ROBUST SELF-SUPERVISED LOSS DESIGN FOR ST4RTRACK ON REAL-WORLD VIDEOS (ST4RTrack+)

Nguyễn Phạm Phương Nam **Hồ Ngọc Luật**

University of Information Technology

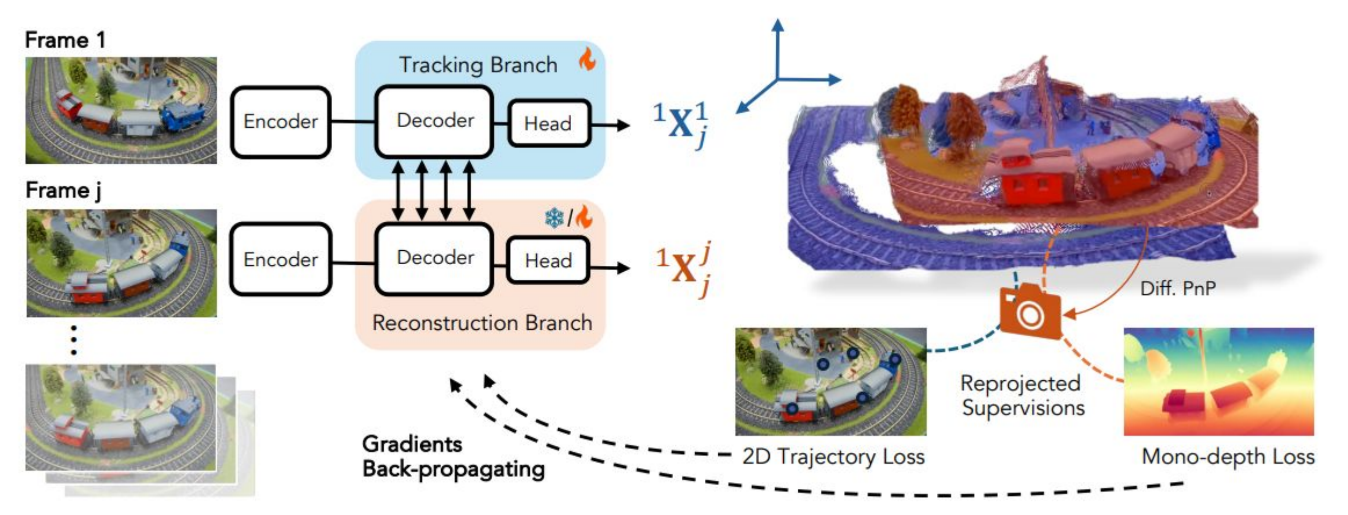
What ?

- Understand the unified 4D pointmap representation (tracking vs reconstruction) and world-frame chaining.
- Reproduce ST4RTrack: ViT encoder + Siamese decoder + 2 pointmap heads; differentiable PnP for camera pose.
- Propose & evaluate ST4RTrack+: depth-based visibility mask + per-point uncertainty weighting; ablations on real videos.

Why ?

- Real-world videos are noisy: occlusion, fast motion, monocular scale/focal errors, and imperfect pseudo-labels (e.g., CoTracker).
- Baseline reprojection losses often penalize all points equally → unstable adaptation (TTA/DA), drift, and ghost/outlier points.
- Robust, geometry- and uncertainty-aware loss design is essential for reliable deployment on real data.

Overview



St4RTrack architecture

St4RTrack+ (robust TTA/DA)
Suppress occlusion & noisy pseudo-labels.
Use visibility mask m□ + weight w□ in losses.

- Occlusion / visibility mask

$$m_n = \mathbf{1}[z_{\text{proj}} \leq z_{\text{surf}}(\hat{x}) + \varepsilon]$$

Apply to L_{traj} , L_{depth}, L_{align} .

- Uncertainty weighting

$$w_n = \exp(-\sigma_n) \quad \text{or} \quad w_n = \frac{1}{\sigma_n^2 + \varepsilon}$$

Weighted loss: $.\frac{1}{\sum_n m_n w_n} \sum_n m_n w_n \, \rho(r_n)$

Description

1) Problem with St4RTrack

St4RTrack predicts two time-dependent pointmaps in a common world frame: (i) Tracking pointmap (frame-1 content transported over time) and (ii) Reconstruction pointmap (per-frame geometry).

Adaptation uses reprojection self-supervision (2D tracks + mono-depth + 3D consistency). On real videos, occlusion/fast motion and noisy pseudo-labels introduce outliers; monocular scale errors destabilize optimization → drift.

2) Problem statement

Design robust self-supervised loss for in-the-wild videos (no 4D GT):
• mask m□ via depth consistency (z-buffer)
• weight w□ via uncertainty σ□ / confidence
• robust L_traj (NLL) + L_depth (MoGe) + L_align
• ablations on occlusion / fast motion

3) Proposed method

3.1) Pretrain
Init from DUSt3R/MonST3R; train on 4D synthetic.
Supervise tracking (mesh vertices) + reconstruction (depth + camera).

3.2) Adapt to real-world videos (TTA/DA)
Estimate K from frame-1 pointmap; solve poses $(\mathbf{R}^j, \mathbf{T}^j)$ via (diff.) PnP+RANSAC.
Self-sup losses:
•L_{traj} : robust NLL reprojection vs CoTracker (mask+σ)
•L_{depth} : mono-depth consistency (MoGe) with scale α*
•L_{align} : 3D self-consistency across branches
Total:

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

4) Expected results

• Robust self-supervised loss for ST4RTrack on real videos (mask + uncertainty).
• More stable adaptation: less drift under occlusion / fast motion.
• Better 3D tracking & reconstruction quality on real data (quant + qual).

Key equations (our robust loss & geometry constraints)
(1) Reprojection: $\hat{x}^{j,n} = \pi\big(K\left(R^j X_j^{1,n} + T^j\right)\big)$
(2) Residual: $r_n = \widehat{x}^{j,n} - x_{\text{trk}}^{j,n}$
(3) Visibility mask: $m_n = 1[z_{\text{proj}} \leq z_{\text{surf}}(\widehat{x}) + \varepsilon]$
(4) Weight: $w_n = \exp(-\sigma_n)$ or $w_n = \frac{1}{\sigma_n^2 + \varepsilon}$
(5) Robust traj NLL:

$$L_{\text{traj}} = \frac{1}{\sum_n m_n w_n} \sum_n m_n w_n \left(\frac{\|r_n\|^2}{2\sigma_n^2} + \log \sigma_n \right)$$

(6) Total:

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

NII **Nguyễn Phạm Phương Nam – Hồ Ngọc Luật -** University of Information Technology
TEL : 0389008002 - 0917121928 **Email : 23520978@gm.uit.edu.vn - 23520900@gm.uit.edu.vn**