深度信号引导学习混合变换器的高性能无监督视频目标分割

苏天康[1,2],宋慧慧[1,2],樊佳庆[3],张开华[1,2]
(1. 南京信息工程大学江苏省大数据分析技术重点实验室,江苏南京 210044;
2. 南京信息工程大学大气环境与装备技术协同创新中心,江苏南京 210044;
3. 南京航天航空大学计算机与科学技术学院,江苏南京 211106)

摘 要: 现存的无监督视频目标分割方法通常使用光流作为运动线索来提升模型性能. 然而,光流的估计常存在误差,这将导致双流网络易对噪声过拟合. 为此,本文提出一种基于混合变换器的无监督视频目标分割算法,通过引入深度信号引导变换器高效融合不同模态数据,以学习更加鲁棒的特征表达,从而减轻模型对噪声的过拟合. 首先,设计一个新颖的混合注意力模块来获得全局感受野并对不同模态的特征进行充分交互,以增强特征的全局语义信息来提升模型的抗干扰能力. 接着,为了进一步感知精细化的目标边缘,设计了一个局部-非局部语义增强模块,将局部语义的归纳偏置引入补充学习非局部语义特征,在提升模型抗干扰力的同时突出更精细化的目标区域. 最后,增强后的特征输入变换器的解码器,预测得到高质量的分割结果. 与最先进的方法相比,本文所提算法在四个标准数据集上都获得了领先的性能,充分表明了本文所提方法的有效性.

关键词: 无监督视频目标分割;混合变换器;混合注意力;多模态;深度估计;鲁棒特征
基金项目: 科技创新 2030-"新一代人工智能"重大项目(No.2018AAA0100400);国家自然科学基金(No.62276141, No.U20B2065)
中图分类号: TP391.41 **文献标识码:** A **文章编号:** 0372-2112(2023)05-1388-08
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20221162

Learning Depth Signal Guided Mixed Transformer for High-Performance Unsupervised Video Object Segmentation

SU Tian-kang[1,2], SONG Hui-hui[1,2], FAN Jia-qing[3], ZHANG Kai-hua[1,2]
(1. *Jiangsu Key Laboratory of Big Data Analysis Technology*, *Nanjing University of Information Science and Technology*, *Nanjing*, *Jiangsu* 210044, *China*;
2. *Collaborative Innovation Center on Atmospheric Environment and Equipment Technology*, *University of Information Science and Technology*, *Nanjing*, *Jiangsu* 210044, *China*;
3. *College of Computer Science and Technology*, *Nanjing University of Aeronautics and Astronautics*, *Nanjing*, *Jiangsu* 211106, *China*)

Abstract: The existing unsupervised video object segmentation methods usually employ optical flow as a motion cue to improve the model performance. However, the estimation of optical flow frequently involves errors, resulting in lots of noise, especially for objects with static or complicated motion interference. The two-stream networks will easily overfit to the noise, which severely degrades the segmentation model. To relieve this, we propose to a novel mixed transformer in unsupervised video object segmentation, which can efficiently fuse different modality data by introducing depth signals to learn more robust feature representation and reduce the model overfitting to noise. In specific, the video frame, optical flow and depth map that are cropped into a set of fixed-size patches and concatenated together, are first composed of a triplet as the transformer input. The linear layer followed by a position-encoding layer is applied on the triplet, producing the features to be encoded. After this, the features are integrated by a novel mixed attention module, which can obtain the global respective field and sufficiently interact with the various modality features, to enhance the global semantic features and improve the anti-interference ability of the model. The local-non-local semantic enhancement module is developed in order to

收稿日期:2022-10-11;修回日期:2023-01-18;责任编辑:覃怀银

further perceive the refined target edge by introducing the inductive bias of local semantic information into supplementary learning of non-local semantic features. In this way, the target region is more refined while improving the anti-interference capability of the model. In the end, the enhanced features as the transformer decoder input to produce the predicted segmentation mask. Extensive experiments on four standard challenging benchmarks demonstrate that the proposed method achieves favorable performance against state-of-the-art methods.

Key words: unsupervised video object segmentation mixed transformer; mixed attention; multimodality; depth estimation; robust features

Foundation Item(s): National Key Research and Development Program of China (No.2018AAA0100400); National Natural Science Foundation of China (No.62276141, No.U20B2065)

1 引言

给定一组视频序列,视频目标分割(Video Object Segmentation,VOS)旨在从中定位并分割出特定的目标.VOS在计算机视觉中是一项极具挑战性的任务,拥有目标跟踪和自动驾驶等应用场景[1-3]现有的VOS技术大致可分为三类:半监督视频目标分割、无监督视频目标分割(Unsupervised VOS,UVOS)和参考视频目标分割.

UVOS任务的主流方法包括双流网络[4-8]、基于记忆的卷积神经网络[9,10]和3D卷积神经网络[11,12]等.其中,双流网络方法利用光流捕获运动信息,并通过信息交互模块融合外观与运动特征,以获得增强后的时空特征.基于记忆的UVOS方法利用过往所有的历史帧信息,将当前帧与历史帧在时空域上做匹配学习,从而关联当前帧和历史帧的特征.基于3D卷积神经网络的方法将连续视频帧沿着时间维度拼接,再输入3D卷积模块中提取时空特征.

