查询编码器的多尺度时序特征与值编码器最后一组残差组提取的1/16高尺度特征进行融合，提升时序特征表示。同时，特征融合模块引入了通道-空间注意力模块(convolutional block attention module，CBAM)[27]，通过全局平均池化(global average pooling，GAP)和全局最大池化(global max pooling，GMP)获取通道描述，通过1×1卷积和Sigmoid函数生成通道注意力权重。空间注意力通过GAP和GMP捕获特征响应，经7×7卷积和Sigmoid处理后生成空间注意力权重，专注于猪只边缘和细小部位的处理。值编码器结构如图6所示。

4)记忆读取模块。为充分利用猪只视频序列中的时空信息，增强模型在猪只的识别和分割方面的

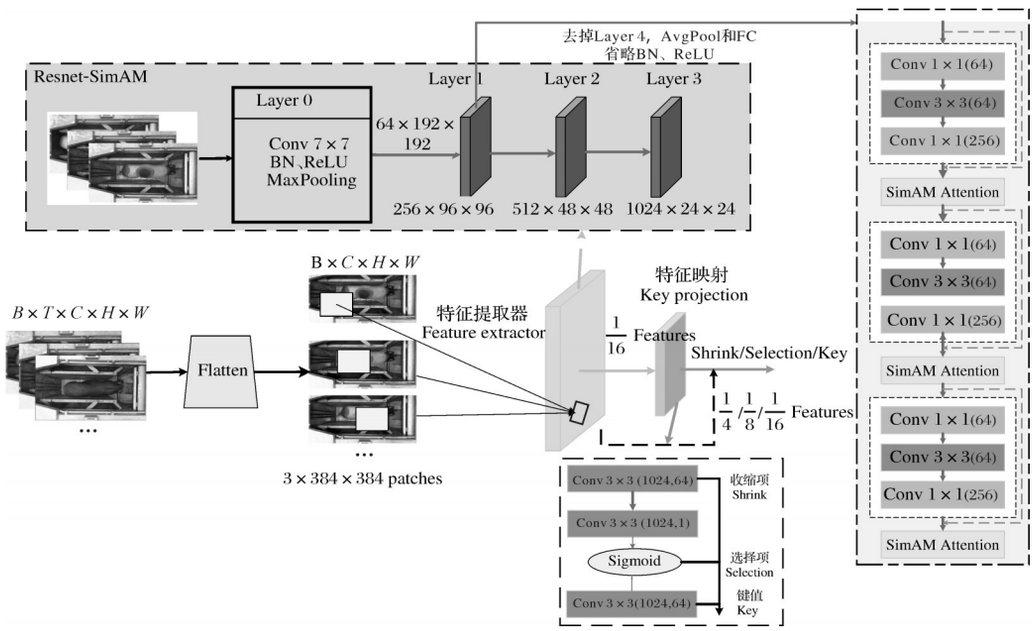


图 5　查询编码器框架图

Fig. 5　Query encoder framework diagram

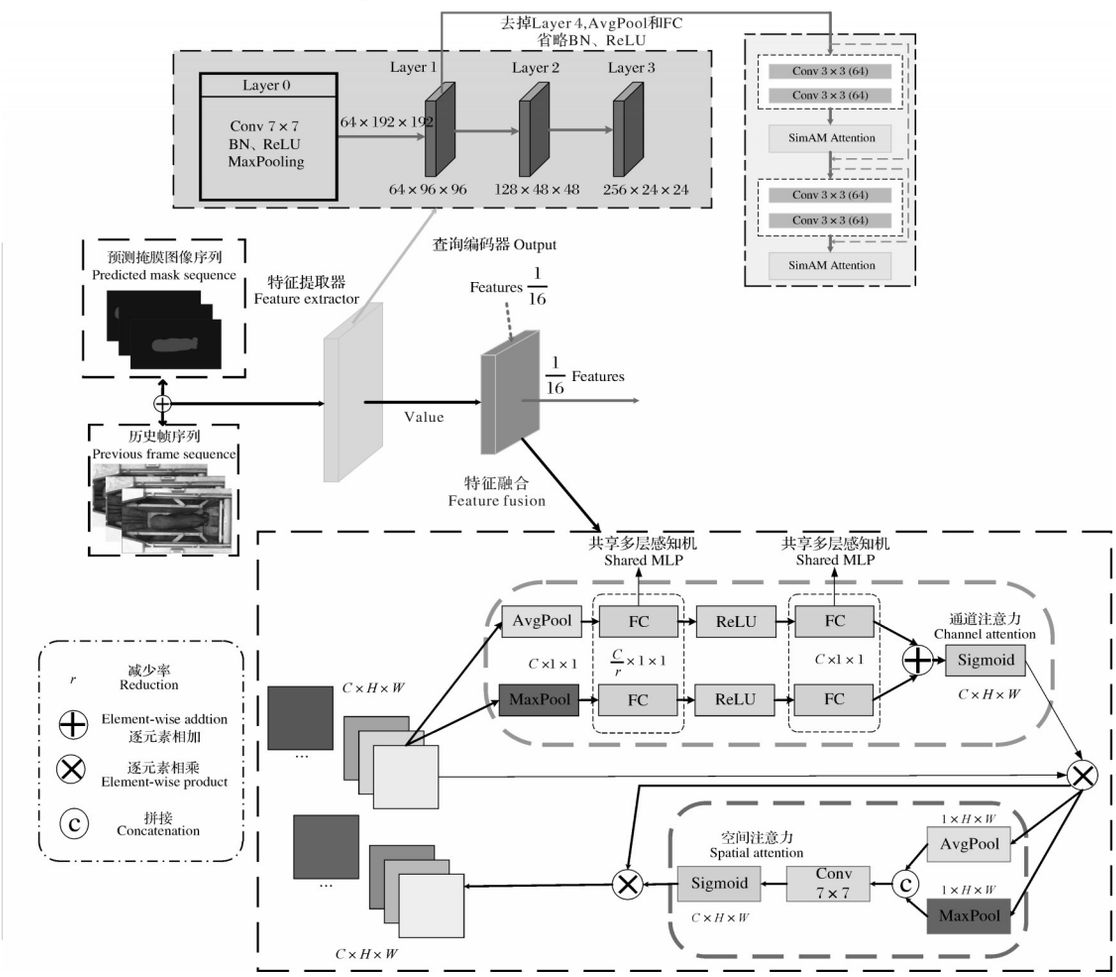


图 6　值编码器框架图

Fig.6　Value encoder framework diagram

能力，引入记忆读取单元。该单元接受5个主要输入，包括：查询编码器输出的键值k、收缩项s、选择项e以及值编码器输出的值v，这些特征分别对应视频序列帧和历史帧的特征信息。记忆读取模块采用相似度动态权重计算机制，有效捕捉视频序列中猪只移动和外观变化的依赖关系。通过计算视频序列帧和历史帧特征的相似度权重，可以识别出哪些视频序列帧像素特征需要重点关注，提高模型对猪只动态生长变化的响应能力，从而更准确生成预测掩码图像。相似度计算公式如下：

$$W(v,k)_{ij}=\frac{exp(a-2a\cdot b+b)\cdot s_i}{\sum_k exp(a-2a\cdot b+b)\cdot s_i} \tag{6}$$

$$a=\sum_{c}^{C^K}(v^2\cdot e_j) \tag{7}$$

$$b=\Sigma(\mathrm{e}_j\cdot k^2) \tag{8}$$

式(6)～(8)中，s、e、i、j分别表示查询编码器的收缩项、选择项、第i个内存帧元素、第j个查询帧元素，v、k、C^K、$exp()$分别表示值编码器键、查询编码器键、键通道数、指数函数。

