

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



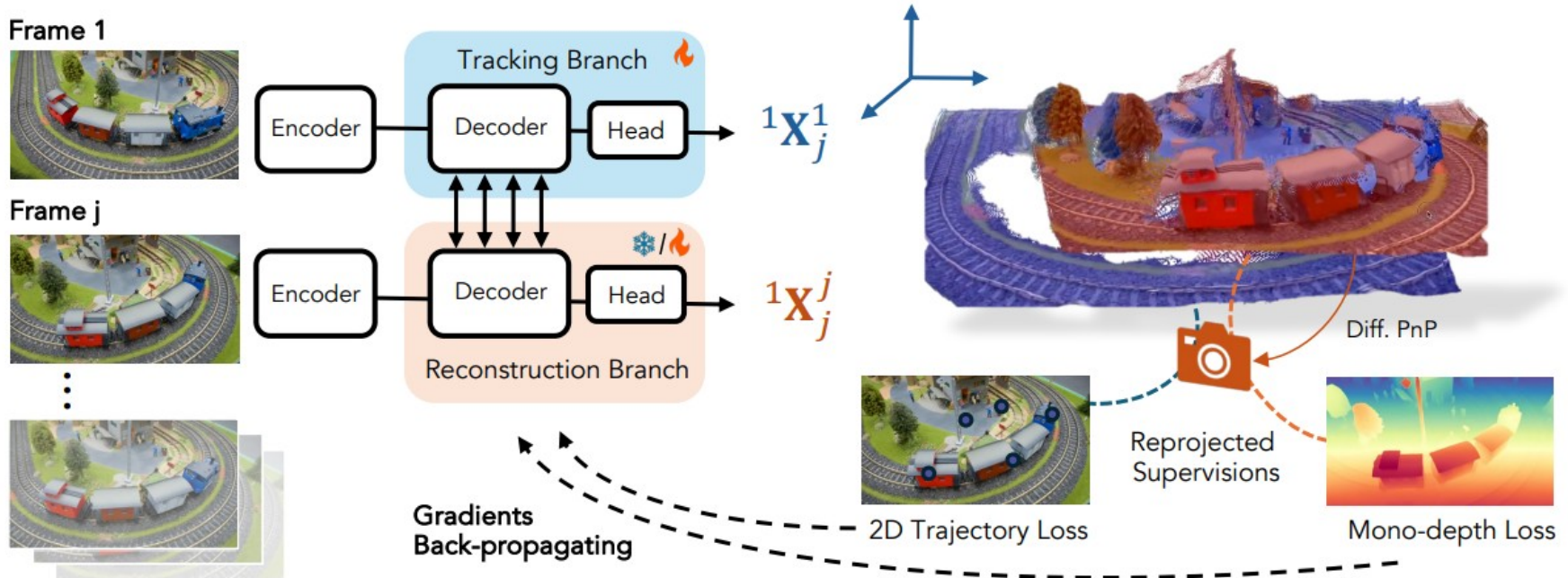
- Các mô hình trước không gắn kết tự nhiên giữa hai bài toán
- Đề xuất: **St4RTrack** – mô hình học sâu feed-forward
 - vừa **Reconstruction 3D**, vừa **Tracking 3D** trực tiếp từ video RGB
 - làm việc trong **hệ tọa độ thế giới (world frame)**, tách được chuyển động camera và chuyển động của vật thể.

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



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 \leftrightarrow một điểm 3D).

- Với cảnh động, pointmap phụ thuộc thời gian:

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

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

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

- ${}^i\mathbf{X}_j^i$: 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:

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

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

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 \rightarrow 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 ($x_{j,n} \leftrightarrow X_{j,n}^j$)

Reprojection loss

- Trajectory loss L_{traj} : chiếu X_j^1 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

$$\alpha^* = \frac{\sum_n z_{j,n}^{\text{proj}} z_{j,n}^{\text{mono}}}{\sum_n (z_{j,n}^{\text{proj}})^2},$$

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

3D self-consistency

- L_{align} : với các điểm từ frame 1 còn thấy ở frame j , so sánh vị trí 3D từ X_j^1 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

- <https://arxiv.org/pdf/2504.13152> : 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