尽管上述UVOS方法取得了较好的性能,但是仍然存在一些不足之处:双流网络结构复杂且计算开销较大,容易对光流噪声信号过拟合;3D卷积网络计算量大,并且感受野受限,无法捕获特征之间的长程依赖关系;基于记忆的UVOS方法需要大量的内存开销来存储历史帧信息.

基于上述分析,本文提出了基于深度信号引导学习混合变换器的UVOS网络.

2 预备知识

为了解决模型对光流噪声过拟合的问题,本文在网络中引入深度信号.直观上,深度信息提供了一帧中所有对象的空间位置信息,而不像光流会受物体静止或复杂运动的干扰.通过深度图,我们可以很容易地区分出不同空间位置的物体,从而精确分割出运动目标区域.此外,针对卷积神经网络中局部感受野受限和网络复杂、计算量大的问题,轻量化的视觉变换器(Vision Transformer,ViT)能够以较小的计算量和参数量来建模长程依赖关系,从而避免引入卷积神经网络的归纳偏置而获得全局感受野.最后,变换器本身具备良好的处理多模态信号的特性,特别适合建模视频帧、光流图、深度图等不同模态数据之间的相互依赖关系,从而高效挖掘出它们之间的共性信息.

3 基于深度信号引导学习混合变换器

3.1 网络整体结构

基于深度信号引导学习混合变换器网络主要包括Transformer编码器和Transformer解码器.在每个Transformer模块中包含层归一化、混合注意力模块、局部-非局部语义增强模块和多层感知机,具体网络结构如图1所示.

给定RGB帧 $\boldsymbol{I}_{\mathrm{r}}\in\mathbb{R}^{H\times W\times 3}$、光流图 $\boldsymbol{I}_{\mathrm{f}}\in\mathbb{R}^{H\times W\times 3}$ 和深度图 $\boldsymbol{I}_{\mathrm{d}}\in\mathbb{R}^{H\times W\times 3}$ 三种模态数据组成的三元组,其中下标r、f、d分别表示视频帧、光流和深度信息.然后,分别将其裁剪为 N 个固定尺寸的图像块,并将每个图像块展成一个 C 维向量 $\boldsymbol{x}_{l,j}$ 构成一个图符(token),其中 l 是r、f、d中的一种模态.随后,通过可学习的投影矩阵 $\boldsymbol{W}\in\mathbb{R}^{C\times D}$ 线性映射后加入位置编码 $\boldsymbol{P}\in\mathbb{R}^{N\times D}$,生成待编码的特征 $\boldsymbol{F}_l\in\mathbb{R}^{N\times D}$:

$$\boldsymbol{F}_l=[\boldsymbol{x}_{l,1}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{x}_{l,2}^{\mathrm{T}}\boldsymbol{W}\dots\boldsymbol{x}_{l,N}^{\mathrm{T}}\boldsymbol{W}]+\boldsymbol{P},l\in\{\mathrm{r},\mathrm{f},\mathrm{d}\}\tag{1}$$

接着,将模态 l 特征 $\boldsymbol{F}_l$ 输入编码器,并在各个阶段抽象产生它们的中间特征.编码器和解码器由多个本Transformer模块组成.其通过利用深度信号引导RGB帧与光流分支学习鲁棒的特征表示,有效解决了感受野受限和光流噪声导致的模型退化问题.随后,通过跳跃连接将逐级编码的特征输入解码器得到特征 $\boldsymbol{S}$.最后,特征 $\boldsymbol{S}$ 输入预测器并预测得到最后的分割掩膜 $\boldsymbol{M}\in\mathbb{R}^{H\times W}$:

$$\boldsymbol{M}=\mathrm{Conv}(\boldsymbol{S})\tag{2}$$

其中,Conv表示 1×1 卷积.

3.2 混合注意力模块

混合注意力模块是追求简洁、紧凑的端到端UVOS的核心设计.如图2所示,首先使用线性映射层将 $\boldsymbol{F}_l$ 映射到对应的查询(query),键(key)和值(value):

$$[\boldsymbol{Q}_l\boldsymbol{K}_l\boldsymbol{V}_l]=[\boldsymbol{F}_l\boldsymbol{W}_{\mathrm{q},l}\boldsymbol{F}_l\boldsymbol{W}_{\mathrm{k},l}\boldsymbol{F}_l\boldsymbol{W}_{\mathrm{v},l}],l\in\{\mathrm{r},\mathrm{f},\mathrm{d}\}\tag{3}$$

