A New Tracking-by-Attention Paradigm

Laura Leal-Taixe
Technical University of Munich

The challenging task of multi-object tracking (MOT) requires simultaneous reasoning about track initialization, identity, and spatiotemporal trajectories. We formulate this task as a frame-to-frame set prediction problem and introduce TrackFormer, an end-to-end MOT approach based on an encoder-decoder Transformer architecture. Our model achieves data association between frames via attention by evolving a set of track predictions through a video sequence. The Transformer decoder initializes new tracks from static object queries and autoregressively follows existing tracks in space and time with the new concept of identity preserving track queries. Both decoder query types benefit from self- and encoder-decoder attention on global frame-level features, thereby omitting any additional graph optimization and matching or modeling of motion and appearance. TrackFormer represents a new tracking-by-attention paradigm and yields state-of-the-art performance on the task of multi-object tracking (MOT17) and segmentation (MOTS20).