计算机科学
COMPUTER SCIENCE

基于跟踪检测时序特征融合的视频遮挡目标分割方法

郑申海, 高茜, 刘鹏威, 李伟生

引用本文

郑申海, 高茜, 刘鹏威, 李伟生. 基于跟踪检测时序特征融合的视频遮挡目标分割方法[J]. 计算机科学, 2024, 51(6A): 230600186-6.

ZHENG Shenhai, GAO Xi, LIU Pengwei, LI Weisheng. Occluded Video Instance Segmentation Method Based on Feature Fusion of Tracking and Detection in Time Sequence [J]. Computer Science, 2024, 51(6A): 230600186-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230600186

基于跟踪检测时序特征融合的视频遮挡目标分割方法

郑申海[1,2]　高　茜[1]　刘鹏威[1]　李伟生[1,2]
1 重庆邮电大学计算机科学与技术学院　重庆 400065
2 图像认知重庆市重点实验室(重庆邮电大学)　重庆 400065

摘　要　视频实例分割是近年来兴起的一项在图像实例分割基础上引入时序特性的视觉任务,旨在同时对每一帧的目标进行分割并实现帧间的目标跟踪。移动互联网和人工智能的迅猛发展产生了大量的视频数据,但由于拍摄角度、快速运动和部分遮挡等,视频中的物体往往会出现分裂或模糊的情况,使得从视频数据中准确地分割目标并对目标进行处理和分析面临着重大挑战。经查阅和实践发现,现有的视频实例分割方法在遮挡情况下的表现较差。针对上述问题,提出了一种改进的遮挡视频实例分割算法——通过融合 Transformer 和跟踪检测的时序特征来改善分割性能。为增强网络对空间位置信息的学习能力,该算法将时间维度引入 Transformer 网络中,并考虑到视频中目标检测、跟踪和分割之间的相互依赖和促进关系,提出了一种能够有效地聚合目标在视频中的跟踪偏移的融合跟踪模块和检测时序特征模块,提升了遮挡环境下的目标分割性能。通过在 OVIS 和 YouTube-VIS 数据集上进行的实验,验证了所提方法的有效性。相比当前的基准方法,该方法展现出了更好的分割精度,进一步证明了其优越性。

关键词:视频实例分割;目标检测;目标跟踪;时序特征;遮挡目标
中图分类号　TP391

Occluded Video Instance Segmentation Method Based on Feature Fusion of Tracking and Detection in Time Sequence

ZHENG Shenhai[1,2],GAO Xi[1],LIU Pengwei[1] and LI Weisheng[1,2]
1 College of Computer Science and Technology,Chongqing University of Posts and Telecommunications,Chongqing 400065,China
2 Chongqing Key Laboratory of Image Cognition(Chongqing University of Posts and Telecommunications),Chongqing 400065,China

Abstract　Video instance segmentation is a visual task that has emerged in recent years,which introduces temporal characteristics on the basis of image instance segmentation. It aims to simultaneously segment objects in each frame and achieve inter frame object tracking. A large amount of video data has been generated with the rapid development of mobile Internet and artificial intelligence. However,due to shooting angles,rapid motion,and partial occlusion,objects in videos often split or blur,posing significant challenges in accurately segmenting targets from video data and processing and analyzing them. After consulting and practicing,it is found that existing video instance segmentation methods perform poorly in occluded situations. In response to the above issues,this paper proposes an improved occlusion video instance segmentation algorithm,which improves segmentation performance by integrating the temporal features of Transformer and tracking detection. To enhance the learning ability of the network for spatial position information,this algorithm introduces the time dimension into the Transformer network and considers the interdependence and promotion relationship between object detection,tracking,and segmentation in videos. A fusion tracking module and a detection temporal feature module that can effectively aggregate the tracking offset of objects in videos are proposed,improving the performance of object segmentation in occluded environments. The effectiveness of the proposed method is verified through experiments on the OVIS and YouTube VIS datasets. Compared to the current benchmark method,the proposed method exhibits better segmentation accuracy,further demonstrating its superiority.

