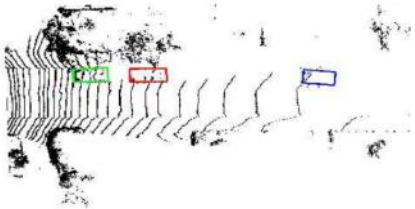
Deep Prediction for Uber Self-Driving Cars

Advisor: Prof. Jeff Schneider

Abhay Gupta (abhayg) Nitin Singh (nitinsin)

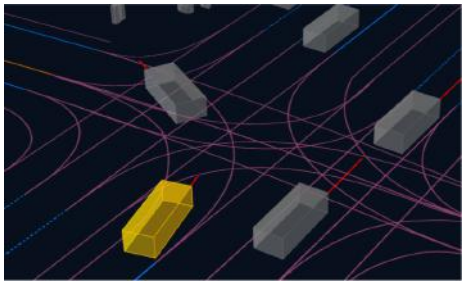
Motivation

Predicting behavior of traffic actors (vehicles/pedestrians/bicyclists) to prevent accidents and aid in better planning for Self-Driving Vehicles (SDVs)



Problem

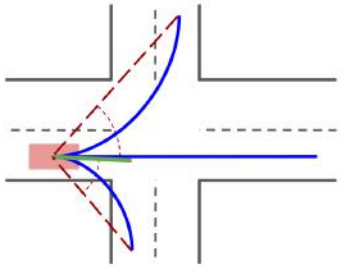
Simultaneously predict all possible trajectories of traffic actors given HD Maps of the surroundings of a SDV



Solution

1. Traditional Methods:
    a. Constant Velocity Model
    b. Unscented/Extended Kalman Filter
2. Deep Learning Methods:
    a. Intermediate Representations
    b. Model interactions of traffic actors
    c. Model non-linear structure of motion

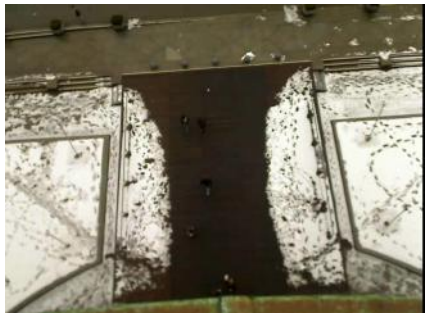
???



