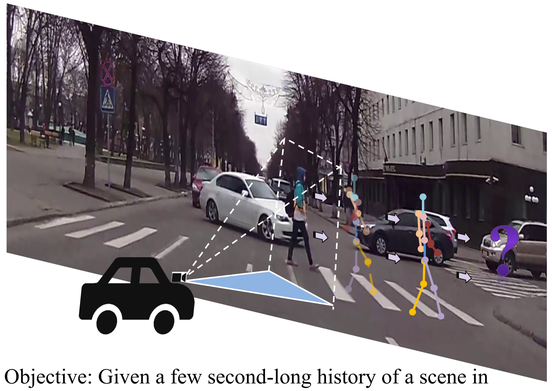
Disentangling Human Dynamics for Pedestrian Locomotion Forecasting with Noisy Supervision

Karttikeya Mangalam[1,3], Ehsan Adeli[1], Kuan-Hui Lee[2], Adrien Gaidon[2], Juan Carlos Niebles[1]

[1]Stanford University [2]Toyota Research Institute [3]University of California, Berkeley

TOYOTA RESEARCH INSTITUTE

Problem Statement & Motivation



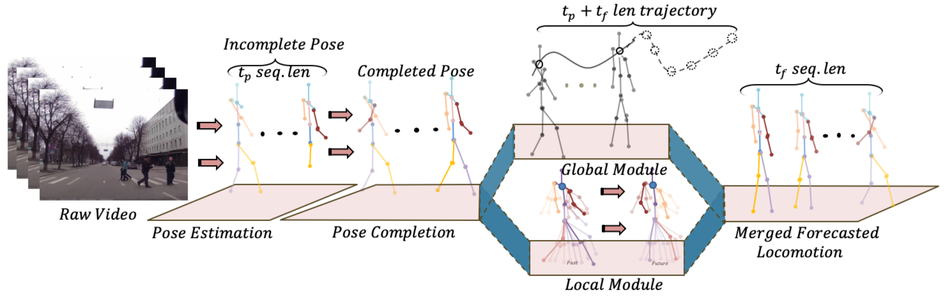
Objective: Given a few second-long history of a scene in form of an egocentric RGB video, predict the future positions of several key-points on the visible pedestrians in camera coordinates.

- Essential for reasoning about pedestrian intent & behavior
- Useful for model predictive control for autonomous cars
- Jointly targets the related tasks of pose prediction & trajectory forecasting that can benefit from mutual learning

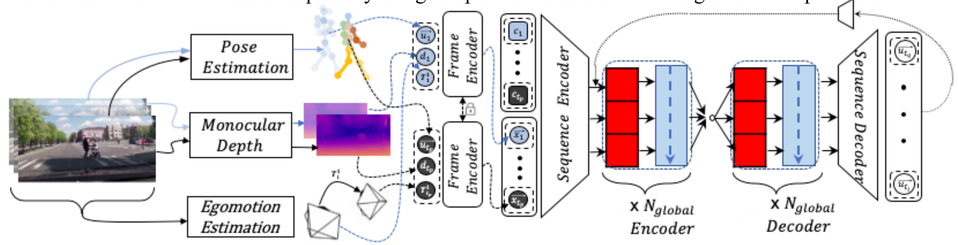
Key Ideas & Contributions

- Overall forecasting pipeline is structured as shown (Figure 1).
- We propose to complete the observed motion and decompose into global and local streams (Figure 4).
- Missing data imputation is performed using autoencoder which fills in low confidence estimates with reconstructions.
- The filled in disentangled streams are processed separately.
- We propose a novel global module utilizing extracted signals for pose, depth and Egomotion for root forecasting (Figure 2).
- The local stream is forecasted using the proposed Quasi RNN module with motion relative to the root node (Figure 3).

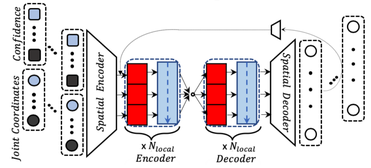
Method (Contd.)



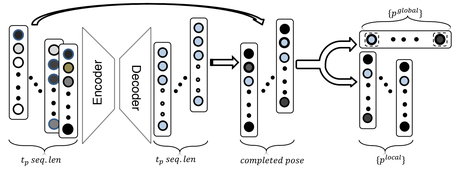
(Fig 1) **Proposed pipeline:** The extracted raw input pose is noisy with several joints missing. The pose completion module fills in the joints' positions and disentangles them into global & local streams to separate concurrent motions. Each stream is forecasted separately using its specific module & then merged for final prediction.



(Fig 2) **Global Module:** An encoder-recurrent-decoder architecture. From the raw scene, noisy estimates of the human pose, monocular depth & the scene transformation matrix are extracted. They are processed with a frame & a sequence level encoder (pretrained) respectively. These representations are then forecasted with a Quasi-RNN and decoded to estimate future positions.



(Fig 3) **Local Module:** A Quasi-RNN based local motion forecasting architecture. The prediction is performed in the latent space induced by the pose completion module. Teacher forcing is used to stabilize training.



(Fig 4) **Pose completion & Disentanglement Module:** The shades represent the confidence in locating the joint (white being missing data). All low confidence detections are replaced with the autoencoder estimates. Imputed data is then split into local and global streams.

Results

Method	Decomposition	KDE
Last Observed-Velocity	-	195.5
Constant-Velocity	-	48.9
Zero-Velocity	-	40.4
LSTM-ED	✗	24.4
TCNN[1]	✗	31.8
GRU-ED	✗	23.8
Structural RNN[2]	✗	31.7
Ours	✓	**10.9**

[1]Lea, Colin et al. CVPR 2018 — Best performing baseline — [2]Jain, Ashesh et al. CVPR 2016

Performance measured by Keypoint Disp. Error. Lower is better.

Stream	Completion	Disentanglement	KDE
Global	✗	-	6.8
Global	✓	-	5.6
Local	✗	-	9.5
Local	✓	-	6.3
Both	✗	✗	22.1
Both	✓	✗	18.2
Both	✗	✓	15.4
Both	✓	✓	**10.9**

Completion affects local stream the most

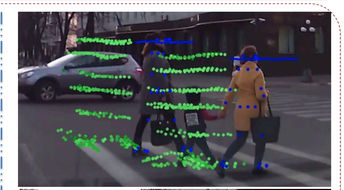
Pose completion is quite important for disentanglement!

Ablation study of various components. As indicated by the deltas, both completion and disentanglement plays a key role for forecasting.

t_f	5	10	15	20	25	30	
GRU-ED	5.7	6.6	9.7	11.4	14.1	18.3	1.5
Our Method	4.3	5.1	5.6	7.8	10.4	11.4	6.9

Gap increases with time

KDE comparison of our method with GRU-ED across time horizons.



Some qualitative results on JAAD dataset for overall locomotion. Green represents the (filled in) pose history, blue represents the predicted motion & red is the ground truth.

Qualitative results for trajectory prediction for several instances in a video. Same color scheme as above.

Discussion & Conclusion

- We propose the task of human locomotion prediction, combining the individual tasks of human trajectory forecasting and pose prediction
- We show that disentanglement combined with pose completion is an effective strategy for reducing problem complexity giving superior performance
- We also posit a novel global stream prediction module utilizing several low and mid level vision signals such as Egomotion, pose and depth.

Corresponding Author: mangalam@berkeley.edu

Paper & Video Results