Keywords　Video instance segmentation,Object detection,Object tracking,Feature in time sequence,Occluded instance

1　引言

近年来,随着计算机图形计算能力的爆炸性增长和人工智能算法的迅猛发展,视频作为新的传播媒介,其使用越来越广泛。计算机视觉的主要研究方向之一就是图像分割,图像分割的特点是根据不同的需求将图像分成不同的部分,并添

基金项目:国家自然科学基金(61902046);重庆市教委科学技术研究计划重点项目(KJZD-K202200606);重庆市自然科学基金(2022NSCQ-MSX3746)
This work was supported by the National Natural Science Foundation of China(61902046),Science and Technology Research Program of Chongqing Municipal Education Commission(KJZD-K202200606) and Natural Science Foundation of Chongqing,China(2022NSCQ-MSX3746).
通信作者:郑申海(zhengsh@cqupt.edu.cn)

加相应的标签。它被认为是图像分类和识别向像素级的拓展，而实例分割会同时进行目标检测和语义分割以完成任务，有着丰富的应用场景，如自动驾驶识别、监控识别、远程会议等。视频实例分割在图像实例分割的基础上引入时序维度，在分割每一帧物体的同时在帧间跟踪这些物体，因此如何利用好视频的时序特征也是该任务的一大难点。因为视频帧存在多种图像质量问题，如运动模糊、对焦不准确、部分或严重遮挡以及外观变化等，所以涉及的视觉任务较多。因此，要建立一个分割精度准确、识别正确率高的视频实例分割模型仍然需要不断探索。

现有的视频实例分割算法在一些数据集中的分割效果表现出色。然而，视频中的物体往往不是孤立出现的，而是面临各种程度的遮挡。研究表明[1]，尽管存在遮挡，人类视觉系统仍能够通过生活经验或先前的观测迅速、准确地区分目标对象的正确边界。然而，对于视频分割系统而言，对有着严重遮挡的视频进行实例分割一直不能取得良好的效果。

2 相关工作

视频实例分割任务最初由 Yang 等[2]提出，他们同时发布 Youtube-VIS 作为该任务的第一个数据集。与数据集同时提出的 MaskTrack R-CNN 算法在 Mask R-CNN[3]的基础上增加了一个跟踪分支。MaskProp[4]则以分割算法 HTC 为基础，设计实例特征传播(Instance Feature Propagation)模块，其中使用帧间特征差值与可变形卷积来提高算法在处理运动模糊和短暂遮挡时的鲁棒性。STEm-Seg[5]利用像素级嵌入表征对一小段视频序列进行聚类，实现序列中同一物体的所有前景像素的聚类。CompFeat[6]利用空间和时间注意力模块从多个视角综合提取多尺度特征。VisTR[7]改进基于 Transformer[8]的端到端检测器 DETR[9]，通过图像序列输入、实例序列输出的方式完成视频实例分割任务。这类方法有的会忽略对时序上下文信息的利用，但大多都集中在 YouTube-VIS 数据集上进行验证和测试，在复杂遮挡场景下容易出现分割效果不佳的情况。

在提出之初，Transformer 常被用于自然语言处理方向[10]。2020 年，Transformer 开始在计算机视觉领域大放异彩。一些工作仅使用 Transformer 的结构完成图像分类任务，ViT[11](Vision Transformer)是 Google 在 2020 年提出的直接将 Transformer 应用于图像分类的模型。另一些工作则将 Transformer 与 CNN 相结合，如 STTR[12]和 LSTR[13]分别用于视差估计和车道形状预测，LRNet[14]和 SAN[15]在减轻自注意力计算方面做出贡献，而 Axial-DeepLab[16]则采用分解全局注意力的方式来减少计算量。

针对上述问题，本文设计一种引入时序特征的联合检测和跟踪的端到端视频实例分割框架(如图 1 所示)，有效地将其他帧检测和跟踪到的目标特征和线索聚合到当前图像特征上。所提方法的贡献有以下 3 个方面：

