Easi3R: Estimating Disentangled Motion from DUSt3R Without Training

Xingyu Chen[1] Yue Chen[1] Yuliang Xiu[1,2] Andreas Geiger[3] Anpei Chen[1,3]
[1]Westlake University [2]Max Planck Institute for Intelligent Systems
[3]University of Tübingen, Tübingen AI Center
easi3r.github.io

Abstract

Recent advances in DUSt3R have enabled robust estimation of dense point clouds and camera parameters of static scenes, leveraging Transformer network architectures and direct supervision on large-scale 3D datasets. In contrast, the limited scale and diversity of available 4D datasets present a major bottleneck for training a highly generalizable 4D model. This constraint has driven conventional 4D methods to fine-tune 3D models on scalable dynamic video data with additional geometric priors such as optical flow and depths. In this work, we take an opposite path and introduce Easi3R, a simple yet efficient training-free method for 4D reconstruction. Our approach applies attention adaptation during inference, eliminating the need for from-scratch pre-training or network fine-tuning. We find that the attention layers in DUSt3R inherently encode rich information about camera and object motion. By carefully disentangling these attention maps, we achieve accurate dynamic region segmentation, camera pose estimation, and 4D dense point map reconstruction. Extensive experiments on real-world dynamic videos demonstrate that our lightweight attention adaptation significantly outperforms previous state-of-the-art methods that are trained or fine-tuned on extensive dynamic datasets.

1. Introduction

Recovering geometry and motions from dynamic image collections is still a fundamental challenge in computer vision, with broad downstream applications in novel view synthesis, AR/VR, autonomous navigation, and robotics. The literature commonly identifies this problem as Structure-from-Motion (SfM) and has been the core focus in 3D vision over decades, yielding mature algorithms that perform well under stationary conditions and wide baselines. However, these algorithms often fail when applied to dynamic video input.

The main reason for the accuracy and robustness gap between static and dynamic SfM is object dynamics, a common component in real-world videos. Moving objects violate fundamental assumptions of homography and epipolar consistency in traditional SfM methods [37, 48]. In addition, in dynamic videos, where camera and object motions are often entangled, these methods struggle to disentangle the two motions, often causing the motion with rich texture to mainly contribute to camera pose estimation erroneously. Recent efforts, such as MonST3R [73] and CUT3R [63], have made strides to address these challenges. However, their success is based on extensive training data [19, 25, 63, 68, 73] or task-specific prior models [22, 25, 73, 74], such as the depth, optical flow, and object mask estimators. These limitations motivate us to innovate further to minimize the gap between static and dynamic reconstruction.

We ask ourselves if there are lessons from human perception that can be used as design principles for dynamic 4D reconstruction: Human beings are capable of perceiving body motion and the structure of the scene, identifying

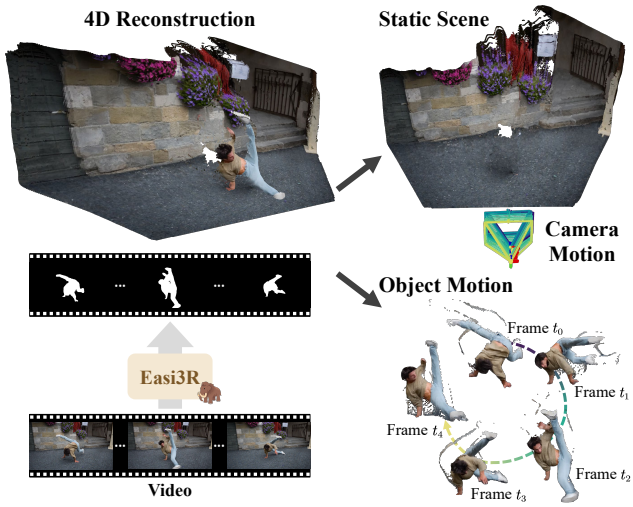

Figure 1. We present Easi3R, a training-free, plug-and-play approach that efficiently disentangles object and camera motion, enabling the adaptation of DUSt3R for 4D reconstruction.

dynamic objects, and disentangling ego-motion from object motion through the inherent attention mechanisms of the brain [58]. Yet, the learning process rarely relies on explicit dynamic labels.

