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/CS519-Q11-KHTN
- Link YouTube video: https://youtu.be/4WvGUZqM1mU

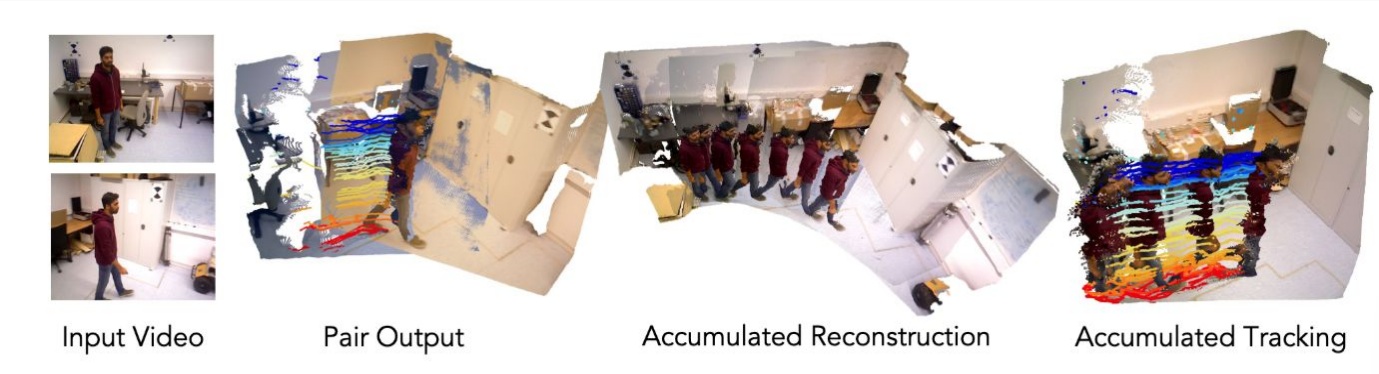
Nguyễn Phạm Phương Nam

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 toạ 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).

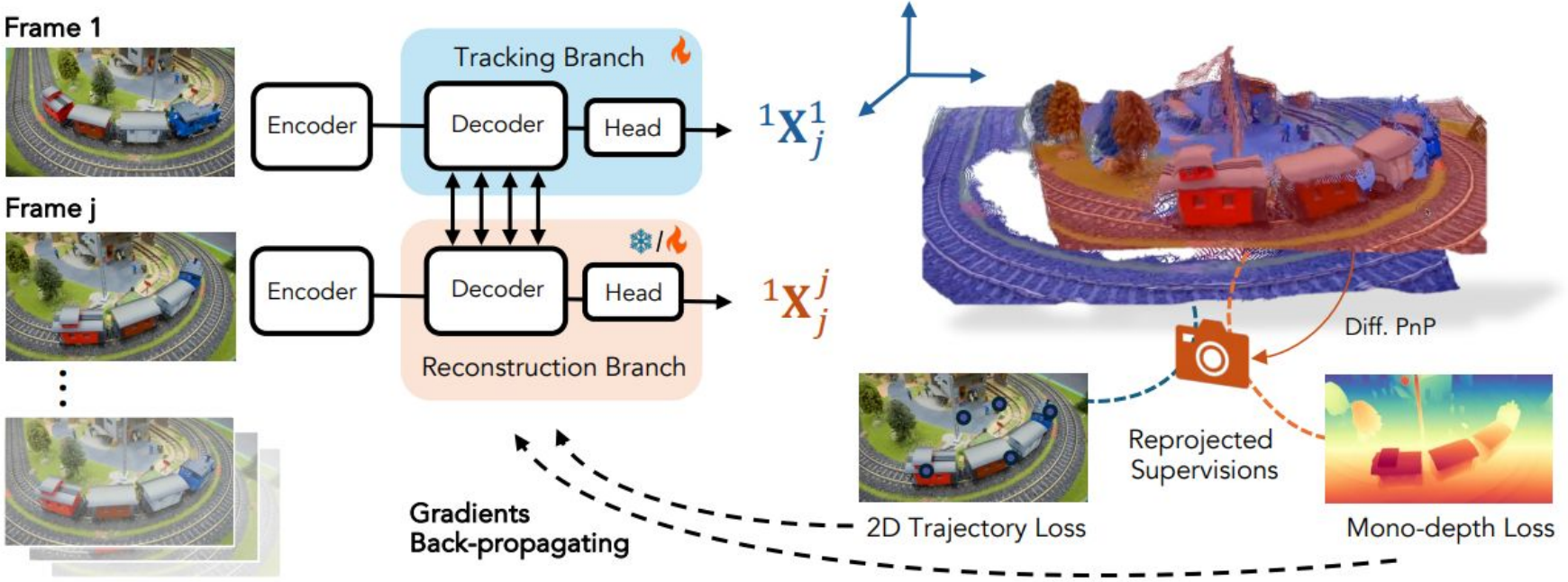
- Xây dựng một khung thống nhất để **vừa reconstruction 3D, vừa tracking 3D** trong **hệ toạ độ thế giới**, tách được **chuyển động camera** và **chuyển động cảnh** từ video RGB.