1)针对遮挡视频实例分割这一任务，提出将跟踪、检测关联的视频实例分割框架，将目标跟踪和检测的特征输入，实现对目标物体在遮挡视频中的实例分割效果的提升；

2)使用更加适合遮挡视频场景的 Transformer 优化主干网络特征提取部分，以实现对复杂真实遮挡场景下的实例分割；

3)在 OVIS 数据集和 YouTube-VIS 数据集中与现有方法进行对比，分割效果取得一定的提升。

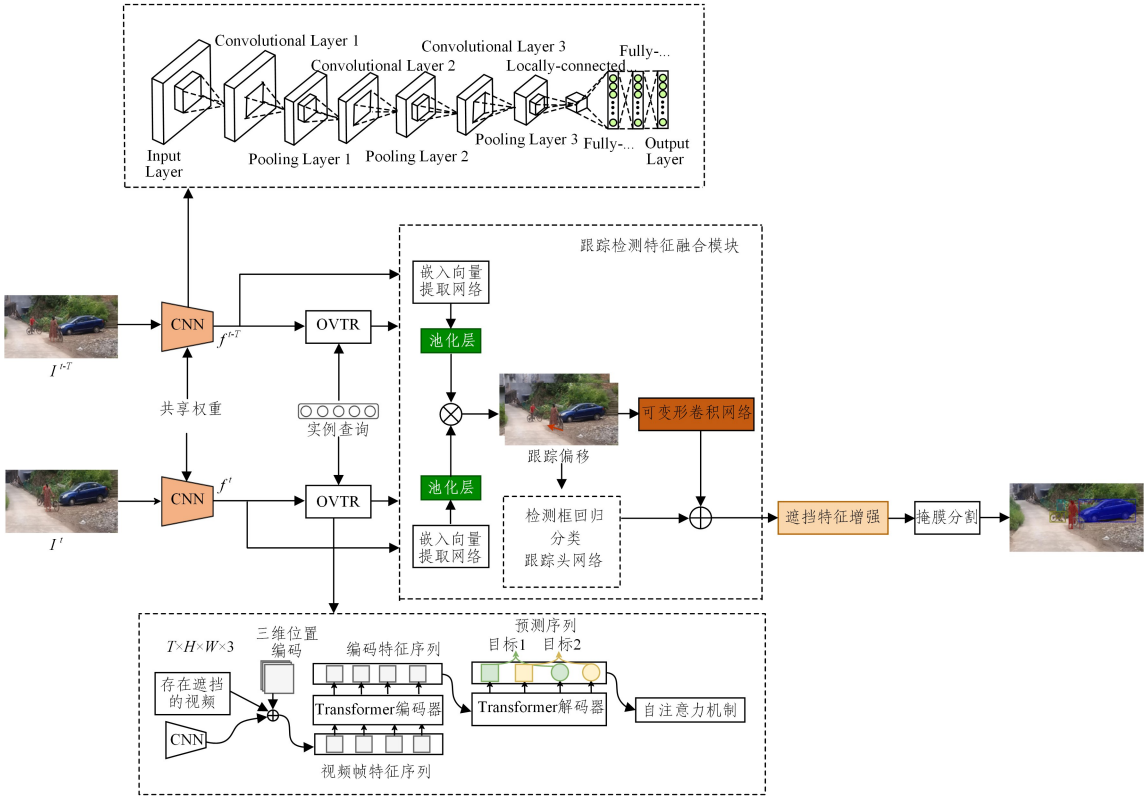

图 1 基于跟踪检测时序特征融合的视频遮挡目标分割方法的网络结构图

Fig. 1 Network structure diagram of video occlusion target segmentation method based on tracking detection time sequence feature fusion

3 特征提取主干网络改进方法

3.1 基本定义

由于卷积操作仅具备相对固定的局部相关性，需要寻找更灵活的视频帧处理方案才能得到更好的分割效果。Transformer是一种基于自注意力机制的深度学习模型。在视频实例分割领域，Transformer架构也被广泛应用，在视频实例分割中，Transformer有序列建模能力，适合对时间序列的建模，从而捕捉视频中的时间依赖性。此外，由于Transformer具有自注意力机制，因此它可以将每一帧与其他所有帧进行交互，从而更好地理解每一帧在整个时间序列中的上下文信息。相对于单张图像而言，视频包含时间维度，比图片有着更多的信息量，例如不同目标的轨迹和运动速度。这些丰富的线索可以帮助解决之前图像实例分割任务中的一些困难，比如两个外观相似的物体由于靠近产生遮挡等问题。

此外，多帧图像能够提供更好的参考帧特征表示，有助于模型更好地跟踪识别物体。出于这些考虑，本文选择Transformer模型，该模型非常擅长对长序列进行建模，可以将多帧图片直接分成固定大小的区域转为长序列模型，因此非常适合应用于严重遮挡视频中对多帧序列的时序信息进行建模。由于Transformer具有自注意力机制，它可以基于两两之间的相似度来学习和更新特征。这些特性使得Transformer成为视频实例分割(VIS)任务的恰当选择。因此，本文方法将时间维度引入Transformer模型中，设计了遮挡视频实例分割的模型的特征提取网络OVTR(Occluded Videos Transformer)，其网络结构如图2所示。

