International Journal of Computer Vision
https://doi.org/10.1007/s11263-022-01703-8

Recurrent Graph Neural Networks for Video Instance Segmentation

Emil Brissman[1,4] · **Joakim Johnander**[1,5] · **Martin Danelljan**[3] · **Michael Felsberg**[1,2]

Received: 14 February 2022 / Accepted: 13 October 2022

Abstract
Video instance segmentation is one of the core problems in computer vision. Formulating a purely learning-based method, which models the generic track management required to solve the video instance segmentation task, is a highly challenging problem. In this work, we propose a novel learning framework where the entire video instance segmentation problem is modeled jointly. To this end, we design a graph neural network that in each frame jointly processes all detections and a memory of previously seen tracks. Past information is considered and processed via a recurrent connection. We demonstrate the effectiveness of the proposed approach in comprehensive experiments. Our approach operates online at over 25 FPS and obtains 16.3 AP on the challenging OVIS benchmark, setting a new state-of-the-art. We further conduct detailed ablative experiments that validate the different aspects of our approach. Code is available at https://github.com/emibr948/RGNNVIS-PlusPlus.

Keywords Detection · Tracking · Segmentation · Video

1 Introduction

Video instance segmentation (VIS) is the task of simultaneously detecting, segmenting, and tracking object instances from a set of predefined classes. This task has a wide range of applications in autonomous driving (Cordts et al., 2016; Yu et al., 2020), data annotation (Izquierdo et al., 2019; Berg et al., 2019), and biology (T'Jampens et al., 2016; Zhang et al., 2008; Burghardt & Ćalić, 2006). In contrast to image instance segmentation, the temporal aspect of its video counterpart poses several additional challenges. Preserving correct instance identities across frames is made difficult by the presence of other similar instances. Objects may be subject to occlusions, fast motion, or major appearance changes. Moreover, the videos can be subject to wild camera motion and severe background clutter.

Prior work on video instance segmentation has taken inspiration from related areas of multiple object tracking, video object detection, instance segmentation, and video object segmentation (Yang et al., 2019; Athar et al., 2020; Bertasius & Torresani, 2020). Most methods adopt the tracking-by-detection paradigm popular in multiple object tracking (Brasó & Leal-Taixé, 2020). In this paradigm, an instance segmentation method provides detections in each frame, reducing the task to the formation of *tracks*. Given a set of already initialized tracks, one must determine for each detection whether it belongs to one of the tracks, if it is a false positive, or if it should initialize a new track. Most approaches (Yang et al., 2019; Cao et al., 2020; Bertasius & Torresani, 2020; Luiten et al., 2019) learn to match pairs of detections and then rely on heuristics to form the final output, *e.g.*, initializing new tracks, predicting confidences, removing tracks, and predicting class memberships.

Communicated by Alexander Schwing.

These authors contributed equally to this work.

✉ Joakim Johnander
joakim.johnander@liu.se

Emil Brissman
emil.brissman@liu.se

Martin Danelljan
martin.danelljan@vision.ee.ethz.ch

Michael Felsberg
michael.felsberg@liu.se

[1] Computer Vision Laboratory, Department of Electrical Engineering, Linköping University, Linköping, Sweden

[2] School of Engineering, University of KwaZulu-Natal, Durban, South Africa

[3] Computer Vision Lab, ETH Zürich, Zürich, Switzerland

[4] Saab, Linköping, Sweden

[5] Zenseact, Göteborg, Sweden

Published online: 18 November 2022