中国科学院软件研究所学术年会’2021
暨计算机科学国家重点实验室开放周

学术论文

A New Siamese Co-attention Network for Unsupervised Video Object Segmentation

张正昊
2021 The 6th International Conference on Computer and Communication Systems
*e-mail：zhenghao2020@iscas.ac.cn, *tel 18610768277

概述

•无监督视频物体分割任务在推理时不提供任何标注信息，需要网络自行判断视频中的显著物体并进行输出。
•目前主流的无监督视频物体分割算法大多基于光流法或者循环神经网络（RNN），往往建立局部依赖，十分容易随着时间的推移积累错误。

目标

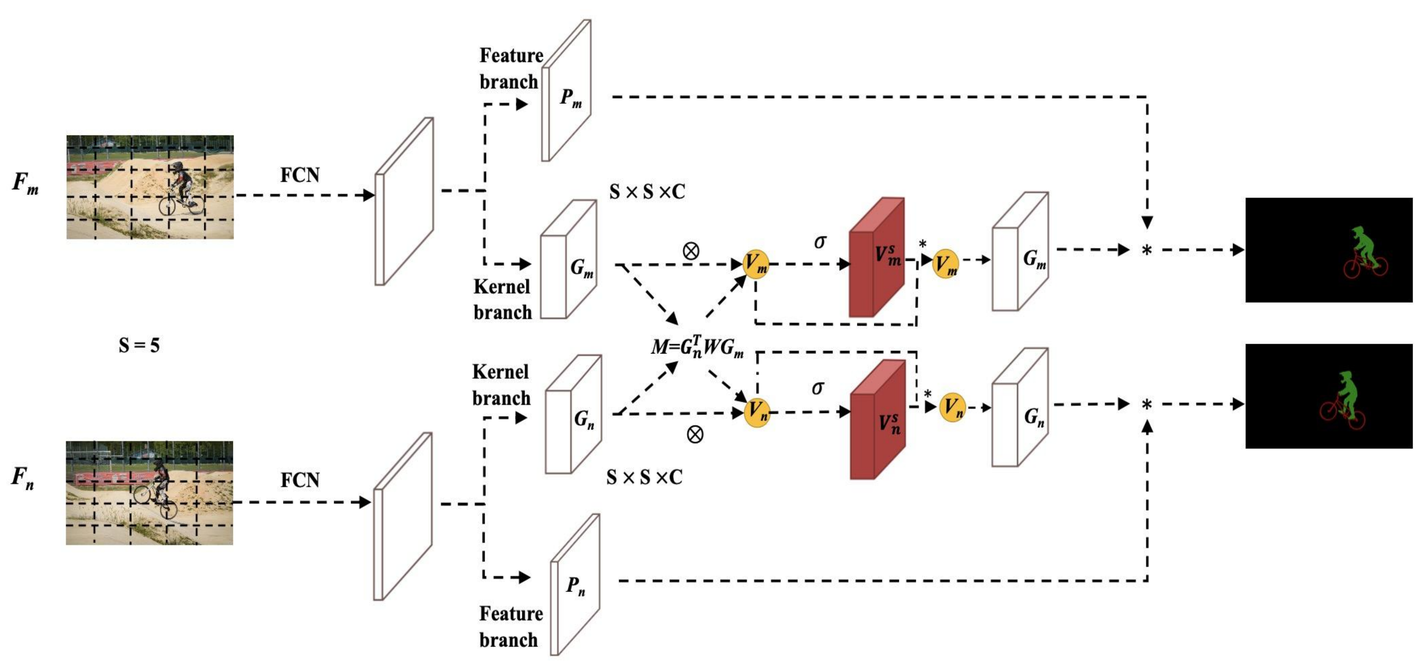
•网络可以准确判断出每个显著性物体，同时有效利用视频全局语义信息，能够较好处理遮挡、消失重现的场景。

思路

- 设计了一种新的孪生分割网络，利用Co-attention机制捕获同一个视频序列中任意两帧之间的长依赖关系，在推理的时候能够利用不同帧辅助分割，从视频全局语义角度出发，较好建模全局依赖关系，有效提高遮挡、消失重现等复杂场景下的分割精度。

