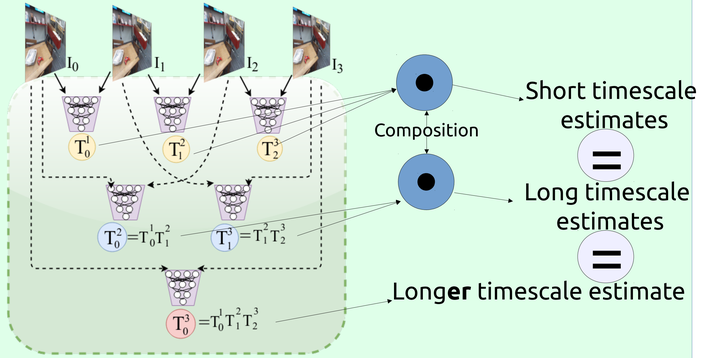
Geometric Consistency for Self-Supervised End-to-End Visual Odometry

Ganesh Iyer[1*], **J. Krishna Murthy**[2*], Gunshi Gupta[1], K. Madhava Krishna[1], Liam Paull[2]
[1]International Institute of Information Technology, Hyderabad, India
[2]MILA, DIRO, Universite de Montreal, Canada

CVPR 2018 SALT LAKE CITY • JUNE 18-22

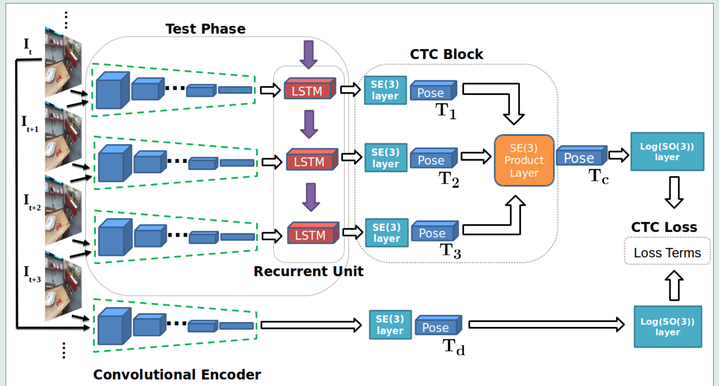
Core Contribution

- Most deep architectures for visual odometry estimation rely on large amounts of precisely labeled data.
- Such labels are extremely expensive to obtain.
- We propose an *unsupervised paradigm* for deep visual odometry learning.
- Using a *noisy teacher*, which could be a standard VO pipeline and a ***geometric consistency*** loss term, we can train accurate models for visual odometry without requiring any ground-truth labels.



We leverage the observation that compounded sequences of transformations over short timescales should be equivalent to a single transformation independently computed over longer timescales. This allows us to create *Composite Transformation Constraints (CTCs) that can be used as supervisory signals for learning visual odometry.*

Network Architecture



$$T_t^{t+1} \cdot T_{t+1}^{t+2} \cdot T_{t+2}^{t+3} = T_t^{t+3}$$
$$T_t^{t+1} \cdot T_{t+1}^{t+2} = T_t^{t+2}$$
$$T_{t+1}^{t+2} \cdot T_{t+2}^{t+3} = T_{t+1}^{t+3}$$
$$\mathcal{L}_{ctc} = \|\xi_d - \xi_c\|_2^2$$
$$\mathcal{L}_{reg} = \|\xi_* - \hat{\xi}_*\|_2^2$$
$$\mathcal{L}_{final} = \alpha\mathcal{L}_{ctc} + \beta\mathcal{L}_{reg}$$
$$\xi = (v^T, \omega^T)^T \in \mathfrak{se}(3)$$

Loss terms used for training CTCNet

End-to-end architecture: An example of Composite Transformation Constraints (CTCs) being applied to 4 successive input images. During training, two estimates are generated from the inputs: one for a sequential pairwise constraint and one for a CTC constraint.
At test time, each frame is only fed into the network once to receive the output pose from the SE(3) layer.

Project Page

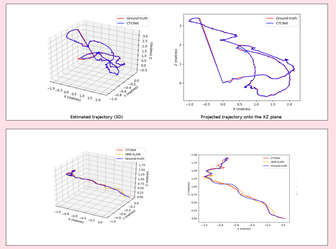
Code, models, and more ...

giyer@andrew.cmu.edu
krrish94@gmail.com
gunshigupta9@gmail.com

CTCNet: Brief Description

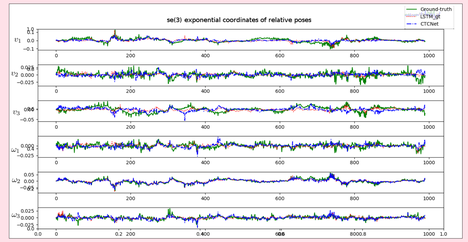
- We use *noisy* odometry estimates from a conventional VO system, such as ORB-SLAM to bootstrap our system.
- CTCs are gradually fed to the network in a curriculum that presents shorter timescale constraints to the network first, and gradually increases the window size.
- The convolutional encoder and the recurrent CTC blocks are trained sequentially.
- We use dropout at the last linear layer, to aid in uncertainty characterization.

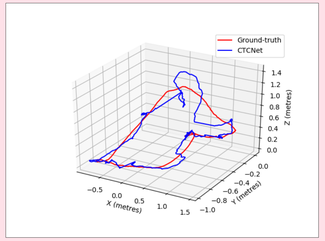
Results



Trajectory estimates on a sequence from the 7-Scenes test split. Top (left to right): Output trajectories are shown in red, against ground-truth trajectories in blue. Bottom: se(3) estimates of relative poses. Each of the 6 se(3) coordinates is plotted independently. On this sequence, CTCNet performs better than the same LSTM architecture trained using ground-truth pose labels.

se(3) error plot:A close-up view of the rotational and translational errors. We compare three quantities: ground-truth, an LSTM model trained using ground-truth labels, and CTCNet





Generalization to unseen data: CTCNet was evaluated on a sequence that was in stark contrast to the kind of sequences it had been presented with during training. Estimated 3D trajectory plotted against ground-truth.