Past

Crowd Scenarios

Pedestrian Datasets



ETH

HOTEL

ZARA

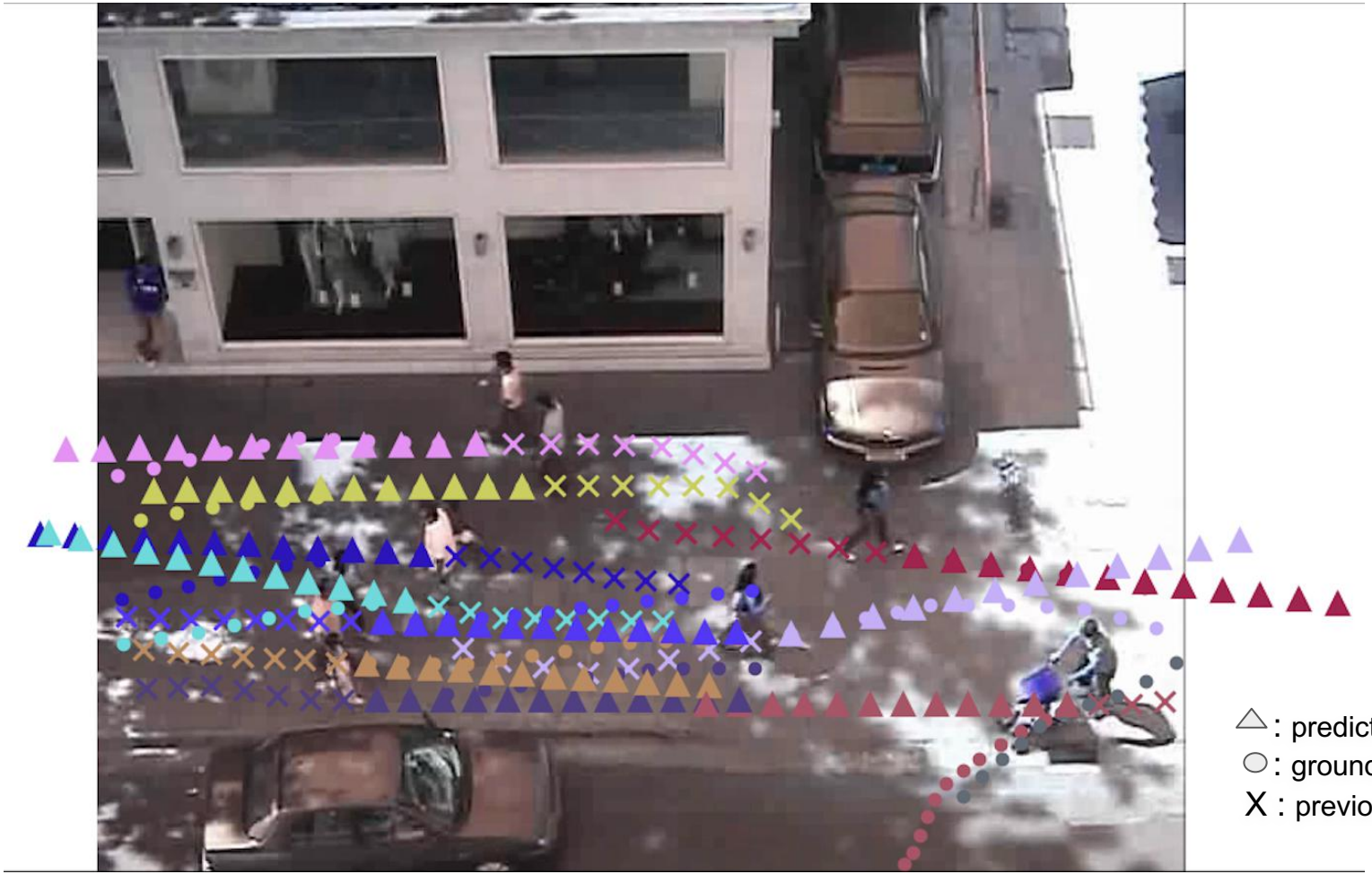
UNIVERSITY

S. Pellegrini, A. Ess, and L. Van Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. In Computer Vision–ECCV 2010, pages 452–465.Springer, 2010
L. Leal-Taix´e, M. Fenzi, A. Kuznetsova, B. Rosenhahn, and S. Savarese. Learning an image-based motion context for multiple people tracking. InCVPR, pages 3542–3549. IEEE,2014

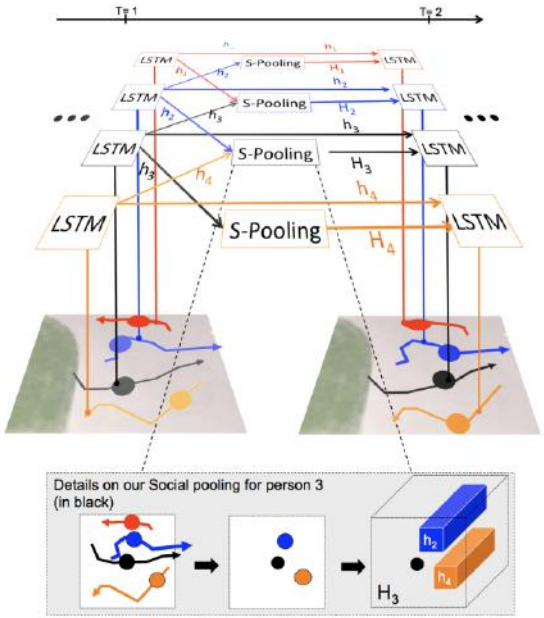
Model 1 - Constant Velocity Model (CVM)