We observe that DUSt3R implicitly learned a similar mechanism, and based on this, we introduce Easi3R, a training-free method to achieve dynamic object segmentation, dense point map reconstruction, and robust camera pose estimation from dynamic videos, as shown in Figure 1. DUSt3R uses attention layers at its core, taking two image features as input and producing pixel-aligned point maps as output. These attention layers are trained to directly predict pointmaps in the reference view coordinate space, implicitly matching the image features between the input views [4] and estimating the rigid view transformation in the feature space. In practice, performance drops significantly when processing pairs with object dynamics [73], as shown in Figure 2. By analyzing the attention maps in the transformer layers, we find that regions with less texture, under-observed, and dynamic objects can yield low attention values. Therefore, we propose a simple yet effective decomposition strategy to isolate the above components, which enables long-horizon dynamic object detection and segmentation. With this segmentation, we perform a second inference pass by applying a re-weighting [17] in the cross-attention layers, enabling robust dynamic 4D reconstruction and camera motion recovery without fine-tuning on a dynamic dataset, all at minimal additional cost to DUSt3R.

Despite its simplicity, we demonstrate that our inference-time scaling approach for 4D reconstruction is remarkably robust and accurate on in-the-wild casual dynamic videos. We evaluate our Easi3R adaptation on the DUSt3R and MonST3R backbones in three task categories: camera pose estimation, dynamic object segmentation, and pointcloud reconstruction in dynamic scenes. Easi3R performs surprisingly well across a wide range of datasets, even surpassing concurrent methods (e.g., CUT3R [63], MonST3R [73], and DAS3R [68]) that are trained on dynamic datasets.

2. Related Work

SfM and SLAM. Structure-from-Motion (SfM) [2, 41, 42, 48, 51, 52] and Simultaneous Localization and Mapping (SLAM) [9, 13, 32, 34] have long been the foundation for 3D structure and camera pose estimation. These methods are done by associating 2D correspondences [5, 10, 28, 32, 47] or minimizing photometric errors [12, 13], followed by bundle adjustment (BA) [3, 6, 55, 57, 59, 62] to refine structure and motion estimates. Although highly effective with dense input, these approaches often struggle with limited camera parallax or ill-posed conditions, leading to performance degeneracy. To overcome these limitations, DUSt3R [64] introduced a learning-based approach that directly predicts two pointmaps from an image pair in the coordinate space of the first view. This approach inherently matches image features and rigid body view transformation. By leveraging a Transformer-based architecture [11] and direct point supervision on large-scale 3D datasets, DUSt3R establishes a robust Multi-View Stereo (MVS) foundation model. However, DUSt3R and the follow-up methods [27, 33, 56, 61] assume primarily static scenes, which can lead to significant performance degradation when dealing with videos with dynamic objects.

Pose-free Dynamic Scene Reconstruction. Modifications to SLAM for dynamic scenes involve robust pose estimation to mitigate moving object interference, dynamic map management for updating changing environments, including techniques like semantic segmentation [72], optical flows [75], enhance SLAM's resilience in dynamic scenarios. Another line of work focuses on estimating stable video depth by incorporating geometric constraints [29] and generative priors [18, 49]. These methods enhance monocular depth accuracy but lack global point cloud lifting due to missing camera intrinsics and poses. For joint pose and depth estimation, optimization-based methods such as CasualSAM [74] fine-tune a depth network [45] at test time using pre-computed optical flow [66]. Robust-CVD [22] refines pre-computed depth [45] and camera pose by leveraging masked optical flow [16, 66] to improve stability in occluded and moving regions. Concurrently, MegaSaM [25] further enhances pose and depth accuracy by integrating DROID-SLAM [57], optical flow [66], and depth initializations from [40, 71], achieving state-of-the-art results. Alternatively, point-map-based approaches like MonST3R [73] extend DUSt3R to dynamic scenes by fine-tuning with dynamic datasets and incorporating optical flow [66] to infer dynamic object segmentation. DAS3R trains a DPT [44] on top of MonST3R, enabling feedforward segmentation estimation. CUT3R [63] fine-tunes MASt3R [24] on both static and dynamic datasets, achieving feedforward reconstruction but without predicting dynamic object segmentation, thereby entangling the static scene with dynamic objects. Although effective, these methods require costly training on diverse motion patterns to generalize well.

In contrast, we take an opposite path, exploring a training-free and plug-in-play adaptation that enhances the generalization of DUSt3R variants for dynamic scene reconstruction. Our method requires no fine-tuning and comes at almost no additional cost, offering a scalable and efficient alternative for handling real-world dynamic videos.

Motion Segmentation. Motion segmentation aims to predict dynamic object masks from video inputs. Classical approaches generally rely on optical flow estimation [26, 31, 67, 70] and point tracking [7, 21, 35, 50, 69] to distinguish moving objects from the background. Being trained solely on 2D data, they often struggle with occlusions and