

# Unveiling the Power of Deep Tracking

Goutam Bhat<sup>1</sup>, Joakim Johnander<sup>1,2</sup>, Martin Danelljan<sup>1</sup>,  
Fahad Shahbaz Khan<sup>1,3</sup>, and Michael Felsberg<sup>1</sup>

<sup>1</sup> CVL, Department of Electrical Engineering, Linköping University, Sweden

<sup>2</sup> Zenuity, Sweden

<sup>3</sup> Inception Institute of Artificial Intelligence, Abu Dhabi, UAE

**Abstract.** In the field of generic object tracking numerous attempts have been made to exploit deep features. Despite all expectations, deep trackers are yet to reach an outstanding level of performance compared to methods solely based on handcrafted features. In this paper, we investigate this key issue and propose an approach to unlock the true potential of deep features for tracking. We systematically study the characteristics of both deep and shallow features, and their relation to tracking accuracy and robustness. We identify the limited data and low spatial resolution as the main challenges, and propose strategies to counter these issues when integrating deep features for tracking. Furthermore, we propose a novel adaptive fusion approach that leverages the complementary properties of deep and shallow features to improve both robustness and accuracy. Extensive experiments are performed on four challenging datasets. On VOT2017, our approach significantly outperforms the top performing tracker from the challenge with a relative gain of 17% in EAO.

## 1 Introduction

Generic object tracking is the problem of estimating the trajectory of a target in a video, given only its initial state. The problem is particularly difficult, primarily due to the limited training data available to learn an appearance model of the target online. Existing methods rely on rich feature representations to address this fundamental challenge. While handcrafted features have long been employed for this task, recent focus has been shifted towards deep features. The advantages of deep features being their ability to encode high-level information, invariant to complex appearance changes and clutter.

Despite the outstanding success of deep learning in a variety of computer vision tasks, its impact in generic object tracking has been limited. In fact, trackers based on handcrafted features [1, 7, 8, 22, 37] still provide competitive results, even outperforming many deep trackers on standard benchmarks [16, 36]. Moreover, contrary to the trend in image classification, object trackers do not tend to benefit from deeper and more sophisticated network architectures (see figure 1). In this work, we investigate the reasons behind the limited success of deep networks in visual object tracking.