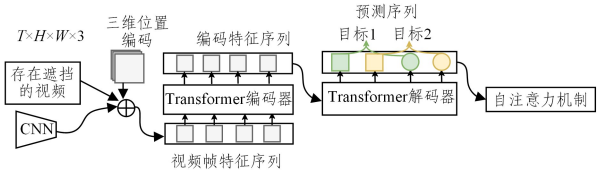

图2 OVTR示意图

Fig. 2 OVTR illustration

3.2 网络结构

最初，对于输入的视频片段，主干网络会提取像素级特征。设初始T帧分辨率为$H_0\times W_0$的视频片段记为$x\in R^{T\times 3\times H_0\times W_0}$。针对每一帧，主干网络会将每一帧的特征连接成特征图$f_0\in R^{T\times C\times H\times W}$。为了建模所有像素级特征之间的相似性，使用Transformer编码器。通过对特征图进行1×1卷积，得到新的特征图$f_1\in R^{T\times C\times H\times W}$，其维度从$C$降至$d(d<C)$。然后，将$f_1$的空间和时间维度平化为一维，从而得到一个大小为$d\times(T\cdot H\cdot W)$的2D特征图，以便将其输入Transformer编码器中进行处理。Transformer架构是置换不变的，而分割任务需要精确的位置信息，为了弥补这一缺陷，本文使用了固定的位置编码信息来补充特征，该编码信息包含视频片段的三维位置信息。本文将原Transformer中的位置编码进行调整以适应视频的三维情况。位置编码是一种在输入特征中注入位置信息的方法，它利用正弦和余弦函数在不同的位置编码中进行交替计算，以确保不同位置的编码之间的区别是一致的。这样，在特征被Transformer编码后，它们就具有了空间和时间的位置信息，从而可以更准确地执行视频分割任务。Transformer解码器的主要任务是解码每一帧中实例级别的特征，这些特征反映每个实例的顶部像素信息。本文方法受到DETR的启发，引入固定数量的输入嵌入，用于从像素特征中查询实例特征。实例查询的数量由模型自动学习，且与像素特征的维度相同。假设模型每帧需要解码n个实例，那么对于T帧的序列，实例查询的数量为$N=n\cdot T$。Transformer解码器接收Transformer编码器的输出和实例查询作为输入，然后输出实例特征。整体预测遵循输入帧的顺序，不同图像中实例的预测顺序相同，通过将对应的索引项直接链接起来，就可以实现对不同帧实例的跟踪。

4 跟踪检测时序特征融合模块

视频实例分割还需要实现目标的关联匹配，涉及目标的检测、分割和跟踪等多个任务的学习。针对检测、分割和跟踪等多个任务，目前的目标跟踪方法遵循两种主要的范式：基于检测的跟踪和联合检测跟踪。基于检测的跟踪范式将检测和跟踪作为两个独立的任务，首先利用现成的目标检测器检测，接着再用其他网络进行数据关联。这种基于检测的跟踪范式通常效率不高，并且两步骤流程不能实现端到端的优化。为了解决这个问题，最近的解决方案中出现了联合目标检测和跟踪范式，这种范式类似于将检测和跟踪在单个前向传播中完成。然而，联合跟踪检测范式仍然存在问题，虽然大多数联合跟踪检测网络的骨干网络是共享的，但是检测部分还是单独的，没有利用到跟踪信息。

本文提出跟踪检测特征融合模块(Feature Fusion Module of Tracking and Detection，FTD)，实现提取跟踪和检测的特征并将其结果输入分割头网络中，以提升最终的遮挡视频分割效果。

如图1所示，主干网络提取的特征图为f^t和f^{t-T}，将提取的特征传至embedding提取网络中，提取每张图像的embedding特征，embedding提取网络由3个卷积层构成，得到的特征图为$e^t=\sigma(f^t)\in R^{H_F\times W_F\times 128}$。然后通过计算相似矩阵来保存两帧之间特征图上每两个点之间的匹配相似度，用最大池化对embedding特征图进行下采样，得到$e'\in R^{H_c\times W_c\times 128}$，其中$H_c=H_F/2$，$W_c=W_F/2$，即每个特征图上有$H_c\times W_c$个目标的embedding向量。因此需要计算得到两个特征图上任意两个a点之间的相似矩阵$\boldsymbol{C}\in R^{H_c\times W_c\times H_c\times W_c}$，代表帧$I^t$和$I^{t-T}$之间的相似度。$O\in R^{H_c\times W_c\times 2}$中的每个元素的计算式为：