5)解码器模块。为进一步提升猪只视频序列分割的精度和效率，本研究构建解码器模块。解码器整合查询编码器输出的多尺度(1/4、1/8、1/16)特征图、记忆读取模块的相似度权值以及值编码器的融合特征(1/16)。特征融合模块对其进行融合，以丰富猪只时序特征的表达。随后，通过2级上采样模块，融合后的特征逐步上采样至猪只视频帧相同的分辨率。在这一过程中，每一级上采样结合查询编码器的特征(1/4、1/8)，通过远眺连接(skip-connection)增强局部区域细节信息；上采样到原始分辨率后，特征图经3×3卷积处理，生成猪只视频序列掩码图像。为实现解码器轻量化并优化性能，本研究借鉴Cutie网络[28]解码器策略，通过减半上采样通道数，用转置卷积替代双线性插值的方法生成预测的掩码图像，适应猪只动态移动的数据处理需求。

1.5 损失函数

本研究采用Dice损失函数和Bootstrapped交叉熵损失函数相结合的方法，以优化模型在猪只视频数据集中的分割性能。

1)Dice损失函数。Dice Loss用于衡量2个样本相似度。在视频对象分割任务中，Dice损失弥补了像素级损失函数(如交叉熵损失)在处理小对象或者边界区域的不足，计算公式如下：

$$\mathrm{Dice}(y,\hat{y})=1-\frac{2\sum(y_i\cdot\hat{y}_i)+1}{\sum y_i+\sum\hat{y}_i+1} \tag{9}$$

式(9)中，y_i为真实标签，第i个像素的真实值；$\hat{y}_i$为预测标签，第i个像素的预测值。

2)Bootstrapped损失函数。在猪只视频对象分割任务中，采用Bootstrapped交叉熵损失函数，重点处理难以分割的样本，从而提高模型在复杂视频序列中对猪只的分割准确度，基础交叉熵损失和Bootstrapped交叉熵损失计算公式分别如式(10)～(11)所示：

$$\mathrm{CE}(y,\hat{y})=-\frac{1}{N}\sum_{i=1}^{N}(y_i\log(\hat{y}_i)+(1-y_i)\log(1-\hat{y}_i)) \tag{10}$$

$$\mathrm{BootstrappedCE}(y,\hat{y})=\frac{1}{k}\sum_{i=1}^{k}\mathrm{CE}(y_i,\hat{y}_i) \tag{11}$$