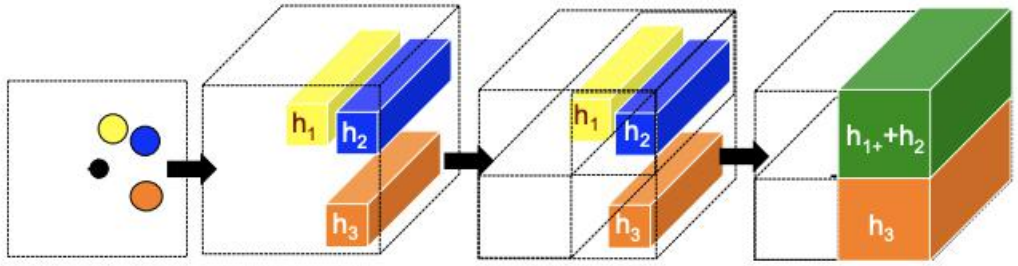
1. Assumes pedestrians walk with same velocity and in the same direction as their previous two timesteps.
2. We compute the velocity vector and propagate it for the future timesteps.

Schöller, Christoph, et al. "The Simpler the Better: Constant Velocity for Pedestrian Motion Prediction." *arXiv preprint arXiv:1903.07933* (2019).


△: predicted
○: ground truth
X : previous time-steps

Model 2 - Social LSTM[1]





$$L^i = -\sum_{t=T_{obs}+1}^{T_{pred}} \log\left(\mathbb{P}\left(x_t^i, y_t^i | \sigma_t^i, \mu_t^i, \rho_t^i\right)\right)$$

i – pedestrian index

$(x_t^i, y_t^i) - \mathcal{N}(\mu_t^i, \sigma_t^i, \rho_t^i)$

T_{obs} – history observed till here

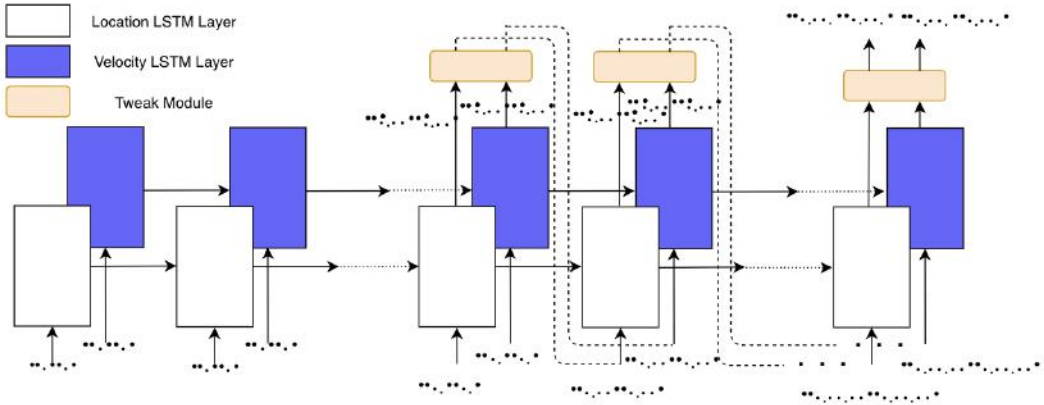
T_{pred} – predictions made till here

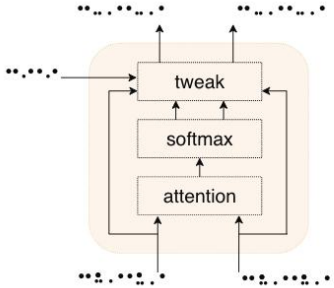
The probabilities are modeled using Mixture Density Networks[2]

1 - Alahi, Alexandre, et al. "Social lstm: Human trajectory prediction in crowded spaces." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

2 - Bishop, Christopher M. *Mixture density networks*. Technical Report NCRG/4288, Aston University, Birmingham, UK, 1994.