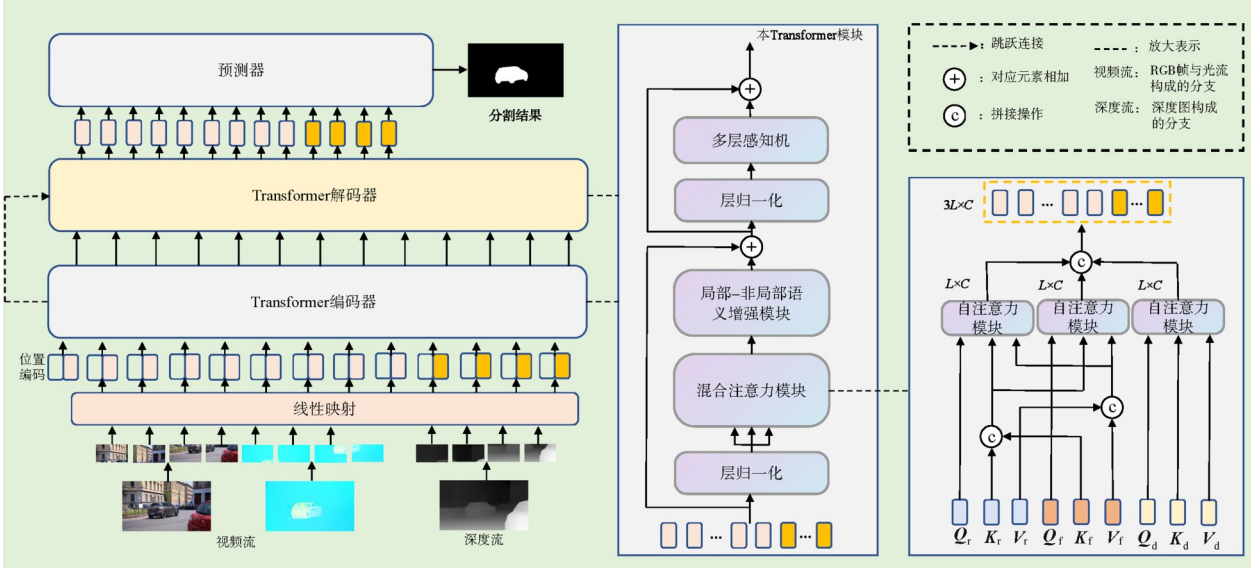


图1 基于深度信号引导学习混合变换器网络结构

其中，$\boldsymbol{Q}_l, \boldsymbol{K}_l, \boldsymbol{V}_l \in \mathbb{R}^{N\times D}$ 分别表示对应的查询、键和值. $\boldsymbol{W}_{q,l}, \boldsymbol{W}_{k,l}, \boldsymbol{W}_{v,l} \in \mathbb{R}^{D\times D}$. 表示可学习的投影矩阵. 然后，分别将 $\boldsymbol{K}_r$ 与 $\boldsymbol{K}_f$ 拼接，$\boldsymbol{V}_r$ 与 $\boldsymbol{V}_f$ 拼接，得到RGB帧和光流分支融合的key和value. 分别表示为 $\boldsymbol{K}_{video} \in \mathbb{R}^{2N\times D}$ 与 $\boldsymbol{V}_{video} \in \mathbb{R}^{2N\times D}$：

$$[\boldsymbol{K}_{video}\boldsymbol{V}_{video}]=[\mathrm{Cat}(\boldsymbol{K}_r,\boldsymbol{K}_f),\mathrm{Cat}(\boldsymbol{V}_r,\boldsymbol{V}_f)] \tag{4}$$

接着，分别对它们做如下注意力操作：

$$[\boldsymbol{A}_r\boldsymbol{A}_f\boldsymbol{A}_d]=\mathrm{Softmax}\left(\frac{[\boldsymbol{Q}_r\boldsymbol{K}_{video}^{\mathrm{T}}\boldsymbol{Q}_f\boldsymbol{K}_{video}^{\mathrm{T}}\boldsymbol{Q}_d\boldsymbol{K}_d^{\mathrm{T}}]}{\sqrt{d}}\right) \tag{5}$$

其中，$\boldsymbol{A}_r, \boldsymbol{A}_f \in \mathbb{R}^{N\times 2N}, \boldsymbol{A}_d \in \mathbb{R}^{N\times N}$ 分别表示RGB特征、光流特征和深度特征的注意力权值，d 表示key的维度. 得到后的注意力权值与对应的value相乘得到增强后的特征：

$$[\boldsymbol{Y}_r\boldsymbol{Y}_f\boldsymbol{Y}_d]=[\boldsymbol{A}_r\boldsymbol{V}_{video}\boldsymbol{A}_f\boldsymbol{V}_{video}\boldsymbol{A}_d\boldsymbol{V}_d] \tag{6}$$

其中，$\boldsymbol{Y}_r, \boldsymbol{Y}_f, \boldsymbol{Y}_d \in \mathbb{R}^{N\times D}$ 分别表示RGB、光流和深度图增强后的特征. 最后，将 $\boldsymbol{Y}_r, \boldsymbol{Y}_f, \boldsymbol{Y}_d$ 拼接得到输出 $\boldsymbol{Y}$：

$$\boldsymbol{Y}=\mathrm{Cat}(\boldsymbol{Y}_r,\boldsymbol{Y}_f,\boldsymbol{Y}_d) \tag{7}$$

其中，$\boldsymbol{Y} \in \mathbb{R}^{3N\times D}$ 表示充分挖掘不同模态之间共性信息后的融合特征.

图2展示了不同特征的热力图. 从中可见，第一行的光流效果较好，外观特征融合光流特征可以有效关注目标区域，融合深度特征可以提供额外的目标定位信息. 第二行到第四行展示了当光流信息包含噪声时，导致模型对噪声信号过拟合. 但是，当融合深度特征后，模型可以利用深度信号提供的空间位置信息，更加准确的关注运动目标. 第五行展示了当目标静止时，光流线索产生负面影响，导致分割效果变差. 但是，当引入深度特征时，模型易于区分出不同空间位置的物体，从而精确分割出主要目标. 上述实验结果表明了本文

图2 不同特征热力图

提出的混合注意力模块能够充分交互学习RGB帧、光流和深度图之间的特征，从而充分挖掘它们之间的共性信息，增强了模型的抗干扰能力. 传统变换器中采用的自注意力机制只能获取空间位置信息，而未能有效捕获通道方向的语义信息，导致其泛化能力受限，难以有效建模输入图符与图符之间潜在的依赖关系. 因此，本文设计了一个局部-非局部语义增强模块，将局部语义的归纳偏置引入补充学习非局部语义特征，在提升模型抗干扰力的同时获取更加精细的目标区域.