式(10)～(11)中，y_i为真实标签，第i个像素的真实值；$\hat{y}_i$为预测标签，第i个像素的预测值；N为样本总数；k为选取样本数量。

1.6 操作环境

本研究在Windows11操作系统下，使用Pytorch框架进行，核心硬件为2块24 GB显存的GeForce GTX3090Ti显卡，CUDA版本为11.5，Python版本为3.10.12，Pytorch版本为1.11.0，开发环境为Pycharm 2022.3.3。

1.7 参数设置

实验过程中模型训练轮数为10 000轮，初始学习率为1×10^{-5}，在8 000轮次，学习率衰减为原来的1/10；优化器采用AdamW，权重衰减系数为0.05，参考帧(reference-frames)为3，训练批次(batch-size)为8，1个训练批次视频序列帧为8。

1.8 评估指标

本研究使用区域相似度J(region similarity Jaccard)、轮廓准确度F(contour accuracy，F)、Dice系数以及平均度量值$J\&F$，对模型效果进行评估。

区域相似度J和Dice系数计算公式为：

$$J=\frac{\hat{M}\cap M}{\hat{M}\cup M}\times100 \tag{12}$$

$$\mathrm{Dice}=\left(2\times\frac{\hat{M}\cap M}{\hat{M}+M}\right)\times100 \tag{13}$$

式(12)～(13)中，$\hat{M}$为预测的掩码图像，M为真实标签。

轮廓准确度F计算公式为：

$$F=\frac{2P_cR_e}{P_c+R_e}\times100 \tag{14}$$

式(14)中，P_c为轮廓精确度，R_e为召回率。

平均度量值$J\&F$计算公式为：

$$J\&F=\frac{J+F}{2} \tag{15}$$

其中区域相似度J（也称为交并比，intersection over union，IoU）用于评估预测分割区域与真实区域的相似性。J值越接近100，反映模型对猪只分割性能越好；轮廓准确度F是基于精确度（precision）和召回率（recall）的调和平均，用于评估预测分割轮廓与真实轮廓的吻合程度。F值越接近100，表明模型在捕捉猪只边缘细节方面表现出色；平均度量值$J\&F$是J和F的平均值，值越接近100，表示模型在整体分割和边缘细节上均表现良好；Dice系数则评估真实标签与预测掩码图像的重叠程度，采用百分比的形式省略百分号。值越接近100，表明模型的分割精确度越高。

2　结果与分析

2.1　XMem网络分割效果的对比分析

XMem网络和XMem-SimAM网络在猪只视频数据集上的分割结果如表2所示，XMem-SimAM网络在各项评估指标均优于XMem网络。验证集上的J、F和$J\&F$评估指标分别达到了97.0、96.1和97.8，Dice系数为98.0；测试集上，J、F和$J\&F$评估指标分别达到了96.9、95.8和98.0，Dice系数为98.0。XMem-SimAM在验证集上的性能平均提升了0.3、0.2、0.3和0.1，在测试集上则分别提升了0.6、0.6、0.7和0.5。在处理速度（frames per second，FPS）方面，XMem-SimAM网络在验证集和测试集上的帧率分别达到了54.85帧/s和58.50帧/s，相比之下，XMem网络的帧率为51.43帧/s和56.59帧/s，显示出更快的处理速度。这些结果证明了XMem-SimAM网络在处理猪只视频数据集时的高效性和优异的分割效果。

表2　网络评估指标对比

Table 2　Comparison of evaluation indicators for validation set networks

评估指标 Evaluation index	XMem		XMem-SimAM	
	验证集 Valid set	测试集 Test set	验证集 Valid set	测试集 Test set
$J\&F$	96.7	96.3	97.0	96.9
J(Jaccard)	95.9	95.2	96.1	95.8
F	97.5	97.3	97.8	98.0
Dice	97.9	97.5	98.0	98.0
处理速度/(帧/s) FPS	51.43	56.59	54.85	58.50

猪只视频数据集501耳标猪只视频序列的分割对比结果如图7所示，XMem-SimAM网络在猪只视频对象分割任务中具有更强的泛化能力和更高的鲁棒性。XMem-SimAM网络在猪只耳朵等部位的分割效果优于原始网络，掩码图像的轮廓和细节更清晰、更准确，与真实标签吻合度更高。XMem-SimAM网络能够更好地捕捉猪只的边界轮廓，特别是在动态变化和复杂环境中，提高了分割精度。

2.2　视频对象分割模型对比分析

为全面评估XMem-SimAM网络在猪只视频数据集上的泛化能力，本研究将其与几种先进视频对象分割模型进行对比。为了确保评估的公平性，MiVOS、STCN、DEVA以及XMem++模型均采用相同的超参数设置和固定的训练模式。

不同半监督视频对象分割模型在猪只视频数据集上的评估指标得分、处理速度、最大内存分配（max allocated memory）、模型复杂度和计算复杂度如表3、表4所示，XMem-SimAM的评估指标优于其他对比模型。XMem-SimAM在验证集和测试集上的$J\&F$、J、F和Dice系数比STCN网络分别高1.5、1.3、1.6、0.7和1.3、1.6、1.0、1.0；比DEVA网络分别高1.1、1.3、0.7、0.7和0.9、1.4、0.3、0.9；比MiVOS网络分别高1.7、1.5、1.8、0.8和1.5、1.6、1.4、1.0；比XMem++网络分别高2.0、2.3、1.7、1.2和2.6、3.1、2.1、1.8，表明XMem-SimAM在处理具有动态变化和复杂背景的猪只数据时，能够保持高度的分割准确性和稳定性，展示出明显的优势。