$$C_{i,j,k,l}=e'^{t}_{i,j}e'^{t-T}_{k,l} \tag{1}$$

其中，$C_{i,j,k,l}$代表帧t上的点(i,j)和帧$t-T$上的点(k,l)之间的embedding相似度。基于相似矩阵$\boldsymbol{C}$，可以计算跟踪偏移矩阵$\boldsymbol{O}\in R^{H_c\times W_c\times 2}$，这个矩阵存储$t$时刻每个点相对于其在$t-T$时刻的时空位移。

跟踪检测特征融合模块通过embedding网络提取每帧图像之间的embedding特征，经过矩阵计算，推理得到运动偏移，构成了追踪线索。设Y为监督训练的标签，当t帧上的(i,j)位置的目标出现在帧$t-T$上的(k,l)位置时，令$Y_{ijkl}=1$，否则$Y_{ijkl}=0$。Embedding提取网络的损失函数为：

$$L_E=\frac{-1}{\sum_{ijkl}Y_{ijkl}}\sum_{ijkl}\begin{cases}\alpha_1\log(C^{W}_{i,j,l})+\alpha_2\log(C^{H}_{i,j,k}), & \text{if } Y_{ijkl}=1\\ 0, & \text{else}\end{cases} \tag{2}$$

其中，$\alpha_1=(1-C^{w}_{i,j,l})^{\beta}$，$\alpha_2=(1-C^{H}_{i,j,k})^{\beta}$，$\beta$ 为超参数。本文提出的基于跟踪检测时序特征融合的视频遮挡目标分割方法的总损失函数为 $L=L_E+L_{\text{mask}}+L_{\text{box}}+L_{\text{cls}}$。由于跟踪偏移是基于外观相似度计算得到的，所以它能在较大的运动范围内跟踪目标，因此也能作为非常有效的运动线索，从而起到提升遮挡目标分割效果的作用。

5 实验

5.1 数据集

本文方法在 YouTube-VIS 2019[2] 和 OVIS[17] 数据集上进行了实验。YouTube-VIS 数据集包含 2 883 段视频，40 个物体类别，4 883 个物体，该数据集在视频实例分割任务中成为了一个重要的基准数据集。近期，OVIS 数据集被提出，该数据集是一种具有高度挑战性的视频实例分割数据集，体现出现实场景的复杂性和多样性。OVIS 数据集包括 296 000 个高质量实例掩码（大约是 YouTube-VIS-2019 数据集的 2 倍），每个视频 5.80 个实例（大约是 YouTube-VIS-2019 数据集的 3.4 倍），来自 25 个语义类别，其中经常出现对象遮挡的情况。

5.2 实验设置

使用 ResNet-50[18] 作为本文方法的主干网络，并使用与 DETR 相同的超参数。对于 Transformer，该方法使用 6 个编码器，6 个解码器层，宽度为 256。本模型基于 PyTorch-1.6 实现，使用 AdamW[19] 优化器以初始 Transformer 参数进行训练，初始学习率为 10^{-4}，骨干网络学习率为 10^{-5}。模型进行 18 个 epochs 的训练，其中学习率在前 12 个 epochs 中衰减为原来的 1/10。本文方法使用在 COCO[20] 数据集上预训练的 DETR 模型的参数来初始化骨干网络。

本文选取数个主流的视频实例分割网络作为基线，分别是 STEm-Seg[7]，CrossVIS[21]，FEELVOS[22]，IoUTracker+[23]，MaskTrack R-CNN[2]，TraDeS[24] 和 SipMask[25]。为了公平比较，所有方法都在相同的测试环境下进行实验，使用 NVIDIA 2080Ti 训练测试数据，图像都统一处理缩放到 640×360 像素。使用平均精度（AP）和召回率（AR）来评估视频实例分割任务。在 OVIS 验证集上，对比方法的结果如表 1 所列。

表 1 在 OVIS 验证集上与现有方法的分割精度比较

Table 1 Segmentation accuracy comparison with existing methods on OVIS validation set

