SFU SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

NCFRN NSERC Canadian Field Robotics Network
RCCRT Réseau canadien CRSNG pour la robotique de terrain

AUTONOMY LAB
autonomylab.org

DeepIntent: Learning to Model Pedestrian Intent in Autonomous Driving Scenarios

Pratik Gujjar and Richard Vaughan
Autonomy Lab, Simon Fraser University
pgujjar@sfu.ca

Motivation

- Drivers and pedestrians engage in non-verbal and social cues.
- Modeling these scenarios is key for increasingly anthropomorphic interactions.
- High variability of interaction archetypes and dearth of labelled sequences makes this a hard problem.
- Predicting adverse actions 400ms (12 frames) in the future supplements a car travelling at 60kmph with 6.67m of stopping distance.

Spatiotemporal Model

- 3D convolutions derive spatiotemporal features across 10 frames of input.
- Relatively large convolutional kernel on the encoder learns reconstructable spatiotemporal features.
- Strided convolutions; no max pooling.
- Adam optimizer with learning rate scheduling.
- Asymmetric network architecture to model motion. Symmetric networks learn regularity across frames.

Analysis

- Predictions are coherent and plausible.
- Model predictions show accurate and fair colours. (Fig 3a)
- Generalizes to changing weather conditions. (Figs 3b and 3c)
- Acceptable distinction in pedestrian shapes and accuracy in motion predictions. (Fig 3d)
- Network incorporates motion-origin changes towards the final decoder stages. (Fig 2)
- Insufficient accuracy in predictions for fast movement in input frames while testing. (Fig 4a)

Future Work

- Address variance in speed and environments.
- Learn scan patterns for scenes with visual attention models.
- Semi-supervised classification of pedestrian actions/behaviours.

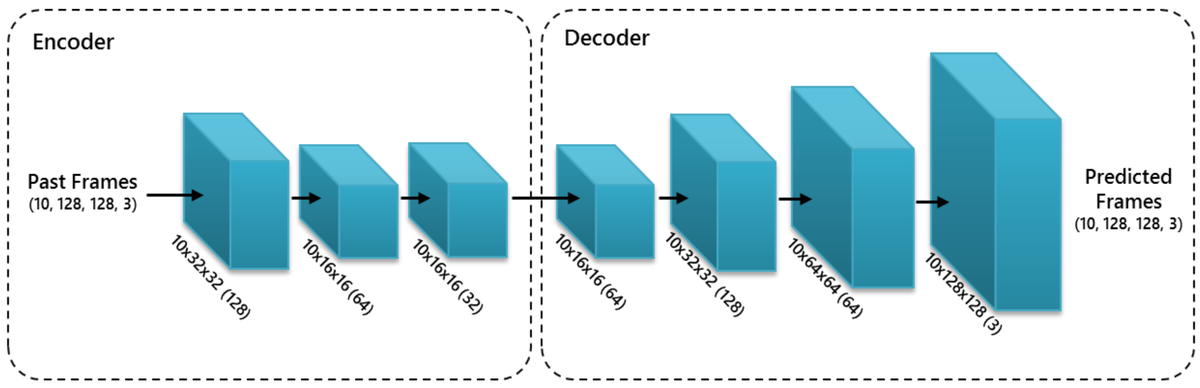

Fig 1. Our spatiotemporal model

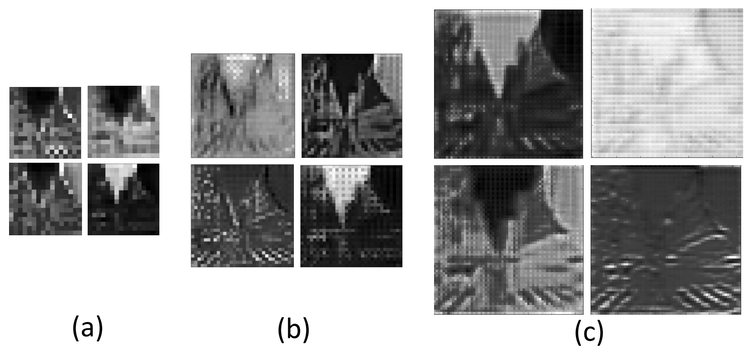

Fig 2. Visualizing learned convolutional filters in the decoder. (a) At the 10x16x16 stage (b) At the 10x32x32 stage and (c) At the 10x64x64 stage