在模型的运行效率与资源占用方面，XMem-SimAM网络虽然在处理速度和内存分配方面并非最优，但其整体评估指标仍十分突出。在处理速度方面，XMem-SimAM在验证集上的达到了54.85帧/s，测试集为58.5帧/s，虽然略低于XMem++，但显著高于DEVA和MiVOS，表明XMem-SimAM在保持较高分割准确度的同时也具备较高的处理速度，能够满足实时处理的需求。最大内存分配方面XMem-SimAM在验证集上为836 MB，测试集为795 MB，这一数据低于STCN和MiVOS，略高于DEVA，与XMem++相当，表明XMem-SimAM在处理效率和资源利用具有良好的平衡性既能确保较高的分割精度，又能有效控制资源的占用。

XMem-SimAM在模型复杂度和计算复杂度的对比中展现出良好的平衡。尽管该方法的计算复杂度和参数量并非最低，但在所有评估指标上依然表现最佳，有效平衡了性能与资源消耗，实现了性能与资源消耗的双重优化。

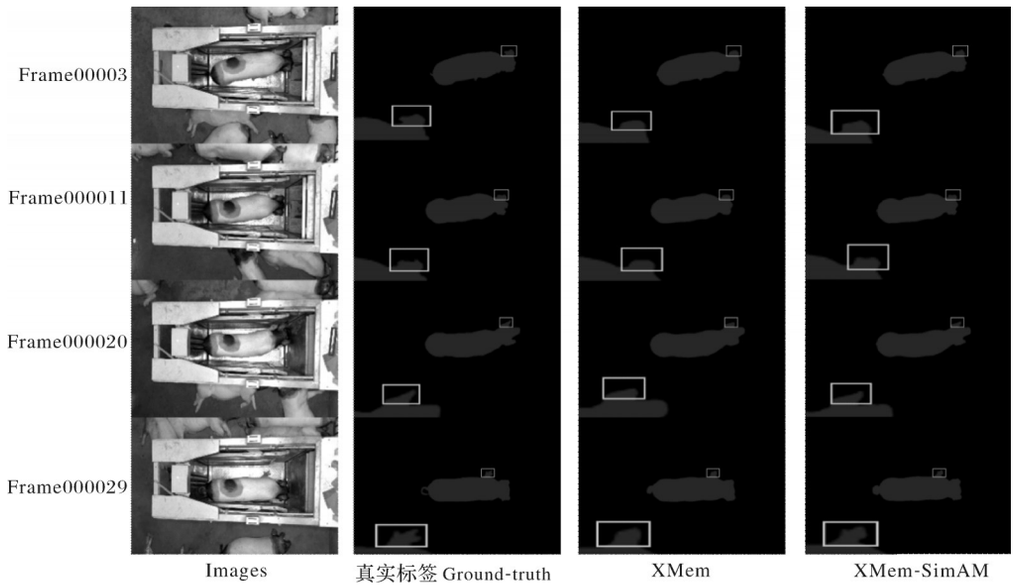


图7 分割效果对比

Fig. 7 Comparison of segmentation effects

表3 评估指标对比

Table 3 Comparison of evaluation indicators

评估指标 Evaluation index	XMem-SimAM		STCN[21]		DEVA[22]		MiVOS[23]		XMem++[24]	
	验证集 Valid set	测试集 Test set	验证集 Valid set	测试集 Test set	验证集 Valid set	测试集 Test set	验证集 Valid set	测试集 Test set	验证集 Valid set	测试集 Test set
J&*F*	**97.0**	**96.9**	95.5	95.6	95.9	96.0	95.3	95.4	95.0	94.3
J(Jaccard)	**96.1**	**95.8**	94.8	94.2	94.8	94.4	94.6	94.2	93.8	92.7
F	**97.8**	**98.0**	96.2	97.0	97.1	97.7	96	96.6	96.1	95.9
Dice	**98.0**	**98.0**	97.3	97.0	97.3	97.1	97.2	97.0	96.8	96.2
处理速度/(帧/s) FPS	54.85	58.50	51.83	57.90	47.39	53.81	34.67	35.90	**67.71**	**67.31**
最大内存分配/MB Max allocated memory	836	795	1 251	1 153	**724**	**671**	1 240	1 160	839	799

表4 复杂度对比

Table 4 Comparison of complexity

复杂度 Complexity	XMem-SimAM	STCN[21]	DEVA[22]	MiVOS[23]	XMem++[24]
模型参数 Model parameters	53.71×10^6	54.43×10^6	53.68×10^6	40.13×10^6	62.19×10^6
FLOPs	348.42×10^9	355.74×10^9	349.57×10^9	260.73×10^9	430.42×10^9

