

# St4RTrack: RECONSTRUCTION VÀ TRACKING 4D MỘT CÁCH ĐỒNG THỜI

Môn học: CS519 - Phương pháp luận NCKH

Lớp: CS519.Q11.KHTN

GVHD: PGS.TS Lê Đình Duy

# Tóm tắt

- Link Github của nhóm:  
<https://github.com/NguyenPhamPhuongNam/CS519.Q11.KHTN.git>
- Link YouTube video: <https://youtu.be/4WvGUZqM1mU>

Nguyễn Phạm Phương Nam



Hồ Ngọc Luật



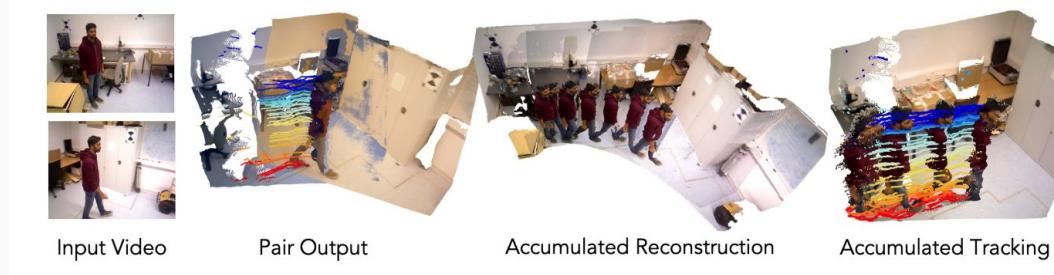
# Giới thiệu

## Reconstruction:

Tái tạo ảnh 3D đối với tọa độ World

## Tracking:

Theo dõi vị trí 3D của 1 điểm theo thời gian



- Các mô hình trước không gắn kết tự nhiên giữa hai bài toán
- Đề xuất: **St4RTrack** – mô hình học sâu feed-forward
  - **vừa Reconstruction 3D, vừa Tracking 3D** trực tiếp từ video RGB
  - làm việc trong **hệ tọa độ thế giới (world frame)**, tách được chuyển động camera và chuyển động của vật thể.

# Mục tiêu

- Khai thác lại sự cộng hưởng giữa **reconstruction 3D** và **correspondence 2D** ngay cả trong **cảnh động**, bằng cách đưa thêm thông tin **chuyển động 3D dày đặc** (3D point tracking).
- Xây dựng một khung thống nhất để **vừa reconstruction 3D, vừa tracking 3D** trong **hệ toạ độ thế giới**, tách được **chuyển động camera** và **chuyển động cảnh** từ video RGB.
- Cho phép **huấn luyện và thích nghi** trên video **in-the-wild** không có nhãn 4D đầy đủ
- Thiết lập **chuẩn đánh giá** trong **world frame** cho bài toán tracking và reconstruction 3D, hướng tới một hệ nhận thức **4D đa nhiệm, không phụ thuộc tác vụ**.

# Nội dung và Phương pháp

## Overview

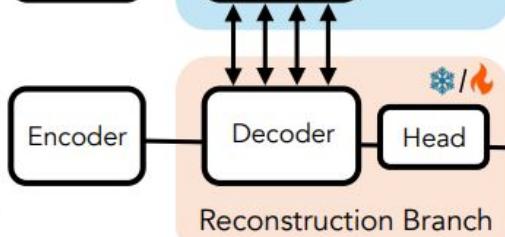
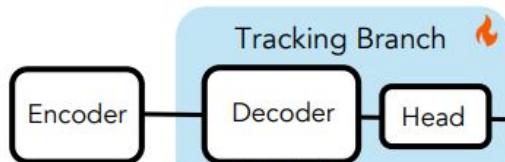
Frame 1



