SpatialTracker: Tracking Any 2D Pixels in 3D Space

Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, Nan Xue, Sida Peng, Yujun Shen, Xiaowei Zhou

*Submitted on 5 Apr 2024 (*Accepted to CVPR 2024)

What is the problem?

The problem addressed in the paper is the challenge of estimating dense and long-range pixel motion in videos, particularly due to issues like occlusions, discontinuities in 2D motion, and the complexities that arise from projecting 3D motion into 2D. Traditional 2D tracking methods, like optical flow and feature tracking, do not fully capture motion across a sequence of video frames, especially in complex scenarios with frequent occlusions or out-of-plane rotations.

- Estimating dense and long-range pixel motion in videos is challenging due to:

- Occlusions and discontinuities in the 2D motion domain.
- Complexity arising from projecting 3D motion into 2D space.
- Traditional 2D tracking methods, such as optical flow and feature tracking, fail in scenarios involving frequent occlusions or out-of-plane rotations.

What has been done earlier?

- Prior approaches involved:
- **Optical flow:** Computed pixel-level motion between adjacent frames but was limited to short-term tracking.
- **Feature tracking:** Tracked sparse points but struggled with dense and long-range motion.
- **Particle Video:** Used semi-dense particles for long-range tracking but had difficulties recovering from occlusions.
- Recent supervised learning models showed promise but still faced limitations with complex motion scenarios like deformations and self-occlusions

However, these approaches either lack generalization to real-world conditions or struggle with occlusions and complex deformations. Several supervised learning-based models have been developed to track any point in videos but still faced limitations when handling complex scenarios.

What are the remaining challenges? What novel solution is proposed by the authors to solve the problem?

Despite progress, existing methods still struggle with complex motion scenarios, especially cases involving deformation and frequent self-occlusions. A key reason for this is that they only track motion in 2D space, ignoring the intrinsic 3D nature of motion. This limitation makes it difficult to accurately represent rotation and occlusion in 2D, leading to challenges in feature correlation and spatial reasoning near occlusion boundaries.

➢ Major challenges include:

- Existing methods struggle with complex deformations and frequent occlusions.
- They only track in 2D space, ignoring the 3D nature of motion, making it difficult to represent motion accurately during rotations or occlusions

➢ Suggested solutions

The authors propose **SpatialTracker**:

- It tracks 2D pixels in 3D space by lifting 2D pixels to 3D using monocular depth estimators.
- Utilizes **triplane representation** to encode the 3D scene of each frame efficiently.
- Uses a transformer for iterative 3D trajectory prediction, aided by an **as-rigid-as-possible (ARAP)** constraint to handle both rigid and deformable motions.
- Achieves state-of-the-art performance in complex scenarios such as out-of-plane rotations and occlusions.

They also incorporate an iterative trajectory prediction process using transformers, combined with an "as-rigid-as-possible" (ARAP) constraint to better handle rigid and deformable motions. The triplane representation and the ARAP rigidity embedding allow for tracking through complex motion, occlusions, and out-of-plane rotations, achieving state-of-the-art performance.

Ankit Saha, B421007

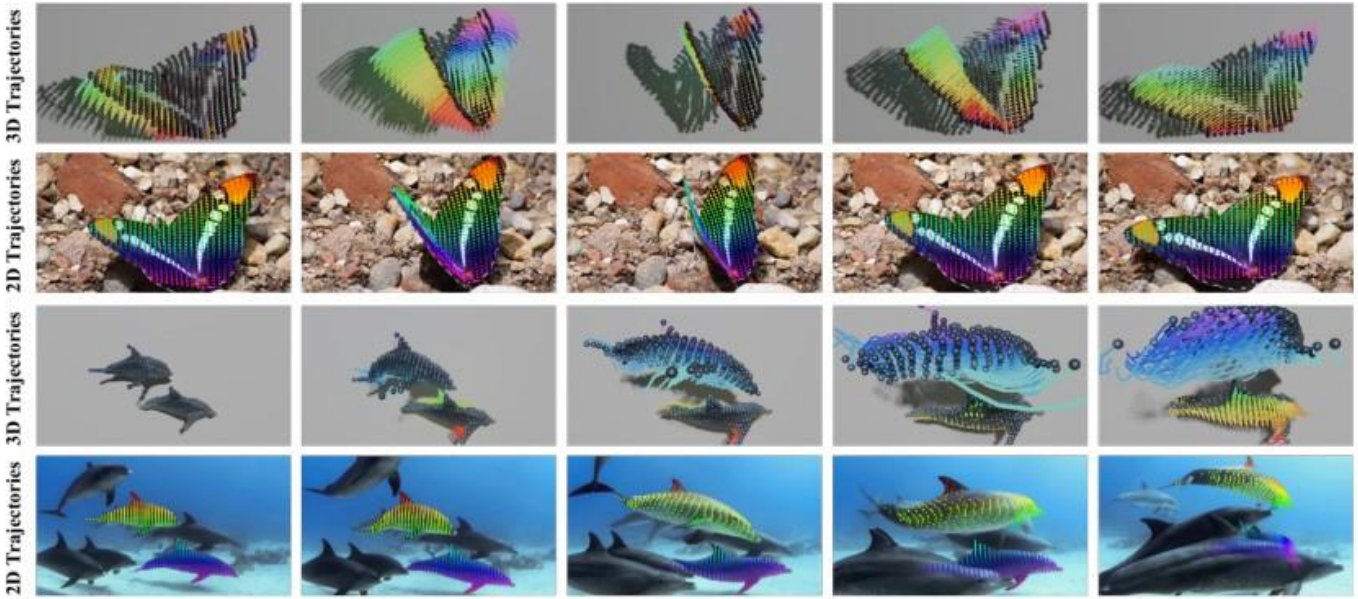

Figure 1. **Tracking 2D pixels in 3D space.** To estimate 2D motion under the occlusion and complex 3D motion, we lift 2D pixels into 3D and perform tracking in the 3D space. Two cases of the estimated 3D and 2D trajectories of a waving butterfly (top) and a group of swimming dolphins (bottom) are illustrated.