通过对比分析XMem-SimAM与STCN、DEVA、MiVOS以及XMem++半监督视频对象分割模型的预测掩码图像，进一步验证了XMem-SimAM在猪只视频序列帧分割任务中的优势。272耳标猪只视频序列帧分割结果如图8所示，XMem-SimAM在耳朵、鼻子等关键细节部位分割效果展现了卓越的性能。相较之下，其他视频对象分割对比模型在处理这些细节时出现漏分割的现象，这些区域的高精度分割为后续猪只体尺测量等下游任务奠定了基础。

为了深入分析模型在小目标(如耳朵、鼻子、尾巴等关键部位)上的分割性能，研究对比了XMem-SimAM与其他对比模型在这些区域的表现。图9展示了各模型在猪只头部和尾部关键部位的分割结果。结果显示，XMem-SimAM在耳朵和鼻子等小目标区域的分割效果显著优于其他模型，能够更准确地捕捉这些关键部位的细节。这种高精度的分割对于后续的猪只体尺测量至关重要，确保了测量的准确性和可靠性。相较之下，其他模型在这些细节处理上表现出漏分割或边界模糊的现象，进一步证明了XMem-SimAM在细节处理上的优势。

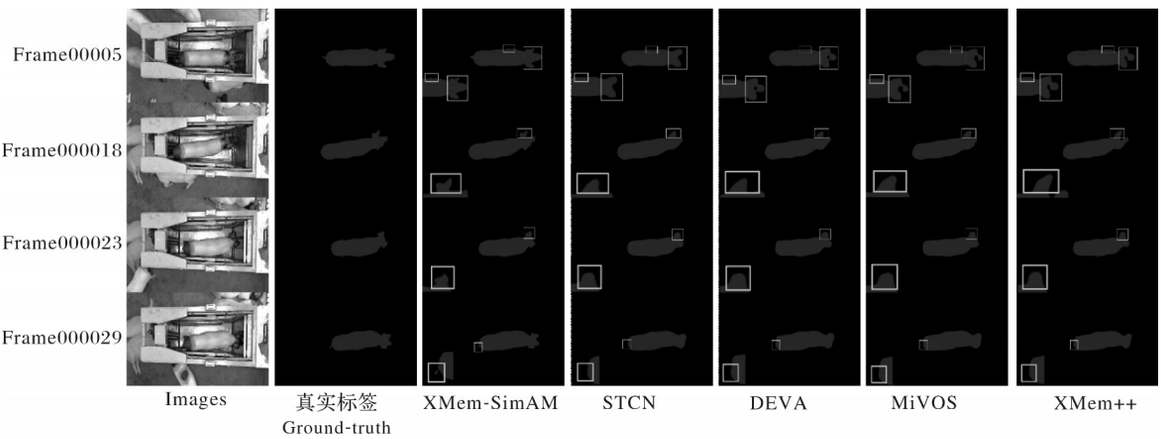


图 8　XMem-SimAM 与半监督视频对象分割网络的分割效果对比

Fig. 8　Comparison of segmentation performance between XMem-SimAM and semi-supervised video object segmentation network

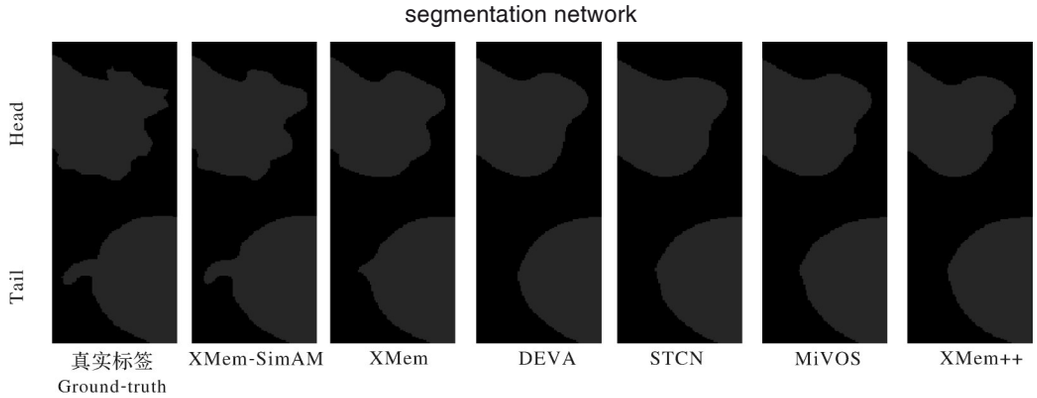


图 9　小目标区域分割结果对比

Fig. 9　Comparison of segmentation results for small target areas

如表 5 所示，XMem-SimAM 在小目标区域分割任务上取得了最优性能，平均指标达到 97.18。相比于 XMem 方法提升了 0.6，相比 DEVA、STCN、MiVOS 和 XMem＋＋等方法分别提升了 0.88、1.23、1.38 和 2.40。这些量化结果进一步验证了 XMem-SimAM 在处理小目标区域分割时的卓越表现，特别是在需要精细分割的关键部位上具有明显优势。