网络架构设计

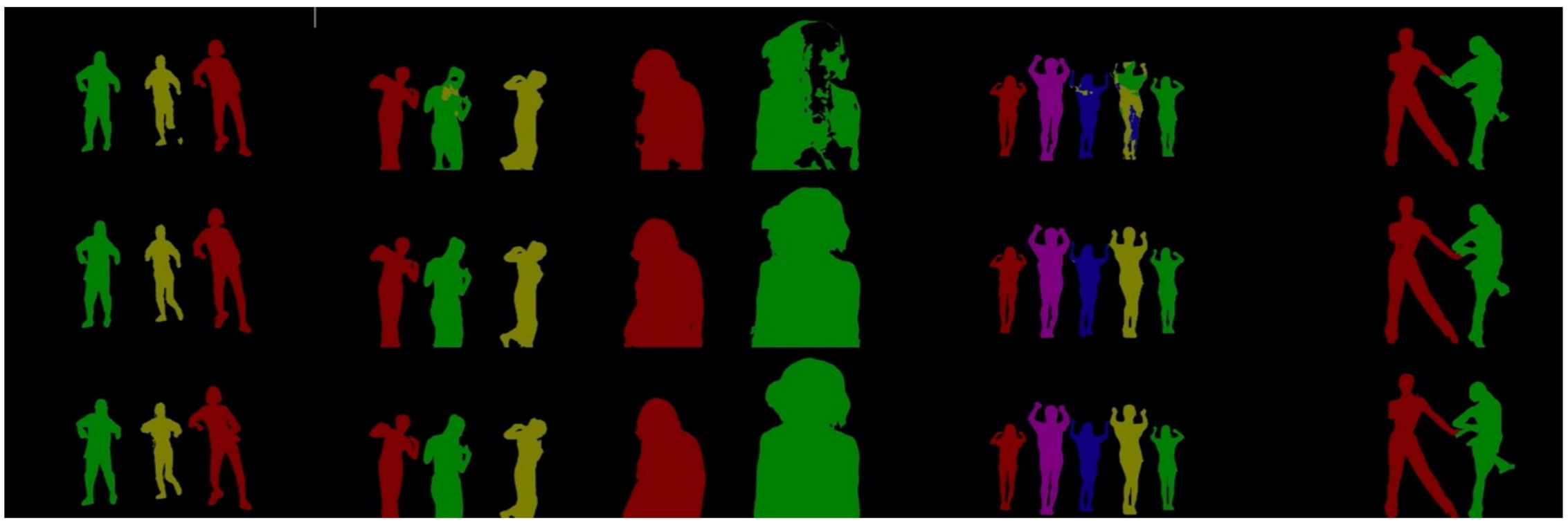
- 以SOLOv2为基础分割器，训练时，选取同一视频中任意两帧图像，经FCN提取得到两帧各自对应的特征分支与核分支，核分支编码了视频中物体的位置信息与高级语义信息，对两个核分支特征利用Co-attention模块，分别得到相应的相似度矩阵，利用相似度矩阵对两帧分别进行协同感应，建模两帧的物体位置与语义信息关联，从而有效辅助相互分割。



Co-attention设计

•对两帧对应的核分支特征G_m^i和G_n^i提取相似度矩阵 $M = {G_m^i}^T W G_n^i \epsilon \mathbb{R}^{(S^2 \times S^2)}$
•对相似度矩阵M分别进行行归一化与列归一化 $M^c = sigmoid(M), \quad M^r = sigmoid(M^T)$
•计算两帧分别对应的协同注意力特征图 $V_m = G_m \otimes M^c \ \epsilon \mathbb{R}^{(C \times S^2)}, V_n = G_n \otimes M^r \ \epsilon \mathbb{R}^{(C \times S^2)}$
•拼接协同注意力特征图与原始核分支特征图并经过卷积，组归一化与ReLU激活，得到新加权核分支特征图。
•$G_m^i = f([G_m^i, V_m]) \epsilon \mathbb{R}^{(S \times S \times C)}$
•$G_n^i = f([G_n^i, V_n]) \epsilon \mathbb{R}^{(S \times S \times C)}$

结果



从上到下的分割结果分别为SOLOv2，CosNet，ground truth