PAGE-4D: Disentangled Pose and Geometry Estimation for 4D Perception

Kaichen Zhou · Yuhan Wang · Grace Chen · Xinhai Chang · Gaspard Beaudouin · Fangneng Zhan · Paul Pu Liang † · Mengyu Wang †

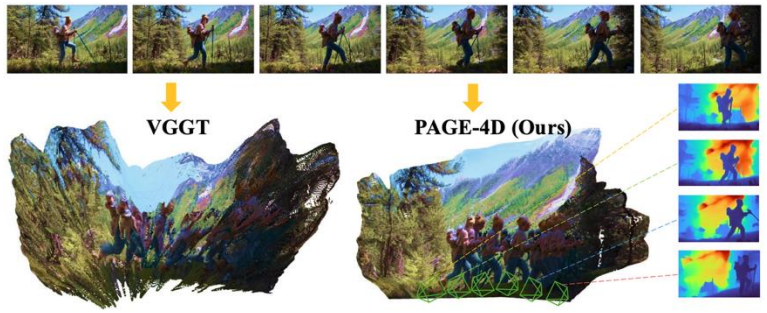
Harvard . MIT . Imperial College . ENPC

INTRODUCTION

The Problem

Static-only training → Dynamic-scene failure

Camera pose errors & geometry distortions when dynamic agents (people, vehicles) are present.



MOTIVATION - 1

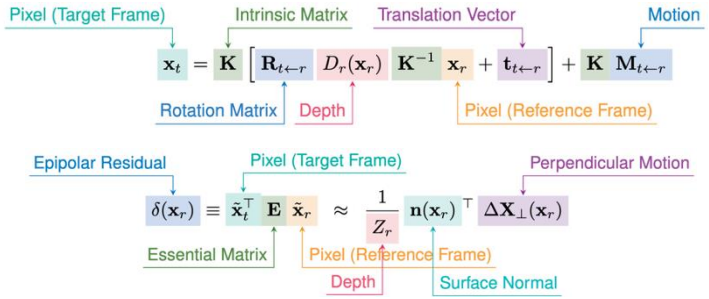
Conflicting Task Objectives

Multi-task 4D reconstruction faces a fundamental tension:

Pose Estimation
Needs to suppress dynamic regions

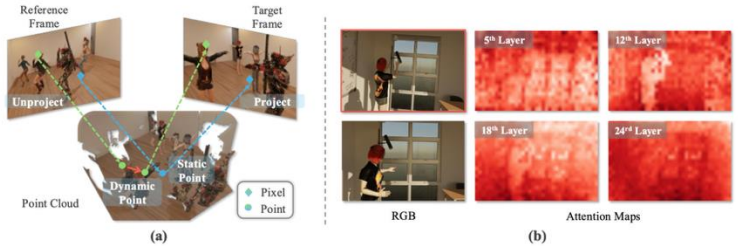
⟺

Geometry Recon.
Needs to model dynamic regions



MOTIVATION - 2

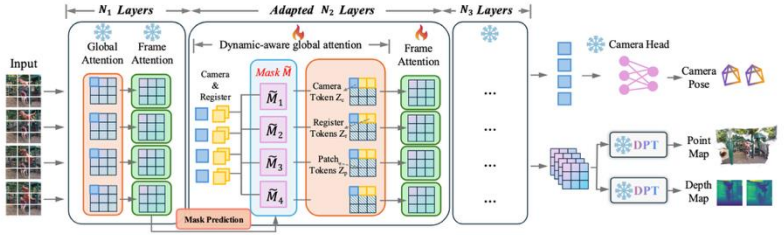
Empirical Observation



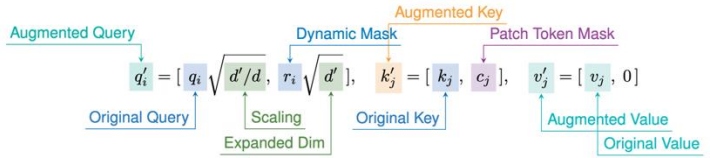
- Dynamic regions exhibit weaker activations compared to static ones, suggesting that static-only model tends to ignore dynamic content.
- Masking dynamic regions during the feedforward process can improve pose estimation performance.

METHOD: PAGE-4D ARCHITECTURE

Overview



- Training — Training only the middle layers to improve efficiency.
- Memory-Efficient Mask - Replaces the quadratic mask with a linear-memory additive mask while preserving attention equivalence.



EXPERIMENTAL RESULTS

Quantitative Results

>30% Pose Estimation	**>18%** Video Depth Estimation
>17% Monocular Depth Estimation	**>4%** View Synthesis

Qualitative Results