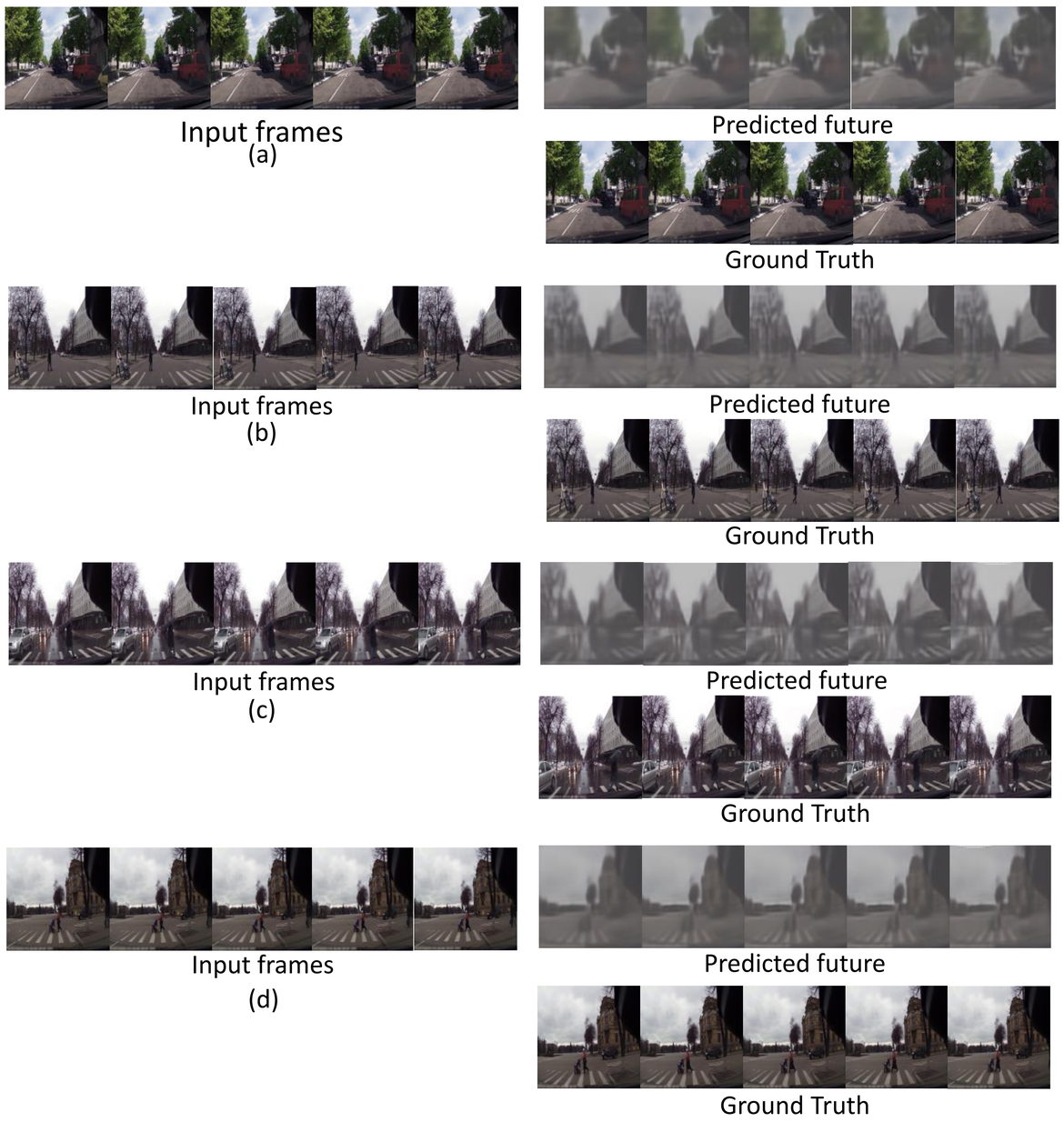

Fig 3. Example predictions for 10 frames of input from the JAAD Dataset. Intermediate frames removed for brevity.

Fig 4. More examples from the KITTI Dataset. Network is not trained on these samples. (a) Coherent but inaccurate predictions. (b) Consistent pedestrian motion predictions.

[1] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, “Agreeing to cross: How drivers and pedestrians communicate,” CoRR, vol. abs/1702.03555, 2017.
[2] “Light Vehicle Brake Systems,” Transport Canada,” TECHNICAL STANDARDS DOCUMENT No 135, Revision 2R, 2005.
[3] C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. In *Neural Information Processing Systems (NIPS)*. 2016.
[4] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning Spatiotemporal Features with 3D Convolutional Networks. arXiv:1412.0767, 2014.