Frame j



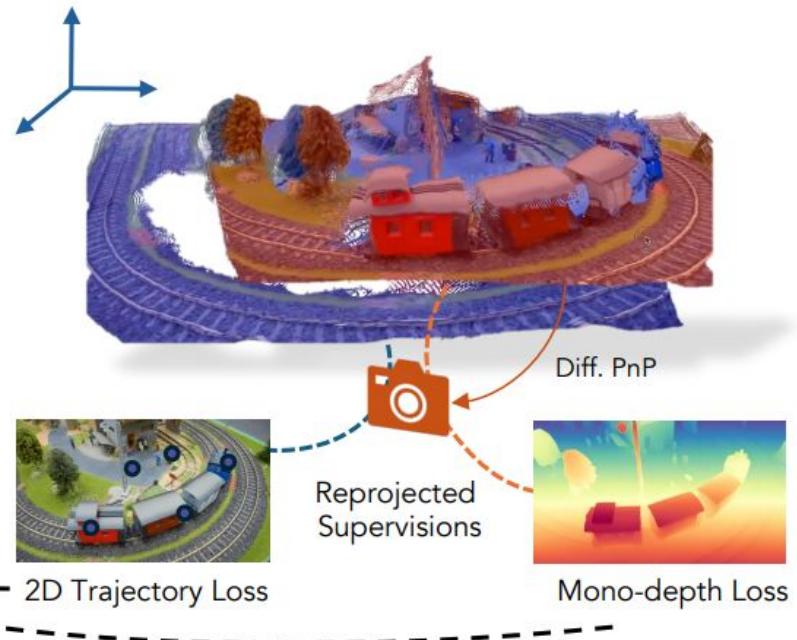
⋮



Gradients  
Back-propagating

$1X^1_j$

$1X^j$



# Nội dung và Phương pháp

## Biểu diễn 4D

Hợp nhất **reconstruction + long-term 3D tracking** bằng cách dự đoán **hai pointmaps phụ thuộc thời gian** trong cùng world frame.

- Input: video RGB, neo  $I_1$  làm world frame.

- Output:  $f(I_1, I_j) = \left( {}^1X_j^1, {}^1X_j^j \right)$

${}^1X_j^1$  : tracking (điểm frame 1 tại time j ).

${}^1X_j^j$  : reconstruction (frame j tại time j).

- Kết quả:

Tracking dài hạn:  $\{{}^1X_t^1\}_{t=1}^T$

Tái tạo động:  $\{{}^1X_t^j\}_{t=1}^T$

Mạng: ViT + Siamese Transformer + 2 head.

# Nội dung và Phương pháp

## Phương pháp huấn luyện và thích nghi video thật

### Pretrain trên dữ liệu 4D tổng hợp:

- Quy mô nhỏ, chuyển động & hình học chưa đa dạng như video thật.
  - Pointmap phải “di chuyển tự do” trong world frame → cần fine-tuning trên dữ liệu thực.
- 
- Loss tổng quát:  $L = L_{\text{traj}} + \lambda_1 L_{\text{depth}} + \lambda_2 L_{\text{align}}$

### Áp dụng lên dữ liệu thực:

- Giải camera pose  $(R_j, T_j)$  từ  ${}^1X_j^j$  (PnP+RANSAC, differentiable).
- Reproject điểm tracking  $\hat{x}_{j,n} = \pi(K(R_j^{-1}X_{j,n}^1 + T_j))$
- 2D track consistency: so với pseudo 2D tracks (CoTracker)
- Depth consistency: so với mono-depth (MoGe)
- 3D self-consistency: đồng bộ tracking với reconstruction tại time j

# Nội dung và Phương pháp

## Điểm cải tiến so với phương pháp gốc

- Trong video thật, occlusion/fast motion làm pseudo-label nhiễu, khiến việc tối ưu kéo lêch pose và pointmap -> mask che khuất và độ tin cậy theo điểm để chỉ học từ các điểm đáng tin.
- Tạo mask occlusion từ 2 pointmaps : reproject điểm tracking sang frame  $j$ , rồi so sánh depth của điểm đó với depth bề mặt visible từ reconstruction tại cùng pixel :  $m_n = \mathbf{1}[z_{j,n}^{\text{trk}} \leq z_j^{\text{rec}}(\hat{x}_{j,n}) + \tau]$
- Confidence/uncertainty theo điểm: Mạng dự đoán  $\sigma_n^2$  cho mỗi điểm: điểm kém tin ->  $\sigma_n^2$  lớn -> giảm trọng số khi tính loss
- Loss robust:  $L_{\text{traj}}^{\text{robust}}$  : reprojection 2D so với CoTracker, được lọc bởi  $m_n$  và được trọng số bởi  $\sigma_n^2$ .  
 $L_{\text{align}}^{\text{robust}}$  : nhất quán 3D giữa tracking và reconstruction, được lọc bởi  $m_n$  và được trọng số bởi  $\sigma_r^2$   
 $L_{\text{depth}}$  : depth consistency so với MoGe ( giữ nguyên).

# Kết quả dự kiến

- Xây dựng pipeline thông nhát tái tạo 3D động và tracking 3D dài hạn trong **world coordinate** từ video RGB.
- Cải tiến giai đoạn thích nghi video thật bằng **occlusion-aware mask + confidence weighting** để giảm outlier/occlusion, giúp tracking ổn định hơn.
- Đánh giá trên reconstruction & world-frame tracking với so sánh/ablation (baseline vs proposed) và trực quan hóa kết quả.

# Tài liệu tham khảo

- <https://arxiv.org/pdf/2504.13152.pdf> : St4RTrack: Simultaneous 4D Reconstruction and Tracking in the World
- Dust3r: Geometric 3d vision made easy. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 20697– 20709, 2023.
- <https://arxiv.org/abs/2410.03825> : MONST3R: A simple approach for estimating geometry in the presence of motion