- 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



Nội dung và Phương pháp

Biểu diễn 4D

- **Time-dependent pointmap**
  - Mỗi ảnh I sinh ra một pointmap

$$\mathbf{X} \in \mathbb{R}^{H \times W \times 3}$$

    (mỗi pixel ↔ một điểm 3D).
  - Với cảnh động, pointmap phụ thuộc **thời gian**:

$${}^{a}\mathbf{X}_t^b : \text{hình học 3D của nội dung từ frame } b, \text{ tại thời điểm } t, \text{ trong hệ toạ độ frame } a.$$

  - Ví dụ: ${}^{i}\mathbf{X}_j^i$ = nội dung của I_i nhưng ở thời điểm j → quỹ đạo 3D của điểm trong I_i.
- **Hàm f_θ và hai pointmaps**

$$f_\theta(I_i, I_j) = \left({}^{i}\mathbf{X}_j^i,\ {}^{i}\mathbf{X}_j^j\right)$$

  - ${}^{i}\mathbf{X}_j^i$: **tracking** – điểm của I_i tại thời điểm j (bám điểm 3D).
  - ${}^{i}\mathbf{X}_j^j$: **reconstruction** – hình học 3D của I_j tại thời điểm j, trong hệ toạ độ I_i.
- **World frame cho cả video**
  - Chọn I_1 làm chuẩn:

$$\{f_\theta(I_1, I_1), f_\theta(I_1, I_2), \ldots, f_\theta(I_1, I_T)\}.$$

  - **Tracking world-frame**: $\{{}^{1}\mathbf{X}_1^1, {}^{1}\mathbf{X}_2^1, \ldots, {}^{1}\mathbf{X}_T^1\}$.
  - **Reconstruction world-frame**: $\{{}^{1}\mathbf{X}_1^1, {}^{1}\mathbf{X}_2^2, \ldots, {}^{1}\mathbf{X}_T^T\}$.

Nội dung và Phương pháp

Phương pháp huấn luyện

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.

Áp dụng lên dữ liệu thực:

Nội tại K: suy ra từ pointmap frame 1

Ngoại tại $P_j = [R_j \mid T_j]$: từ tương ứng 2D–3D $(x_{j,n} \leftrightarrow X^j_{j,n})$

Reprojection loss

- **Trajectory loss** L_{traj}: chiếu X^1_j lên frame j với (K, R_j, T_j), so với track 2D từ CoTracker (scale-invariant).
- **Mono-depth loss** L_{depth}: depth từ X^j_j (sau khi transform sang camera j) so với mono-depth MoGe, với hệ số tỉ lệ

$$\alpha^* = \frac{\sum_n z^{\text{proj}}_{j,n} z^{\text{mono}}_{j,n}}{\sum_n \left(z^{\text{proj}}_{j,n}\right)^2},$$

$$L_{\text{depth}} = \frac{1}{N} \sum_{n=1}^{N} \left(\alpha^* z^{\text{proj}}_{j,n} - z^{\text{mono}}_{j,n}\right)^2.$$

3D self-consistency

- L_{align}: với các điểm từ frame 1 còn thấy ở frame j, so sánh vị trí 3D từ X^1_j và X^j_j tại cùng timestamp.

Tổng loss tự giám sát

$$L_{\text{reproj}} = L_{\text{traj}} + \lambda_1 L_{\text{depth}} + \lambda_2 L_{\text{align}}.$$

Kết quả dự kiến

- Xây dựng một khung thống nhất để **vừa reconstruction 3D, vừa tracking 3D**.
- Đạt kết quả chấp nhận được trên cả bài toán reconstruction 3D và tracking 3D
- Cung cấp benchmark mới cho bài toán tracking và reconstruction 3D trên World Coordinate, giúp thống nhất hệ tọa độ vật lý.

Tài liệu tham khảo

- [1] Haiwen Feng, Junyi Zhang, Qianqian Wang, Yufei Ye, Pengcheng Yu, Michael J Black, Trevor Darrell, and Angjoo Kanazawa. St4rtrack: Simultaneous 4d reconstruction and tracking in the world. arXiv preprint arXiv:2504.13152, 2025
- [2] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In CVPR, 2024.
- [3] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. arXiv preprint arxiv:2410.03825, 2024.