（%）

方法	评价指标				
	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
STEm-Seg[7]	13.8	32.1	11.9	9.1	20.0
CrossVIS[21]	14.9	32.7	12.1	10.3	19.8
FEELVOS[22]	9.6	22.0	7.3	7.4	14.8
IoUTracker+[23]	7.0	16.9	5.3	5.7	14.3
MaskTrack R-CNN	10.8	25.3	8.5	7.9	14.9
TraDeS[24]	11.4	26.5	9.4	7.0	13.8
SipMask[25]	10.2	24.7	7.8	7.9	15.8
本文方法	**16.9**	**33.5**	**12.3**	**10.6**	**21.5**

根据表 1 的结果可以得出，本文方法在 OVIS 验证集上的视频实例分割效果优于现有的基准方法。与 CrossVIS 相比，本文方法的平均准确度 AP 提高了 2.0%，AP_{50} 提高了 0.8%，AP_{75} 提高了 0.2%。相比之下，现有的方法如 MaskTrack R-CNN，忽略了对其他帧的时序线索的使用，仅提取当前图像的特征，难以实现帧间的跟踪检测特征在分割中的使用。

在 YouTube-VIS 验证集上，对比方法的定量结果如表 2 所列。根据表 2 的结果，本文方法在 YouTube-VIS 数据集上取得了最好的表现，具有较高的分割精度。

表 2 在 YouTube-VIS 验证集上与现有方法的分割精度比较

Table 2 Segmentation accuracy comparison with existing methods on YouTube-VIS validation set

方法	评价指标				
	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
STEm-Seg	30.6	50.7	33.5	37.6	37.1
CrossVIS	36.3	56.8	38.9	35.6	40.7
FEELVOS	26.9	42.0	29.7	29.9	33.4
IoUTracker+	23.6	39.2	25.5	26.2	30.9
MaskTrack R-CNN	30.3	51.1	32.6	31.0	35.5
TraDeS	32.6	52.6	32.8	32.4	37.6
SipMask	32.5	53.0	33.3	33.5	38.9
本文方法	**37.2**	**57.3**	**40.5**	**40.1**	**41.8**

跟踪偏移通过外观嵌入向量相似性预测可以匹配具有剧烈运动或低帧速率的对象，并且可以准确地跟踪具有被遮挡运动的对象，同时对物体的预测跟踪偏移可以指导特征融合模块中的特征传播，从而恢复当前帧中被遮挡和模糊的对象特征。特征融合模块通过特征传播可以提高目标跟踪的精度，减少跟踪器的漂移，并且能够在先前帧中提取的传播特征的支持下，恢复可能丢失跟踪的目标。因此，本文提出的模型方法可以提高目标跟踪的鲁棒性和精度。本文方法在一些评估实例上的表现如图 3 所示，可以看出，本方法成功地跟踪到所有目标，即使是在目标被完全遮挡的极端情况下。第一行的实例中，黄色掩码的汽车在第三帧中完全遮挡红色掩码的汽车，然后在第四帧中被橙色掩码的汽车遮挡，在这种极端情况下，模型成功地跟踪到所有的车辆。在第二行中，模型成功地跟踪到黄色掩码中的熊，尽管其被其他物体（紫色掩码中的熊和背景树部分）遮挡。第四行中展示了一个拥挤目标种类数量繁多的场景，所有的物体都被正确地分割。这些评估实例展示了本文方法在处理遮挡和拥挤场景中的优秀表现。

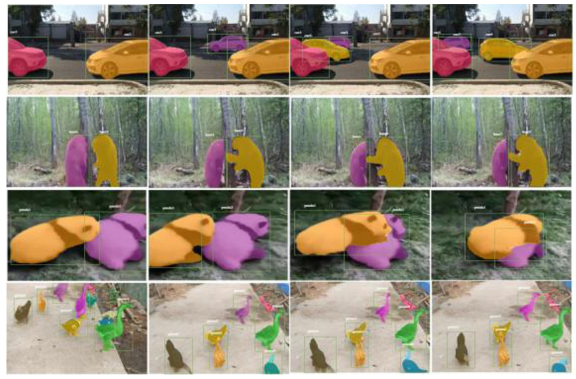

图 3 本文方法在 OVIS 验证数据集上的可视化效果

Fig. 3 Visualization effects of the proposed method on OVIS validation set

在图 4 中，该模型可以在有快速移动的人、滑板和变形的

鹰的场景中准确地检测和分割目标，以及区分猎豹的不同实例，对被树木遮挡的大熊猫仍能进行精确的分割。

图 4　本文方法在 YouTube-VIS 验证数据集上的可视化效果

Fig. 4　Visualization effects of the proposed method on YouTuber VIS validation set