表 5　小目标分割性能对比

Table 5　Comparison of small target segmentation performance

模型 Model	平均指标 Average index	提升 Improvement
XMem-SimAM	97.18	/
XMem	96.58	＋0.60
DEVA	96.30	＋0.88
STCN	95.95	＋1.23
MiVOS	95.80	＋1.38
XMem++	94.78	＋2.40

2.3　消融实验

在 XMem 模型的基础上，本研究通过引入 SimAM 无参注意力和多尺度注意力特征融合模块，进行消融实验，以评估其对分割性能的提升效果。结果如表 6 所示，SimAM 无参注意力模块在几乎不增加模型参数量和仅略微增加 FLOPs 的情况下，使测试集上的 $J\&F$、J、F 和 Dice 系数均提升了 0.2。此外，通过多尺度注意力特征融合模块的引入，测试集上的 $J\&F$、J、F 和 Dice 系数分别较原始模型提升了 0.4、0.4、0.4 和 0.3。综合来看，XMem-SimAM 模型在显著提升分割精度的同时，保持了参数量的合理控制，以较低的计算代价实现了性能的优化，充分证明了本研究所集成的模块在提升模型性能方面发挥了关键作用。

2.4　特征图可视化分析

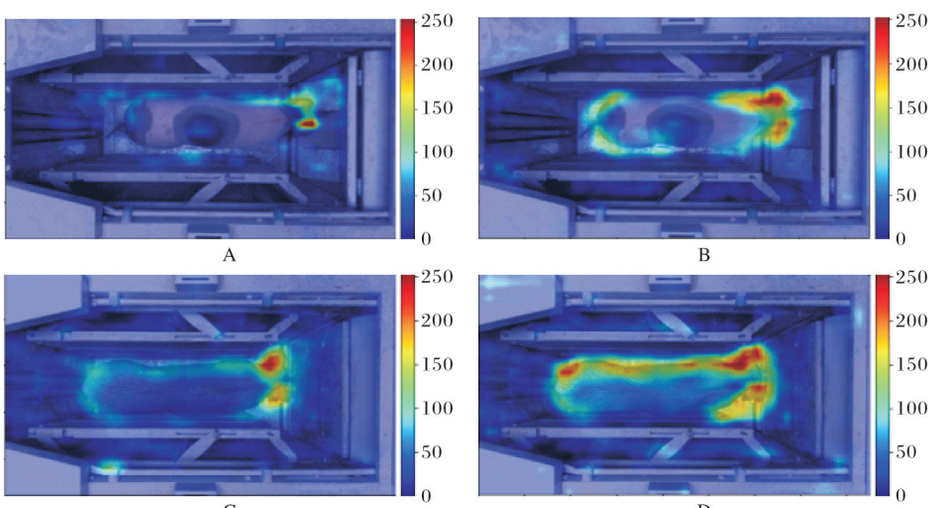
为了深入理解 XMem-SimAM 在猪只视频对象分割任务中的性能表现，本研究采用 Grad-CAM[29]（gradient-weighted class activation mapping，Grad-CAM）进行模型特征图可视化。Grad-CAM 作为一种可视化解释工具，能够揭示网络在作出决策时关注的关键区域。

表6 消融实验结果对比

Table 6 Ablation study results comparison

模型 Model	验证集 Valid set				测试集 Test set				模型参数量 Model parameters	FLOPs
	J&*F*	*J*	*F*	Dice	*J*&*F*	*J*	*F*	Dice		
XMem	96.7	95.9	97.5	97.9	96.3	95.2	97.3	97.5	50.10×10^6	320.00×10^9
＋原始SimAM模块 Original SimAM parameter-free attention mechanism	96.8	96.0	97.6	97.9	96.5	95.4	97.5	97.7	50.10×10^6	320.02×10^9
＋多尺度特征融合模块 Multi-scale feature fusion module	96.9	96.1	97.7	98.0	96.7	95.6	97.6	97.8	53.71×10^6	340.20×10^9
XMem-SimAM	97.0	96.1	97.8	98.0	96.9	95.8	98.0	98.0	53.71×10^6	348.42×10^9

由图10可知，XMem-SimAM在猪只轮廓的边缘区域展现出更加明显的激活区域，尤其是在耳朵和轮廓区域，颜色越接近深红，表示该区域对模型预测的贡献越大，表明模型能够有效地捕捉动态变化的细节特征。通过进一步的热力图对比分析，SimAM注意力机制在耳朵和轮廓区域的关注程度更深，反映了SimAM在细节处理上的显著提升，结果验证了改进策略在猪只视频对象分割任务中的有效性，也表明SimAM注意力机制在捕捉动态变化特征和细节处理中的关键作用。



A:XMem网络查询编码器;B:XMem-SimAM网络查询编码器;C:XMem网络值编码器;D:XMem-SimAM网络值编码器。A:XMem network query_encoder;B:XMem-SimAM network query_ encoder;C:XMem network value_encoder;D:XMem-SimAM network value_encoder