Model 3 - Velocity Model (LVA-LSTM)

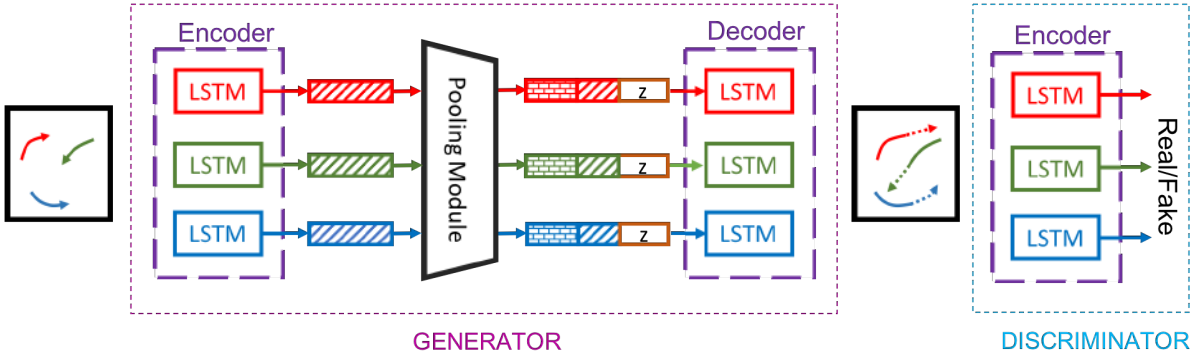




$$x_t = \alpha_t^l \hat{x}_t + \alpha_t^v \left(x_{t-1} + \hat{u}_t\right)$$
$$y_t = \alpha_t^l \hat{y}_t + \alpha_t^v \left(y_{t-1} + \hat{v}_t\right)$$
$$u_t = x_t - x_{t-1}$$
$$v_t = y_t - y_{t-1}$$
$$\alpha_t^l, \alpha_t^v - \text{output computed by softmax layer}$$

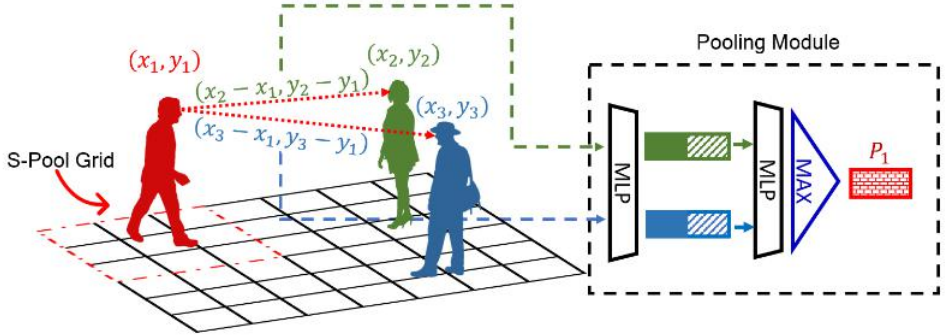
Xue, Hao, Du Huynh, and Mark Reynolds. "Location-Velocity Attention for Pedestrian Trajectory Prediction." *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019.

Model 4 - Social GAN



1. Scene-scale Pooling instead of neighborhood pooling
2. GANs - emulate more natural trajectories

Gupta, Agrim, et al. "Social gan: Socially acceptable trajectories with generative adversarial networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.



1. Max-Pool - helps to learn order invariant symmetric representations similar to PointNet[1]
2. Introduction to Variety Loss

$$\mathcal{L}_{variety} = \min_k \left\| Y_i - \hat{Y}_i^{(k)} \right\|_2$$

Y_i – Ground Truth Prediction

$\hat{Y}_i^{(k)}$ – Model Prediction after sampling $z \sim \mathcal{N}(0, 1)$

k – hyper-prarameter

1 - Qi, Charles R., et al. "Pointnet: Deep learning on point sets for 3d classification and segmentation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