图 5 给出了 3 个具有不同程度遮挡的实例的跟踪和分割结果。从上到下分别为存在轻度遮挡的熊猫、存在中度遮挡的猫和存在重度遮挡的猴子。实验结果表明，本文提出的方法可以全程跟踪并分割存在轻度遮挡的熊猫。在存在中度遮挡的场景中，绿色掩码的猫几乎被紫色掩码的猫完全遮挡，但本文方法仍然能够成功地进行跟踪。在存在严重遮挡的场景中，红色掩码的猴子的躯体被树叶完全遮挡，导致分割不准确，但黄色掩码的猴子和紫色掩码的猫仍能够被准确地分割。

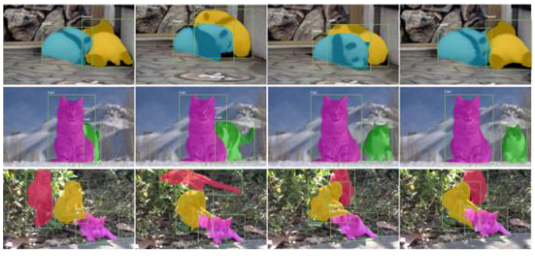

图 5　不同遮挡程度的对比实验

Fig. 5　Comparative experiments of different occlusion degrees

5.3　消融实验

本节进行了消融实验，以说明本文方法的核心因素，比较结果如表 3 和表 4 所列。视频和图像的主要区别在于视频包含时间信息，而如何有效地学习和利用时间信息是视频理解的关键。所有模型都在 OVIS train 上训练 10 个 epoch，并在 OVIS 验证集上使用 ResNet-50 主干网络进行测试。

5.3.1　不同特征提取网络对比实验

本小节主要通过添加不同的主干网络特征在 OVIS 数据集上进行消融实验。

表 3　用于掩码预测的 CNN 编码特征与 OVTR 编码特征

Table 3　CNN encoding features and OVTR encoding features for mask prediction

（%）

方法	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
CNN	13.5	27.2	10.5	7.7	17.7
OVTR	**15.2**	**31.4**	**11.5**	**9.2**	**20.1**

如表 3 所列，基于 CNN 编码的特征，AP 值达到 13.5%，而使用 OVTR 编码的特征后，AP 值提高了 1.7 个百分点。这表明，在 Transformer 通过自注意力机制对特征进行更新时，可以更好地学习特征之间的相互关系。此外，实验结果还表明 OVTR 模型在对空间和时间特征进行整体建模时具有更强的性能优势。

5.3.2　跟踪检测特征融合模块对比实验

本小节通过对网络中跟踪检测特征融合模块进行消融实验，以证明该模块的有效性。

表 4　跟踪检测特征融合模块

Table 4　Tracking and detecting feature fusion module

（%）

方法	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
去除 FTD	14.1	30.4	10.5	9.7	18.7
本文方法	**14.8**	**30.7**	**11.1**	**8.9**	**19.6**

如表 4 所列，FTD 模块能够预测具有大范围运动的物体的跟踪偏移，并提供强大的运动线索，为方法带来约 0.7% AP 的增益，有效地将跟踪检测特征传递给分割模块，提升遮挡目标分割的准确性。

改进后的特征提取网络 OVTR 能够捕捉不同区域之间的长程依赖关系，并且在处理时能够给予不同位置的特征不同的权重，使特征提取网络更好地利用不同帧之间的时序信息，并且捕捉物体在整个视频序列中的演变过程。跟踪检测特征融合模块实现了当前帧和参考帧已跟踪的信息特征融合，从而更好地提升遮挡场景的分割效果。

结束语　针对目前视频实例分割中忽略遮挡现象导致被遮挡视频分割效果不理想的问题，提出了一种基于时序信息跟踪检测特征融合的遮挡视频实例分割模型。该模型将时间维度引入 Transformer 以更好地捕捉其他帧中未遮挡的物体与当前帧中物体之间的精确关联信息。通过计算视频帧中产生的跟踪偏移和利用预测的跟踪偏移矩阵，将跟踪线索从参考帧变换传播到当前帧来完善增强特征图，并提高后续分割任务的有效性。对 OVIS 和 YouTube-VIS 数据集上的模型进行实验，验证了该模型在分割精度方面优于现有的基准方法。未来我们将继续考虑只使用 Transformer 提取特征，协同分割、检测和跟踪任务。

参 考 文 献

[1] QI J Y, GAO Y, HU Y, et al. Occluded video instance segmentation: A benchmark [J]. International Journal of Computer Vision, 2022, 130(8): 2022-2039.