3.3 局部-非局部语义增强模块

图3展示了局部-非局部语义增强模块的设计细节. 其中，对于非局部语义增强模块，首先使用线性投影层将增强后的特征 $\boldsymbol{Y}$ 沿通道维投影到潜在空间 $\boldsymbol{Q}_c \in \mathbb{R}^{3N\times D}$，key $\boldsymbol{K}_c \in \mathbb{R}^{3N\times D}$ 和value $\boldsymbol{V}_c \in \mathbb{R}^{3N\times D}$：

$$[\boldsymbol{Q}_c\boldsymbol{K}_c\boldsymbol{V}_c]=[\boldsymbol{Y}\boldsymbol{W}_{qc}\boldsymbol{Y}\boldsymbol{W}_{kc}\boldsymbol{Y}\boldsymbol{W}_{vc}] \tag{8}$$

其中，$\boldsymbol{W}_{qc}, \boldsymbol{W}_{kc}, \boldsymbol{W}_{vc} \in \mathbb{R}^{D\times D}$ 表示query，key和value的投影矩阵. 然后，对它们执行自注意力操作：

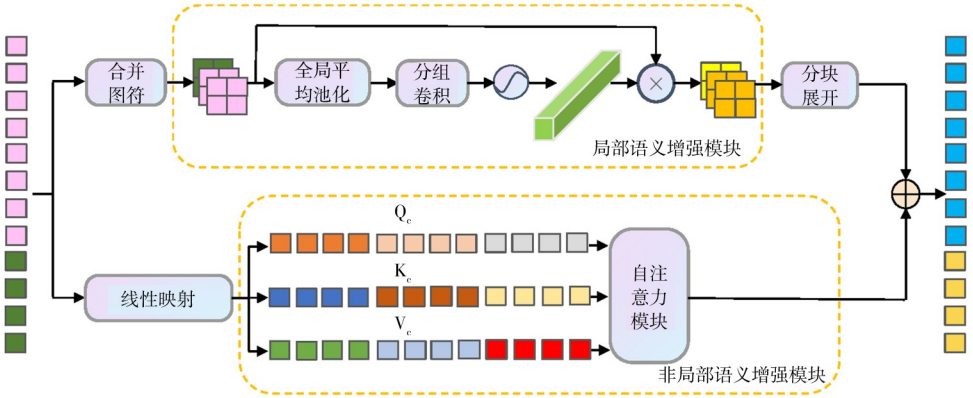


图3　局部-非局部语义增强模块

$$\boldsymbol{A}_{\mathrm{c}}=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}_{\mathrm{c}}{}^{\mathrm{T}}\boldsymbol{V}_{\mathrm{c}}}{\sqrt{d}}\right) \tag{9}$$

其中,$\boldsymbol{A}_{\mathrm{c}}\in\mathbb{R}^{D\times D}$ 表示全局语义注意力权重矩阵,d 表示 key 的维度. 最后,注意力权值与对应的 value 相乘得到增强后的非局部语义特征 $\boldsymbol{O}_{\mathrm{NL}}\in\mathbb{R}^{3N\times D}$:

$$\boldsymbol{O}_{\mathrm{NL}}=(\boldsymbol{A}_{\mathrm{c}}\boldsymbol{V}_{\mathrm{c}}{}^{\mathrm{T}})^{\mathrm{T}} \tag{10}$$

对于局部语义模块,首先将输入特征 $\boldsymbol{Y}$ 合并:

$$\boldsymbol{Y}_1=\mathrm{Merge}(\boldsymbol{Y}) \tag{11}$$

其中,$\boldsymbol{Y}_1\in\mathbb{R}^{T\times H_1\times W_1\times D}$ 表示合并后的特征,Merge(·)表示合并的操作. 合并后的特征经过全局平均池化层聚合空间特征:

$$\boldsymbol{Y}_2=\mathrm{GAP}(\boldsymbol{Y}_1) \tag{12}$$

其中,$\boldsymbol{Y}_2\in\mathbb{R}^{1\times1\times1\times D}$ 表示聚合空间特征,GAP(·)表示全局平均池化操作. 接着,将空间聚合特征经过分组卷积层学习特征,有效缓解了全卷积复杂计算量的问题. 增强后的特征经过 Sigmoid 激活函数层得到通道维度的局部注意力权值:

$$\boldsymbol{Y}_3=\mathrm{Sigmoid}(\mathrm{GroupConv}(\boldsymbol{Y}_2)) \tag{13}$$

其中,$\boldsymbol{Y}_3\in\mathbb{R}^{1\times1\times1\times D}$ 表示局部语义注意力权重向量,GroupConv 表示分组卷积操作,Sigmoid 表示激活函数层. 最后,将得到的局部语义注意力权值与特征 $\boldsymbol{Y}_1$ 对应元素相乘:

$$\boldsymbol{O}_{\mathrm{L}}=\boldsymbol{Y}_3\odot\boldsymbol{Y}_1 \tag{14}$$

其中,$\boldsymbol{O}_{\mathrm{L}}\in\mathbb{R}^{T\times H_1\times W_1\times D}$ 表示增强局部语义信息的特征,⊙表示对应元素相乘. 增强语义信息后的特征分块展开转换为原始特征维度:

$$\boldsymbol{O'}_{\mathrm{L}}=\mathrm{Split}(\boldsymbol{O}_{\mathrm{L}}) \tag{15}$$