图10 Grad-CAM特征可视化

Fig. 10 Grad-CAM feature visualization

3 讨 论

本研究基于测定站内动态生长的猪只视频，构建了一个猪只视频数据集，提出了一种基于半监督学习的猪只视频分割模型XMem-SimAM，该模型通过引入无参注意力机制和多尺度特征融合方法改进编码器-解码器架构，有效捕获猪只视频序列的时序特征。结果表明：XMem-SimAM模型在*J*、*F*、*J*&*F*和Dice系数上分别达到96.9、95.8、98.0和98.0，推理阶段处理速度达到58.5帧/s，内存消耗为795 MB，展现了其在处理动态移动和生长的猪只视频数据的卓越性能。此外，热力图特征可视化和消融实验进一步验证了所提出的改进策略在提升猪只视频分割任务性能方面的显著效果。综上所述，XMem-SimAM能够有效捕获猪只细粒度特征，并实现局部关键区域的高精度分割，使其成为猪只表型测试下游任务的有力工具。

然而本研究仍存在一些不足。首先，模型的复杂度和浮点运算量相对较高，可通过模型剪枝、知识蒸馏和量化等轻量化方式进一步优化，以扩展适应资源受限的边缘设备。其次，当前数据集主要来自测定站内的受控环境，后续研究将扩大数据集规模，增加不同品种和环境下的样本，探索模型轻量化方法，并将模型应用于更复杂的实际养殖场景，进一步

验证其在变化光照、高密度群养等条件下的性能。

参考文献 References

[1] WANG Z Y, SHADPOUR S, CHAN E, et al. ASAS-NANP SYMPOSIUM: Applications of machine learning for livestock body weight prediction from digital images[J/OL]. Journal of animal science, 2021, 99(2): skab022[2024-12-16]. https://doi.org/10.1093/jas/skab022.

[2] 王德福, 黄会男, 张洪建, 等. 生猪养殖设施工程技术研究现状与发展分析[J]. 农业机械学报, 2018, 49(11): 1-14. WANG D F, HUANG H N, ZHANG H J, et al. Analysis of research status and development on engineering technology of swine farming facilities[J]. Transactions of the CSAM, 2018, 49(11): 1-14 (in Chinese with English abstract).

[3] NASIRAHMADI A, EDWARDS S A, STURM B. Implementation of machine vision for detecting behaviour of cattle and pigs[J]. Livestock science, 2017, 202: 25-38.

[4] 耿艳利, 季燕凯, 岳晓东, 等. 基于点云语义分割的猪只体尺测量方法研究[J]. 农业机械学报, 2023, 54(7): 332-338. GENG Y L, JI Y K, YUE X D, et al. Pigs body size measurement based on point cloud semantic segmentation[J]. Transactions of the CSAM, 2023, 54(7): 332-338 (in Chinese with English abstract).

[5] WANG K, GUO H, MA Q, et al. A portable and automatic Xtion-based measurement system for pig body size[J]. Computers and electronics in agriculture, 2018, 148: 291-298.

[6] XIE Q J, WU M R, BAO J, et al. A deep learning-based detection method for pig body temperature using infrared thermography[J]. Computers and electronics in agriculture, 2023, 213: 108200 [2024-12-16]. https://doi. org/10.1016/j. compag.2023.108200.

[7] ZHANG Z Q, ZHANG H, LIU T H. Study on body temperature detection of pig based on infrared technology: a review[J]. Artificial intelligence in agriculture, 2019, 1: 14-26.

[8] HE W, MI Y, DING X D, et al. Two-stream cross-attention vision Transformer based on RGB-D images for pig weight estimation [J/OL]. Computers and electronics in agriculture, 2023, 212: 107986 [2024-12-16]. https://doi. org/10.1016/j. compag.2023.107986.

[9] KWON K, PARK A, LEE H, et al. Deep learning-based weight estimation using a fast-reconstructed mesh model from the point cloud of a pig[J/OL]. Computers and electronics in agriculture, 2023, 210: 107903[2024-12-16]. https://doi.org/10.1016/j.compag.2023.107903.

[10] CHO S, LEE M, LEE S, et al. Treating motion as option to reduce motion dependency in unsupervised video object segmentation[C]//2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). January 2-7, 2023, Waikoloa, HI, USA. Waikoloa: IEEE, 2023: 5129-5138.

[11] PERAZZI F, PONT-TUSET J, MCWILLIAMS B, et al. A benchmark dataset and evaluation methodology for video object segmentation[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). June 27-30, 2016, Las Vegas, NV, USA. Las Vegas: IEEE, 2016: 724-732.

[12] LEE M, CHO S, LEE S, et al. Unsupervised video object segmentation via prototype memory network[C]//2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). January 2-7, 2023, Waikoloa, HI, USA. Waikoloa: IEEE, 2023: 5913-5923.

[13] JAIN S D, GRAUMAN K. Supervoxel-consistent foreground propagation in video [C]//Computer Vision-ECCV 2014. Cham: Springer International Publishing, 2014: 656-671.

[14] PREST A, LEISTNER C, CIVERA J, et al. Learning object class detectors from weakly annotated video[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. June 16-21, 2012, Providence, RI, USA. Providence: IEEE, 2012: 3282-3289.

