CẢI TIẾN HÀM LOSS TỰ GIÁM SÁT NHẰM TĂNG ĐỘ BỀN VỮNG CHO ST4RTRACK TRÊN DỮ LIỆU VIDEO THỰC

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/CS519.Q11
- Link YouTube video: https://youtu.be/YaOfSHIF5Vk

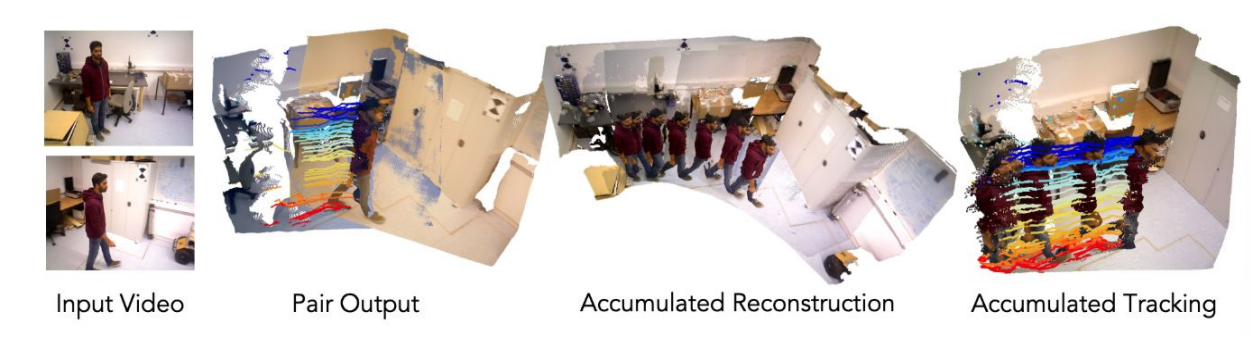
Nguyễn Phạm Phương Nam

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



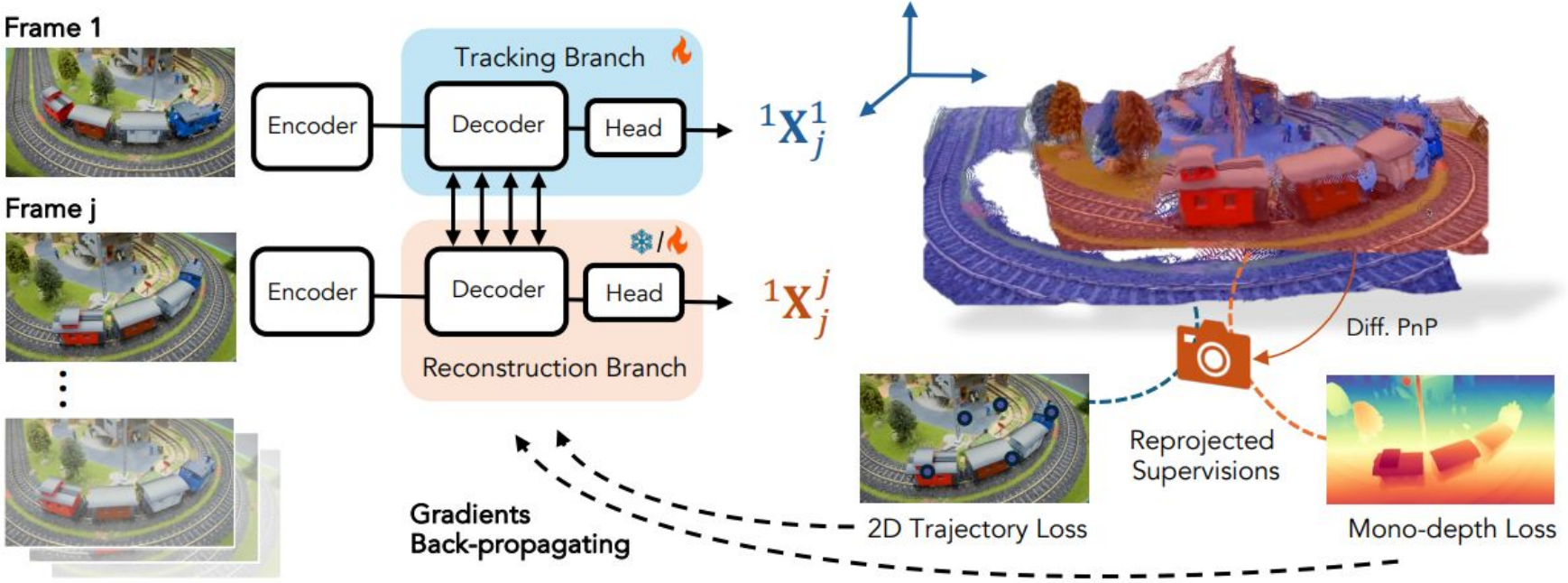
- **St4RTrack** kém ổn định khi áp dụng trên dữ liệu video thực do occlusion, chuyển động nhanh và nhiễu pseudo-label.
- Đề xuất: **St4RTrack+** – mô hình cải tiến về độ bền vững cho St4RTrack, với:
  → **Occlusion/visibility mask dựa trên kiểm tra độ sâu** nhằm loại bỏ các điểm không quan sát được
  → **Confidence/uncertainty weighting theo điểm** để tự động giảm ảnh hưởng của các track không đáng tin cậy trong quá trình tối ưu.