其中,$\boldsymbol{O'}_{\mathrm{L}}\in\mathbb{R}^{3N\times D}$ 表示原始维度的增强语义信息后的特征,Split 表示分块展开操作. 增强后的非局部语义特征 $\boldsymbol{O}_{\mathrm{NL}}$ 和局部语义特征 $\boldsymbol{O'}_{\mathrm{L}}$ 对应元素相加得到最终的特征 $\boldsymbol{O}\in\mathbb{R}^{3N\times D}$:

$$\boldsymbol{O}=\boldsymbol{O}_{\mathrm{NL}}\oplus\boldsymbol{O'}_{\mathrm{L}} \tag{16}$$

其中,⊕表示对应元素相加.

3.4　损失函数

本文使用交叉熵损失函数 $\mathcal{L}_{\mathrm{CE}}$ 和 IoU 损失函数 $\mathcal{L}_{\mathrm{IoU}}$,旨在对目标进行像素级分类. 总的损失函数 $\mathcal{L}$ 定义如下:

$$\mathcal{L}=\mathcal{L}_{\mathrm{CE}}+\mathcal{L}_{\mathrm{IoU}} \tag{17}$$

其中,$\mathcal{L}_{\mathrm{CE}}=-\sum_{i,j}\boldsymbol{G}(i,j)\log\boldsymbol{M}(i,j)$,$\mathcal{L}_{\mathrm{IoU}}=1-\dfrac{|\boldsymbol{M}\cap\boldsymbol{G}|}{|\boldsymbol{M}\cup\boldsymbol{G}|}$,$\boldsymbol{G}\in\mathbb{R}^{H\times W}$ 表示真实标签,$\boldsymbol{M}\in\mathbb{R}^{H\times W}$ 表示预测掩模.

4　实验结果与分析

4.1　实验设置

本文以端到端的方式训练模型. 模型的输入是 RGB 帧、光流和深度图. 其中,深度图和光流分别使用 MiDaS(Mixing DataSets)[13] 和 RAFT(Recurrent All-pairs Field Transforms)[14] 算法生成. 训练的数据集由 DAVIS16[15]、FBMS[16] 和 YouTube-VOS[17] 组成. 本文先在 YouTube-VOS、DAVIS16 和 FBMS 数据集训练 15 个周期,再在 DAVIS16 和 FBMS 数据集上微调 25 个周期. 此外,本文采用了随机旋转、随机裁剪和随机水平翻转的数据增强策略,并将图片统一缩放到的尺寸. 整个网络使用 AdamW 优化器,并初始化学习率为 1e−4,批量大小设置为 4. 使用前 2 个周期预热,剩余周期余弦衰减的优化策略. 实验平台使用一块 64 GB 内存、12 核、2.50 GHz 的 AMD-Ryzen3950X CPU 和两块 11 GB 显存的 RTX2080Ti GPU. 推理阶段,本文在 DAVIS16 和 FBMS 数据集上使用区域相似度 J、边界准确率 F 和平均值 J&F 作为 UVOS 评估指标[15],在 DAVIS16、FBMS、DAVSOD[18] 和 ViSal[19] 数据集上使用平均绝对误差 MAE、$F_{\max}$ 和结构值 S_α 作为视频显著性评估指标[6]. 此外,生成的分割掩膜无需任何后处理. 在前向推理过程中,使用 RAFT 算法估计光流约耗时 0.025 s,使用 MiDaS 算法估计深度图约耗时 0.023 s,所提算法约耗时

0.02 s，总耗时约为0.068 s，即推理速度达到了15 fps.

4.2 定量实验分析

表1列举了本文方法与目前最先进的UVOS方法在DAVIS16和FBMS数据集上的定量比较结果. 其中，红色加粗为最优结果，蓝色加粗为次优结果，绿色加粗为排名第三的结果. 本文所提UVOS网络在DAVIS16数据集上的平均值、区域相似度和边界准确率上分别超过了目前最先进的UVOS算法D2Conv3D 0.4%、0.6%和0.2%. 此外，本文方法在FBMS数据集上的区域相似度大幅领先目前最先进的算. 本文方法在DAVIS16数据集和FBMS数据集上同时取得了最先进的性能，这充分表明了本文方法能够生成高质量的分割掩膜.

表1 在DAVIS16和FBMS数据集上的区域相似度和边界准确率评估结果

算法	DAVIS16			FBMS
	J&*F*	*J*	*F*	*J*
MATNet[5]	81.6	82.4	80.7	**76.1**
3DC-Seg[11]	84.5	84.3	84.7	—
FSNet[6]	83.3	83.4	83.1	—
TransportNet[7]	84.8	84.5	**85.0**	**78.7**
RTNet[8]	**85.2**	**85.6**	84.7	—
CFANet[20]	82.8	83.5	82.0	—
D2Conv3D[12]	**86.0**	**85.5**	**86.5**	—
本文方法	**86.4**	**86.1**	**86.7**	**84.0**

表2列举了本文方法与目前最先进的视频显著性目标检测算法在DAVIS16、FBMS、DAVSOD和ViSal数据集上的定量比较结果. 其中，红色加粗为最优指标，蓝色加粗为次优指标，绿色加粗为第三优指标. 表中可见，本文在四个数据集的11个指标上都取得了最优结果，在一个指标上取得了次优的结果. 与目前最先进的视频显著性检测方法CFANet(Contrastive Features and Attention Network)相比，本文方法在各个评价指标上都领先于CFANet. 特别地，在UVOS方法中，本文方法在各个数据集上优于TransportNet. 这表明了本文方法在UVOS与VSOD(Video Salient Object Detection)任务上均具备优越的性能.