[15] CHO S, LEE M, LEE J, et al. Improving unsupervised video object segmentation via fake flow generation[DB/OL]. arXiv, 2024: 2407.11714 [2024-12-16]. https://doi. org/10.48550/arXiv.2407.11714.

[16] XU N, YANG L J, FAN Y C, et al. YouTube-VOS: sequence-to-sequence video object segmentation[C]// Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 603-619.

[17] YANG Z X, WEI Y C, YANG Y, et al. Associating objects with transformers for video object segmentation [DB/OL]. arXiv, 2021: 2106.02638 [2024-12-16]. https://doi. org/10.48550/arXiv.2106.02638.

[18] 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.

[19] CHENG H K, SCHWING A G. XMem: long-term video object segmentation with an Atkinson-Shiffrin memory model [C]//Computer Vision-ECCV 2022. Cham: Springer Nature Switzerland, 2022: 640-658.

[20] YANG L, ZHANG R Y, LI L, et al. SimAM: a simple, parameter-free attention module for convolutional neural networks[C]//Proceedings of the 38th International Conference on Machine Learning. [S.l.]: PMLR, 2021: 11863-11874.

[21] CHENG H K, TAI Y W, TANG C K, et al. Rethinking space-time networks with improved memory coverage for efficient video object segmentation[DB/OL]. arXiv, 2021: 2106.05210 [2024-12-16]. https://doi.org/10.48550/arXiv.2106.05210.

[22] CHENG H K, OH S W, PRICE B, et al. Tracking anything with decoupled video segmentation[C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). October 1-6, 2023, Paris, France. Paris: IEEE, 2023: 1316-1326.

[23] CHENG H K, TAI Y W, TANG C K. Modular interactive video object segmentation: interaction-to-mask, propagation and difference-aware fusion [C]//2021 IEEE/CVF Confer-

ence on Computer Vision and Pattern Recognition (CVPR). June 20-25, 2021, Nashville, TN, USA. Nashville: IEEE, 2021:5555-5564.

[24] BEKUZAROV M, BERMUDEZ A, LEE J Y, et al. XMem: production-level video segmentation from few annotated frames [C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). October 1-6, 2023, Paris, France. Paris: IEEE, 2023:635-644.

[25] HE K, ZHANG X, REN S, 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, Nevada, USA. Las Vegas: IEEE, 2016: 770-778.

[26] 田甜,程志友,鞠薇,等.基于SimAM-ConvNeXt-FL的茶叶病害小样本分类方法研究[J].农业机械学报,2024,55(3):275-281. TIAN T, CHENG Z Y, JU W, et al. Small sample classification of tea diseases based on SimAM-ConvNeXt-FL [J]. Transactions of the CSAM, 2024, 55(3): 275-281 (in Chinese with English abstract).

[27] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]// Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018:3-19.

[28] CHENG H K, OH S W, PRICE B, et al. Putting the object back into video object segmentation [C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 16-22, 2024, Seattle, WA, USA. Seattle: IEEE, 2024:3151-3161.

[29] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization [J]. International journal of computer vision, 2020, 128(2): 336-359.

XMem-SimAM based semi-supervised video segmentation of pigs

CHEN Mengfang[1,2], XU Dihong[2,3], LI Guoliang[1,2], LIU Xiaolei[4], ZHOU Mingyan[2,3], LI Xuan[2,3]

1.*College of Informatics, Huazhong Agricultural University, Wuhan 430070, China*;
2.*Ministry of Agriculture and Rural Affairs Key Laboratory of Smart Farming for Agricultural Animals, Wuhan 430070 China*; 3.*College of Engineering, Huazhong Agricultural University, Wuhan 430070, China*; 4.*Hubei Hongshan Laboratory, Wuhan 430070, China*

Abstract The dynamic feeding and growth process of breeding pigs during the performance testing was used to solve the problem of accurate segmentation of pigs caused by complex environments in pig farms, dynamic growth of pigs, and changes in body size. A pig video dataset consisting of 234 video sequences was constructed. A XMem-SimAM based semi-supervised video segmentation of pigs was proposed. The ability of model to extract temporal information at different scales was improved and the temporal features of pigs' dynamic movements were captured by introducing SimAM attention for multi-scale feature fusion. The spatial-channel attention module was used to enhance the model's extraction of temporal semantic feature weights. The strategy for multi-scale feature fusion and upsampling module were optimized. The temporal correlation information in video sequences was fully utilized to improve the segmentation accuracy of pigs in videos at a fine-grained level. The results of testing and comparison showed that the Jaccard index, contour accuracy F-score, average metric $J\&F$, and the Dice coefficient of of XMem-SimAM model on the pig video dataset was 96.9, 95.8, 98.0, and 98.0, superior to that of video object segmentation methods including MiVOS, STCN, DEVA, and XMem++, demonstrating its outstanding performance of segmentation. The processing speed reached 58.5 frames per second, with a memory consumption of 795 MB at the stage of reasoning, achieving a good balance between the efficiency of processing and the utilization of resource. The proposed method can be applied to video segmentation of dynamically growing pigs in the complex environments of a pig farm.

Keywords semi supervised; video segmentation; pigs; SimAM Attention

(责任编辑:葛晓霞)