

# Structured Siamese Network for Real-Time Visual Tracking

Yunhua Zhang<sup>[0000–0003–3567–215X]</sup>, Lijun Wang<sup>[0000–0003–2538–8358]</sup>, Jinqing Qi<sup>[0000–0002–3777–2405]</sup>, Dong Wang<sup>[0000–0002–6976–4004]</sup>, Mengyang Feng<sup>[0000–0002–7112–4655]</sup>, and Huchuan Lu<sup>[0000–0002–6668–9758]</sup>

School of Information and Communication Engineering, Dalian University of Technology, China  
{zhangyunhua, wlj, mengyang\_feng}@mail.dlut.edu.cn  
{wdice, jinqing, hchuan}@dlut.edu.cn

**Abstract.** Local structures of target objects are essential for robust tracking. However, existing methods based on deep neural networks mostly describe the target appearance from the global view, leading to high sensitivity to non-rigid appearance change and partial occlusion. In this paper, we circumvent this issue by proposing a local structure learning method, which simultaneously considers the local patterns of the target and their structural relationships for more accurate target tracking. To this end, a local pattern detection module is designed to automatically identify discriminative regions of the target objects. The detection results are further refined by a message passing module, which enforces the structural context among local patterns to construct local structures. We show that the message passing module can be formulated as the inference process of a conditional random field (CRF) and implemented by differentiable operations, allowing the entire model to be trained in an end-to-end manner. By considering various combinations of the local structures, our tracker is able to form various types of structure patterns. Target tracking is finally achieved by a matching procedure of the structure patterns between target template and candidates. Extensive evaluations on three benchmark data sets demonstrate that the proposed tracking algorithm performs favorably against state-of-the-art methods while running at a highly efficient speed of 45 fps.

**Keywords:** Tracking, deep learning, siamese network

## 1 Introduction

Single object tracking is a fundamental problem in computer vision, where the target object is identified in the first video frame and successively tracked in subsequent frames. Although much progress has been made in the past decades, tremendous challenges still exist in designing a robust tracker that can well handle significant appearance changes, pose variations, severe occlusions, and background clutters with real-time speed.