表2 在DAVIS16、FBMS、DAVSOD和ViSal数据集上的S_α，F_{max}和MAE的对比结果

算法	DAVIS16			FBMS			DAVSOD			ViSal		
	$S_\alpha\uparrow$	$F_{max}\uparrow$	MAE↓	$S_\alpha\uparrow$	$F_{max}\uparrow$	MAE↓	$S_\alpha\uparrow$	$F_{max}\uparrow$	MAE↓	$S_\alpha\uparrow$	$F_{max}\uparrow$	MAE↓
SSAV[18]	0.893	0.861	0.028	0.879	0.865	**0.040**	0.724	0.603	0.092	**0.943**	**0.939**	0.020
DFNet[21]	—	0.899	0.018	—	0.833	0.054	—	—	—	—	0.927	**0.017**
3DC-Seg[11]	—	**0.918**	**0.015**	—	0.845	0.048	—	—	—	—	0.922	0.019
CASNet[22]	0.873	0.860	0.032	0.856	0.863	0.056	0.694	—	0.089	**0.820**	0.847	0.029
TransportNet[7]	—	**0.928**	**0.013**	—	0.885	0.045	—	—	—	—	**0.953**	**0.012**
FSNet[6]	**0.920**	0.907	0.020	**0.890**	**0.888**	0.041	**0.773**	**0.685**	**0.072**	—	—	—
CFANet[20]	**0.918**	0.909	**0.015**	**0.909**	**0.915**	**0.026**	**0.753**	**0.662**	**0.083**	—	—	—
本文方法	**0.937**	**0.929**	**0.010**	**0.916**	**0.922**	**0.026**	**0.790**	**0.698**	**0.069**	**0.953**	**0.954**	**0.013**

4.3 消融实验分析

表3列举了本文算法在DAVIS16和FBMS数据集上的消融实验结果. 将深度图引入基线模型中，模型的*J*和*F*指标在DAVIS16数据集上相比于基线模型分别提升了1.9%和1.6%；在FBMS数据集上，模型的*J*指标比基线模型提高了2.6%. 这是因为当目标静止或存在复杂运动干扰时，估计出的光流将含有大量噪声，导致模型对噪声信号过拟合. 深度信号可提供一帧中所有对象的空间位置信息，辅助减轻物体静止或复杂运动的干扰，从而能够得到更高质量的目标分割掩膜. 在原有的基线模型中加入混合注意力模块，在DAVIS16数据集上，模型的*J*和*F*比基线模型提升了2.5%和3.2%；

表3 在DAVIS16和FBMS数据集上消融实验结果

模型变化			DAVIS16		FBMS
深度图	混合注意力模块	局部-非局部语义增强模块	*J*	*F*	*J*
			80.7	80.6	78.2
√			82.6	82.2	80.8
	√		83.2	83.8	81.4
		√	81.3	81.5	79.8
√	√		84.6	84.8	82.6
√		√	83.6	84.0	81.2
	√	√	84.2	84.6	82.0
√	√	√	86.1	86.7	84.0

在FBMS数据集上,模型的指标J在基线模型的基础上提升了3.2%. 这是因为混合注意力模块能够充分交互学习RGB帧、光流和深度图之间的特征,从而充分挖掘它们之间的共性信息,生成更加精确的目标分割掩膜. 在基准模型上加入全局-局部语义调制模块,在DAVIS16和FBMS数据集上的评估指标都获得了一定的提升,这充分表明了全局-局部语义调制模块能够获取更加精细、完整的目标区域,从而进一步提升了分割的精度.

为了有效验证深度图、混合注意力模块和局部-非局部语义增强模块对模型的积极作用,本文通过删除模块的方法来测试对模型性能的影响. 当去除混合注意力模块时,模型在DAVIS16数据集上的J和F比原始版本降低了2.5%和2.7%;在FBMS数据集上,模型的区域相似度J比原始版本降低了2.0%,充分体现了混合注意力模块在挖掘各种模态数据之间共性的优越性. 当去除深度图时,模型在DAVIS16数据集上的J和F比原始版本降低了1.9%和2.1%;在FBMS数据集上,J比原始版本降低了2.0%. 体现了深度图能够有效提供目标的空间位置信息,避免了受物体静止和复杂运动的干扰. 在去除局部-非局部语义增强模块的情况下,模型在DAVIS16和FBMS数据上的性能均下降了一些,侧面体现了局部-非局部语义增强模块能够获取更加精细的目标区域,进一步提高了分割质量.