Performance

- Average Displacement Error (ADE) - The mean square error (MSE) over all estimated points of a trajectory and the true points
- Final Displacement Error (FDE) - The distance between the predicted final destination and the true final destination at end of the prediction period

- Error Reported in meters
- Annotations are done in 0.4 seconds each
- Predictions are done for 2 different lengths: 8/12 timesteps (3.2/4.8 secs)

Prediction Length (8 timesteps) - ADE / FDE

	CVM	Vanilla LSTM	Social LSTM	Social GAN (k=20)	LVA LSTM
BIWI ETH	0.62 / 1.37	0.70 / 1.45	0.73 / 1.48	0.57 / 1.11	0.94/2.25
BIWI Hotel	0.27 / 0.54	0.55 / 1.17	0.49 / 1.01	0.36 / 0.72	1.40/3.05
UCY Zara1	0.25 / 0.56	0.25 / 0.53	0.27 / 0.56	0.21 / 0.41	0.26/0.64
UCY Zara2	0.23 / 0.49	0.31 / 0.65	0.33 / 0.70	0.21 / 0.43	0.23/0.59
UCY University	0.27 / 0.60	0.36 / 0.77	0.41 / 0.84	0.33 / 0.70	0.36/0.91

All errors are reported in meters

Prediction Length (12 timesteps) - ADE / FDE

	CVM	Vanilla LSTM	Social LSTM	Social GAN (k=20)	LVA LSTM
BIWI ETH	0.86 / 2.38	1.09 / 2.41	1.09 / 2.35	0.70 / 1.28	1.16/2.72
BIWI Hotel	0.37 / 0.81	0.86 / 1.91	0.79 / 1.76	0.48 / 1.02	2.15/5.18
UCY Zara1	0.41 / 0.98	0.41 / 0.88	0.47 / 1.00	0.34 / 0.69	0.48/1.14
UCY Zara2	0.36 / 0.82	0.52 / 1.11	0.56 / 1.17	0.31 / 0.65	0.39/0.99
UCY University	0.46 / 1.07	0.61 / 1.31	0.67 / 1.40	0.56 / 1.18	0.68/1.59

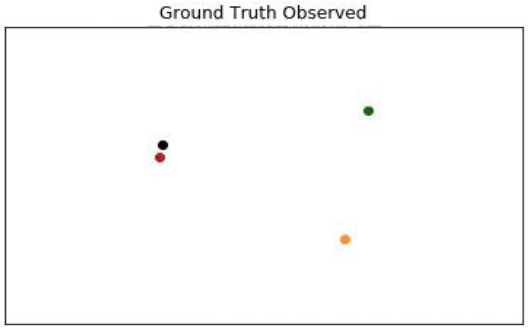
All errors are reported in meters

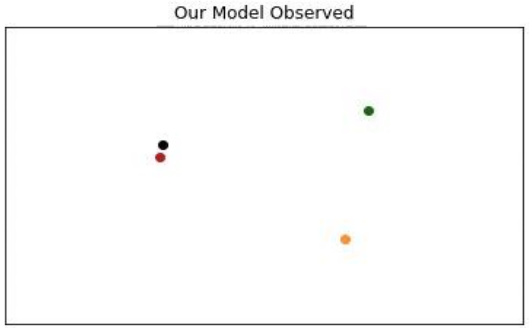
Inference Analysis

	Prediction - 8 steps (in ms)*	**Prediction - 12 steps (in ms)***
CVM	1.09*e-8	1.45*e-8
Vanilla LSTM	5.9	6.2
Social LSTM	6.3	7.1
Social GAN	7	8.5
LVA LSTM	45	57

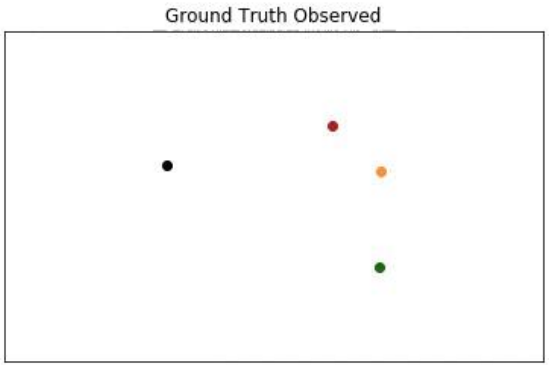
*All numbers are reported on Titan X GPU cards w/ one sample prediction

