EViT: Efficient Vision and Tracking

Prasanth R S
Department of Information Technology, Government Engineering College, Barton Hill, Thiruvananthapuram, India

Gautham J K
Undergraduate Student, Department of Information Technology, Government Engineering College, Barton Hill, Thiruvananthapuram, India

Yedukrishnan S
Department of Information Technology, Government Engineering College, Barton Hill, Thiruvananthapuram, India

J Govind, Ananya K A
Department of Information Technology, Government Engineering College, Barton Hill, Thiruvananthapuram, India

Abstract:

Real-time object segmentation and multi-object tracking are fundamental capabilities in modern computer vision systems used in surveillance, robotics, and autonomous navigation. However, deploying such systems on edge devices remains challenging due to limited computational power, restricted memory capacity, and strict energy constraints. Existing approaches typically trade off between accuracy and efficiency: transformer-based architectures provide high segmentation accuracy but are computationally expensive, while lightweight convolutional models achieve faster inference at the cost of reduced performance. This research introduces EViT (Efficient Vision and Tracking), a hierarchical transformer-based framework designed for efficient real-time segmentation and tracking on resource-constrained edge platforms. The proposed architecture employs overlapping patch embeddings and spatially reduced self-attention within a four-stage hierarchical encoder to extract multi-scale features while significantly lowering computational complexity. A lightweight all-MLP decoder performs efficient multi-scale feature fusion to generate segmentation outputs with minimal parameter overhead. To maintain consistent object identities across frames, a tracking module combining a Kalman filter with Intersection-over-Union (IoU)-based data association is integrated into the pipeline. Experimental evaluations demonstrate competitive segmentation and tracking performance while maintaining real-time inference speeds on edge hardware, achieving approximately 15–20 FPS at 512×512 resolution with around 4M parameters.

Keywords:

Semantic segmentation, vision transformers, edge computing, multi-object tracking, real-time vision.