4.4　定性实验分析

图4展示了一些本文算法的定性分析结果. 从上到下依次是DAVIS16数据集(breakdance,car-roundabout,dance-twirl和libby),DAVSOD数据集(select_0689和select_0669)和FBMS数据集(dogs01). 本文提出的方法在实际场景中表现出色,包括快速移动场景(car-roundabout,libby,select_0689和select_0669)、杂乱背景场景(breakdance,car-roundabout和dance-twirl)、严重遮挡场景(libby,select_0689,select_0669和dogs01)和剧烈形变场景(breakdance,dance-twirl和dogs01). 例如,在第一行和第三行中,即使周围有很多观看者的背景,本文提出的分割算法依然能够准确的定位并分割出舞蹈者. 在第六行中,冲浪的人在浪花中只能看见头部,本文提出的算法依然能够精准的分割出头部轮廓. 在第四行中,小狗被各种围栏、大树遮挡,本文算法仍然可以只分割小狗而不分割遮挡的背景部分. 以上定性分析结果验证了本文算法在各种复杂场景中的有效性.

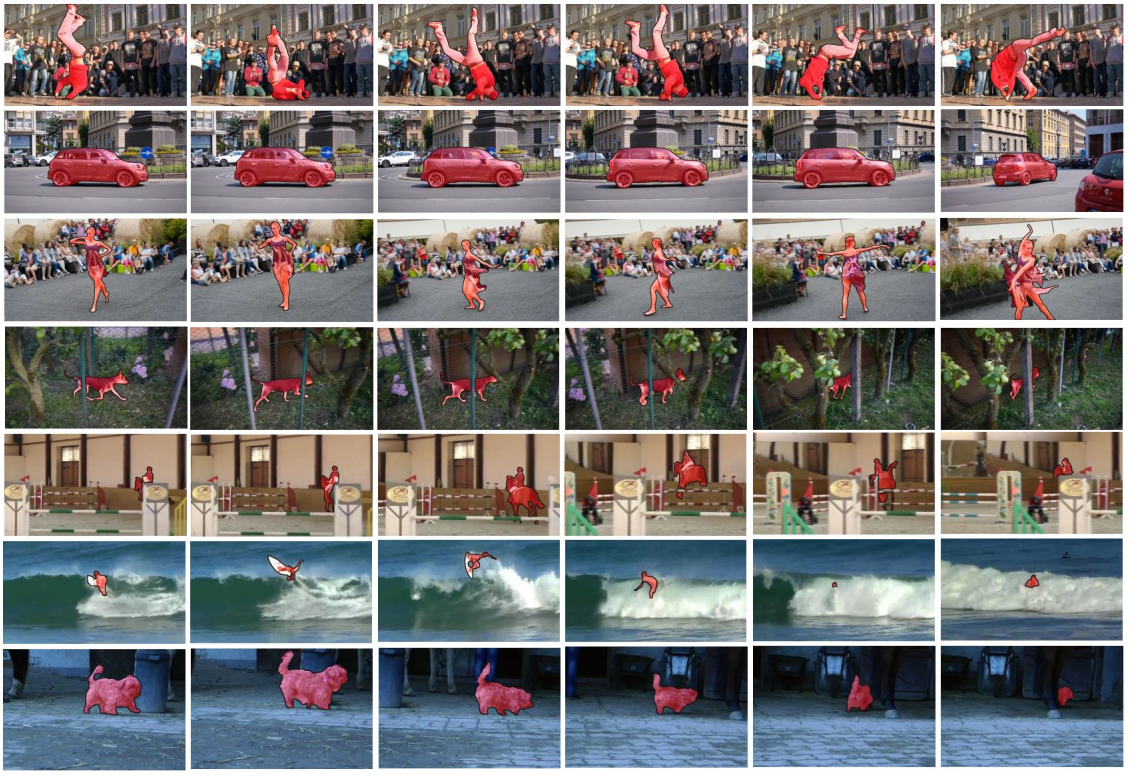


图4　定性实验结果

4.5　模型限制

当深度估计不准确时会存在两种情况,如图5所示. 第一种情况是深度估计不准确且光流存在噪声(见图5前2行),这种条件下,深度信号会对光流噪声起到一定的抑制效果,但是预测结果仍然会存在一些噪声;第二种情况是深度估计不准确但是光流估计较为准确

(见图5后2行),这种条件下,光流估计会占主导作用,因此最终的预测结果效果较为理想. 总之,只有当深度估计不准确且光流存在噪声的情况下,模型预测效果会较差,但深度图仍会对噪声起到一定的抑制效果.

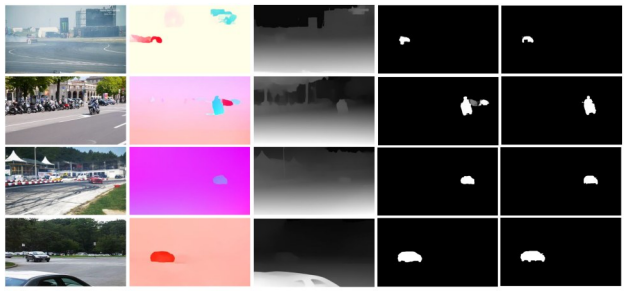


图5 深度估计不准确条件下光流对结果的影响

5 总结

为了解决现有的UVOS框架存在对光流噪声过拟合、感受野受限、网络复杂、计算量大等问题,本文提出了一种基于深度信号引导学习混合变换器的UVOS网络. 实验表明本文所提方法能够有效抑制光流噪声带来的干扰并且推理速度快,能够为视频编辑、目标跟踪和自动驾驶等领域提供更加快速准确的图像处理结果.