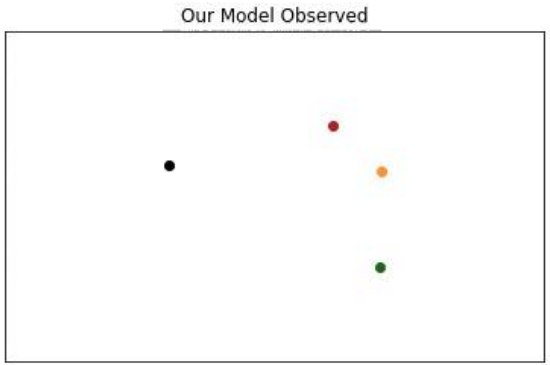
Social GAN - Some results





Social GAN - Some results





Comments from Past

1) Mapping interactions of pedestrians with other actors?
    a) This is not an easy task
    b) The resolutions at which we encode information is very different for different objects
    c) Also, the type of interactions pedestrians and cars have with each other are very different
2) Metrics reported are not good?
    a) ADE / FDE serve as decent metrics when actually using for the autonomous scenarios
    b) But yes, uncertainties can also be predicted to facilitate better learning

Autonomous Vehicle (AV) Scenarios

Datasets

KITTI[1] Dataset

ApolloScape[2] Dataset

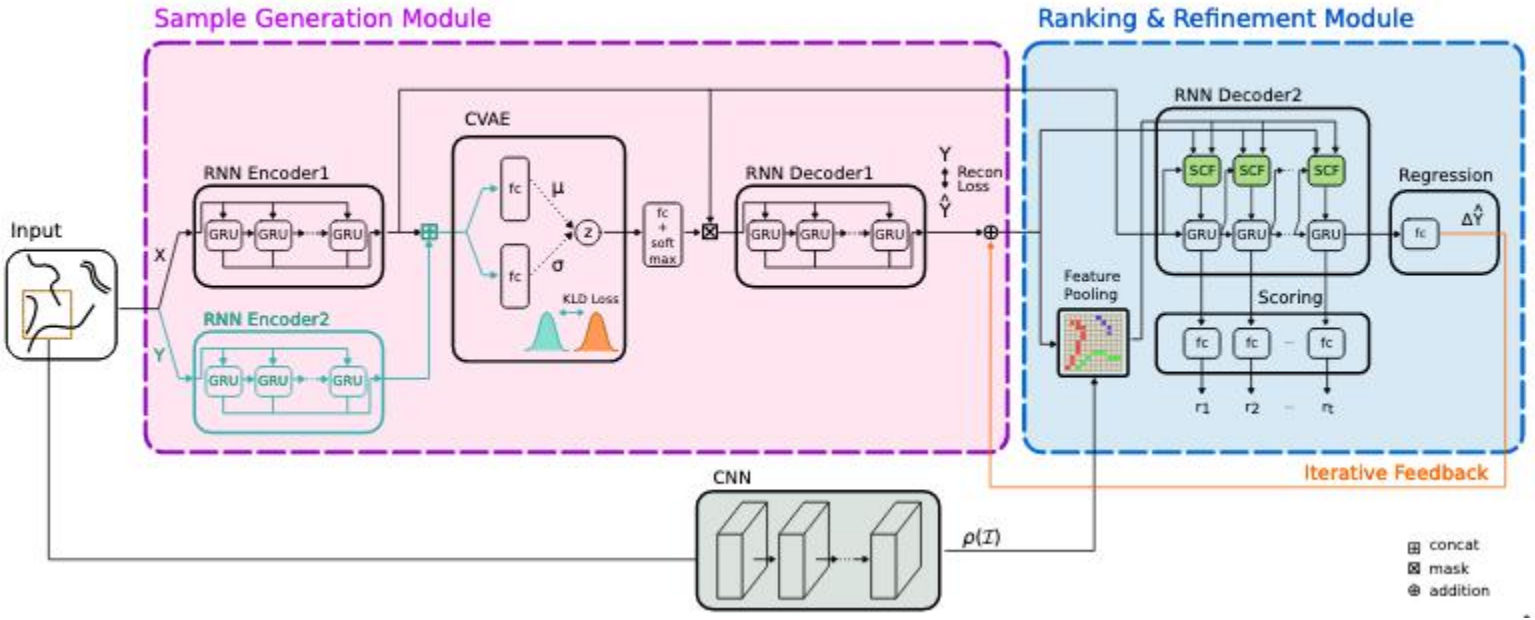
Geiger, Andreas, Philip Lenz, and Raquel Urtasun. "Are we ready for autonomous driving? the kitti vision benchmark suite." *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012.
Huang, Xinyu, et al. "The apolloscape dataset for autonomous driving." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018.