Mục tiêu

- **Thiết kế và cải tiến hệ hàm loss tự giám sát cho ST4RTRACK:** đề xuất St4RTrack+ với các hàm loss reprojection có xét ràng buộc hình học, occlusion và độ bất định nhằm giảm ảnh hưởng của nhiễu và outlier trong quá trình huấn luyện trên dữ liệu video thực.
- **Nâng cao độ bền vững và khả năng thích nghi của mô hình trên dữ liệu thực:** cải thiện tính ổn định khi tối ưu ST4RTRACK trong các kịch bản test-time adaptation và domain-level adaptation, đặc biệt trong điều kiện occlusion, chuyển động phức tạp và sai lệch miền dữ liệu.
- **Đánh giá hiệu quả cải tiến trên các tập dữ liệu video thực tế:** phân tích định lượng và định tính mức độ cải thiện về độ chính xác tracking và tính ổn định của mô hình so với ST4RTRACK gốc, qua đó làm rõ vai trò của từng thành phần loss được đề xuất.

Nội dung và Phương pháp

Overview về St4RTrack



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

- 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.

Áp dụng lên dữ liệu thực:

- Giải camera intrinsic K từ ${}^{1}X_{j}^{1}$ và camera pose (R_j, T_j) từ ${}^{1}X_{j}^{j}$ (PnP solver + RANSAC).
- Reproject điểm tracking $\hat{x}_{j,n} = \pi\Big(K\big(R_j\,{}^{1}X_{j,n}^{1} + T_j\big)\Big)$
- 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

- Loss tổng quát: $L = L_{\text{traj}} + \lambda_1 L_{\text{depth}} + \lambda_2 L_{\text{align}}$

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 lệch 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 : $m_n = \mathbf{1}\left[z_{j,n}^{\mathrm{trk}} \le z_j^{\mathrm{rec}}\left(\hat{x}_{j,n}\right) + \tau\right]$
- Confidence/uncertainty theo điểm: Mạng dự đoán σ_n^2 cho mỗi điểm: điểm kém tin -> σ_n^2 lớn -> giảm trọng số khi tính loss
- Loss robust: $L_{\mathrm{traj}}^{\mathrm{robust}}$: reprojection 2D so với CoTracker, được lọc bởi m_n và được trọng số bởi σ_n^2 .

  $L_{\mathrm{align}}^{\mathrm{robust}}$: nhất quán 3D giữa tracking và reconstruction, được lọc bởi m_n và được trọng số bởi σ_n^2

  L_{depth} : depth consistency so với MoGe (giữ nguyên).

Kết quả dự kiến

- **Đề xuất một hệ hàm loss tự giám sát robust cho ST4RTRACK**, có xét occlusion và độ bất định, giúp giảm ảnh hưởng của nhiễu và outlier khi huấn luyện trên dữ liệu video thực.

- **Cải thiện độ ổn định và khả năng thích nghi của mô hình**, đặc biệt trong các kịch bản test-time adaptation và domain shift so với ST4RTRACK gốc.

- **Nâng cao chất lượng tracking và tái tạo 3D trên dữ liệu thực**, góp phần tăng tính khả thi của ST4RTRACK trong các ứng dụng thực tế.

Tài liệu tham khảo

- 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
- Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, Jerome Revaud. DUSt3R: Geometric 3D Vision Made Easy. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
- Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and MingHsuan Yang. MonST3R: A simple approach for estimating geometry in the presence of motion. In International Conference on Learning Representations (ICLR), 2025.