参考文献

[1] 谢青松, 刘晓庆, 安志勇, 等. 基于前景优化的视觉目标跟踪算法[J]. 电子学报, 2022, 50(7): 1558-1566.
XIE Q S, LIU X Q, AN Z Y, et al. Visual object tracking algorithm based on foreground optimization[J]. Acta Electronica Sinica, 2022, 50(7): 1558-1566. (in Chinese).

[2] 付利华, 赵宇, 姜涵煦, 等. 基于前景感知视觉注意的半监督视频目标分割[J]. 电子学报, 2022, 50(1): 195-206.
FU L H, ZHAO Y, JIANG H X, et al. Semi-supervised video object segmentation based on foreground perception visual attention[J]. Acta Electronica Sinica, 2022, 50(1): 195-206. (in Chinese).

[3] 付利华, 赵宇, 孙晓威, 等. 基于孪生网络的快速视频目标分割[J]. 电子学报, 2020, 48(4): 625-630.
FU L H, ZHAO Y, SUN X W, et al. Fast video object segmentation based on Siamese networks[J]. Acta Electronica Sinica, 2020, 48(4): 625-630. (in Chinese).

[4] FAN J, ZHANG K, ZHAO Y, et al. Unsupervised video object segmentation via weak user interaction and temporal modulation[J]. Chinese Journal of Electronics, 2022, 32: 1-13.

[5] ZHOU T F, LI J W, WANG S Z, et al. Matnet: Motion-attentive transition network for zero-shot video object segmentation[J]. IEEE Transactions on Image Processing, 2020, 29: 8326-8338.

[6] JI G P, FU K R, WU Z, et al. Full-duplex strategy for video object segmentation[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2022: 4922-4933.

[7] ZHANG K H, ZHAO Z C, LIU D, et al. Deep transport network for unsupervised video object segmentation[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 8781-8790.

[8] REN S, LIU W, LIU Y, et al. Reciprocal transformations for unsupervised video object segmentation[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 15455-15464.

[9] TOKMAKOV P, ALAHARI K, SCHMID C. Learning video object segmentation with visual memory[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 4481-4490.

[10] LU X K, WANG W G, DANELLJAN M, et al. Video object segmentation with episodic graph memory Networks [C]//Computer Vision - ECCV 2020. Cham: Springer International Publishing, 2020: 661-679.

[11] Mahadevan S, Athar A, Ošep A, et al. Making a case for 3d convolutions for object segmentation in videos[EB/OL]. (2020-08-26) [2022-11-01]. arXiv preprint arXiv: 2008.11516, 2020.

[12] SCHMIDT C, ATHAR A, MAHADEVAN S, et al. D2conv3d: Dynamic dilated convolutions for object segmentation in videos[C]//2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE, 2022: 1200-1209.

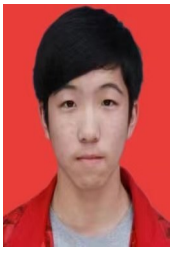
[13] RANFTL R, LASINGER K, HAFNER D, et al. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 44(3): 1623-1637.

[14] TEED Z, DENG J. RAFT: Recurrent all-pairs field transforms for optical flow[C]//Computer Vision - ECCV 2020. Cham: Springer International Publishing, 2020: 402-419.

[15] 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). Pisca-

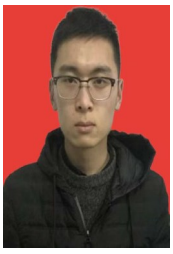
taway: IEEE, 2016: 724-732.
[16] OCHS P, MALIK J, BROX T. Segmentation of moving objects by long term video analysis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 36 (6): 1187-1200.
[17] XU N, YANG L J, FAN Y C, et al. Youtube-vos: Sequence-to-sequence video object segmentation [C]//Computer Vision - ECCV 2018. Cham: Springer International Publishing, 2018: 585-601.
[18] FAN D P, WANG W G, CHENG M M, et al. Shifting more attention to video salient object detection[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 8554-8564.
[19] WANG W G, SHEN J B, SHAO L. Consistent video saliency using local gradient flow optimization and global refinement[J]. IEEE Transactions on Image Processing, 2015, 24(11): 4185-4196.
[20] CHEN Y W, JIN X J, SHEN X H, et al. Video salient object detection via contrastive features and attention modules[C]//2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE, 2022: 1320-1329.
[21] ZHEN M M, LI S W, ZHOU L, et al. Learning discriminative feature with crf for unsupervised video object segmentation[C]//Computer Vision - ECCV 2020. Cham: Springer International Publishing, 2020: 445-462.
[22] JI Y Z, ZHANG H J, JIE Z Q, et al. CASNet: A cross-attention Siamese network for video salient object detection [J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(6): 2676-2690.

作者简介



苏天康　男,1999年生,安徽芜湖人,硕士,研究生,主要研究方向为无监督视频目标分割.
E-mail: tiankangsu@gmail.com

宋慧慧(通讯作者)　女,1986年生,山东省聊城市人,博士,教授,主要研究领域为视频目标分割和图像超分.
E-mail: songhuihui@nuist.edu.cn



樊佳庆　男,1994年生,江苏南通人,博士在读,主要研究领域为视频目标分割和视觉跟踪.
E-mail: jqfan@nuaa.edu.cn

张开华　男,1983年生,山东省日照市人,博士,教授,主要研究领域为协同显著性检测和视觉跟踪.
E-mail: zhkhua@gmail.com