Model 1 - Constant Velocity Model (CVM)

1. Assume cars move with same velocity and in the same direction as their previous timesteps.
2. We compute the velocity vector and propagate it for the future timesteps.

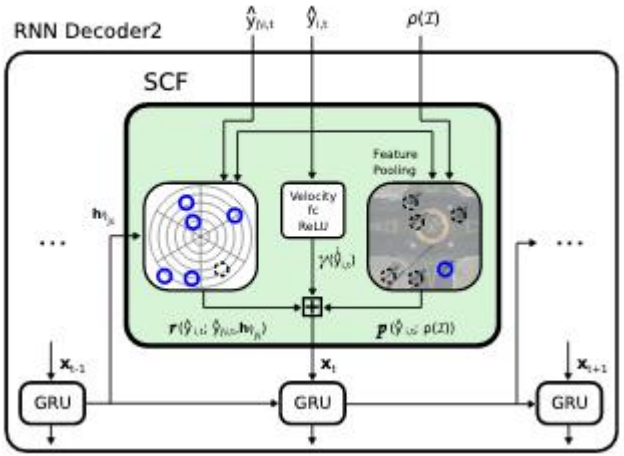
Schöller, Christoph, et al. "The Simpler the Better: Constant Velocity for Pedestrian Motion Prediction." *arXiv preprint arXiv:1903.07933* (2019).

Model 2 - DESIRE



Lee, Namhoon, et al. "Desire: Distant future prediction in dynamic scenes with interacting agents." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