[2] YANG L J, FAN Y C, XU N. Video instance segmentation [C]// International Conference on Computer Vision. 2019: 5188-5197.

[3] HE K M, GKIOXARI G, DOLLAR P, et al. Mask R-CNN[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2961-2969.

[4] BERTASIUS G, TORRESANI L. Classifying, segmenting, and tracking object instances in video with mask propagation[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2020: 9739-9748.

[5] DAI J F, QI H Z, XIONG Y W, et al. Deformable convolutional networks[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2017: 764-773.

[6] ATHAR A, MAHADEVAN S, OSEP A, et al. Stem-seg: Spatio-temporal embeddings for instance segmentation in videos[C]// European Conference on Computer Vision. 2020: 158-177.

[7] FU Y, YANG L J, LIU D, et al. Complete: Comprehensive feature aggregation for video instance segmentation[C]// Conference on Artificial Intelligence. 2021, 35(2): 1361-1369.

[8] WANG Y Q, XU Z L, WANG X L, et al. End-to-end video instance segmentation with transformers[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2021: 8741-8750.

[9] PARMAR N, VASWANI A, USZKOREIT J, et al. Image transformer[C]// International Conference on Machine Learning. 2018: 4055-4064.

[10] ZHU X Z, SU W J, LU L W, et al. Deformable DETR: Deformable transformers for end-to-end object detection [J]. arXiv: 2010.04159, 2021.

[11] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale [J]. arXiv: 2010.11929, 2021.

[12] LI Z S, LIU X T, DRENKOW N, et al. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers[C]// IEEE International Conference on Computer Vision. 2021: 6197-6206.

[13] LIU R J, YUAN Z J, LIU T, et al. End-to-end lane shape prediction with transformers[C]// IEEE Winter Conference on Applications of Computer Vision. 2021: 3694-3702.

[14] LIU CHANG, YUAN W J, WEI Z Q, et al. Location-aware predictive beamforming for UAV communications: A deep learning approach [J]. IEEE Wireless Communications Letters, 2020, 10(3): 668-672.

[15] ZHAO H S, JIA J Y, KOLTUN V. Exploring self-attention for image recognition[C]// IEEE International Conference on Computer Vision. 2020: 10076-10085.

[16] WANG H Y, ZHU Y K, GREEN B, et al. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation [C]// European Conference on Computer Vision. 2020: 108-126.

[17] QI J Y, GAO Y, HU Y, et al. Occluded video instance segmentation: A benchmark [J]. International Journal of Computer Vision, 2022, 130(8): 2022-2039.

[18] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C]// IEEE International Conference on Computer Vision. 2016: 770-778.

[19] LOSHCHILOV I, HUTTER F. Fixing weight decay regularization in adam[C]// International Conference on Learning Representations. 2018.

[20] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common objects in context[C]// European Conference on Computer Vision. 2014: 740-755.

[21] YANG S S, FANG Y X, WANG X G, et al. Crossover learning for fast online video instance segmentation[C]// IEEE International Conference on Computer Vision. 2021: 8043-8052.

[22] VOIGTLAENDER P, CHAI Y, SCHROFF F, et al. Feelvos: Fast end-to-end embedding learning for video object segmentation[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2019: 9481-9490.

[23] BOCHINSKI E, EISELEIN V, SIKORA T. High-speed tracking-by-detection without using image information[C]// IEEE International Conference on Advanced Video and Signal Based Surveillance. 2017: 1-6.

[24] WU J L, CAO J L, SONG L C, et al. Track to detect and segment: An online multi-object tracker[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2021: 12352-12361.

[25] CAO J L, ANWER R M, CHOLAKKAL H, et al. Sipmask: Spatial information preservation for fast image and video instance segmentation[C]// European Conference on Computer Vision. 2020: 1-18.

ZHENG Shenhai, born in 1988, Ph. D, associate professor. His main research interests include machine learning, pattern recognition and medical image computing.