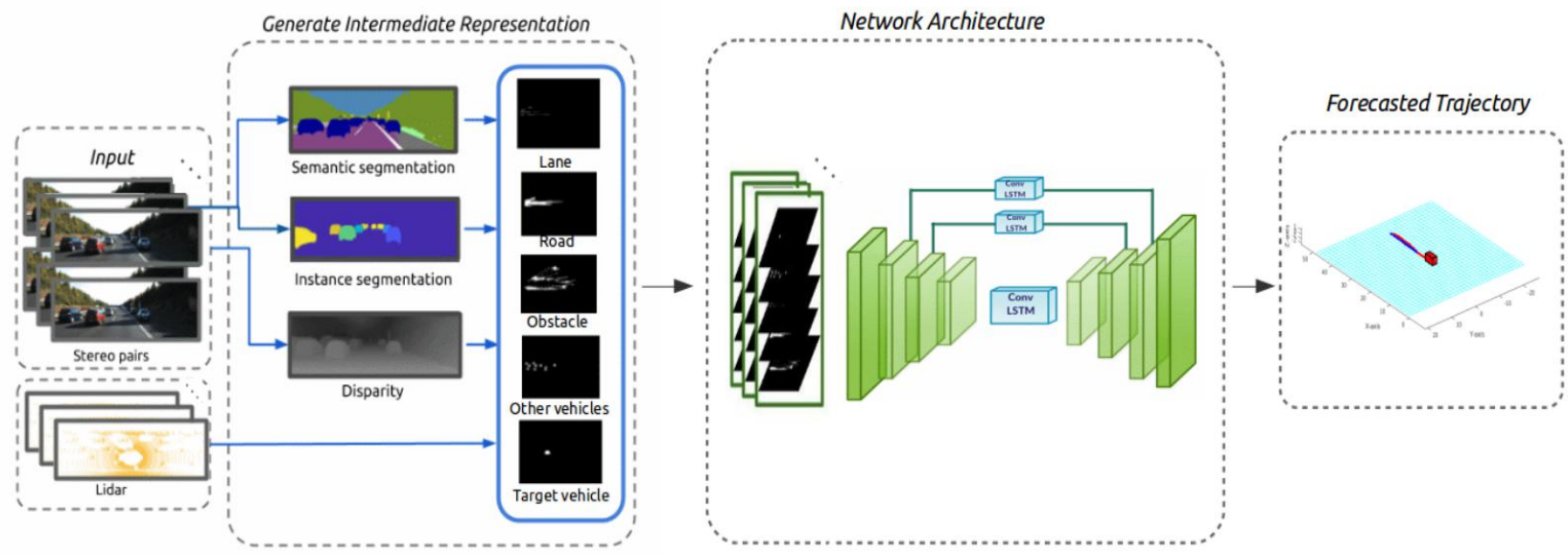
Model 2 - DESIRE



1. Module incorporates multi-contextual cues
    a. Scene information
    b. Velocity Cues
    c. Interactions among agents using multiple cues
2. To map interactions
    a. Log Polar grid is taken
    b. Average pooling done over the grid

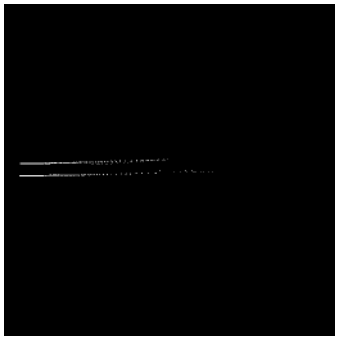
Lee, Namhoon, et al. "Desire: Distant future prediction in dynamic scenes with interacting agents." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

Model 3 - INFER



Srikanth, Shashank, Junaid Ahmed Ansari, and Sarthak Sharma. "INFER: INtermediate representations for FuturE pRediction." *arXiv preprint arXiv:1903.10641* (2019).

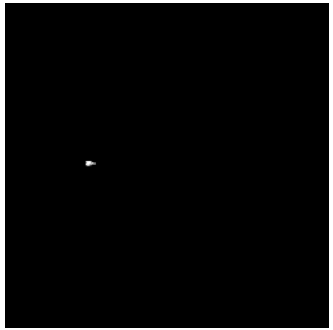
INFER Intermediate Representations



Lane

Obstacles

Road



Target Vehicle.

Other Vehicles

Performance

- Average Displacement Error (ADE) - The mean square error (MSE) over all estimated points of a trajectory and the true points

- Error Reported in meters
- Predictions are done for 4 seconds out
- History of 2 seconds fed to the model
- To match metrics across papers, errors reported at each 1s interval

Prediction Length – 4s (ADE)

	1s	2s	3s	4s
CVM	0.70	1.41	2.12	2.99
DESIRE-SI	0.31	0.70	1.39	2.12
INFER (ConvLSTM)	0.76	1.23	1.60	1.96
INFER (SkipLSTM)	0.53	0.89	1.22	1.56

All errors are reported in meters

Inference Analysis

	Training Time* (in hrs)	Prediction Time* (in ms)
CVM	N/A	0.1
DESIRE-SI	96	224
INFER (ConvLSTM)	27	153
INFER (SkipLSTM)	39	189

1. Slow inference - not really transferrable to actual cars
2. But a good baseline to start working for new directions

*All numbers reported on Titan X GPU cards w/ prediction time per car

The Good, The Bad & The Ugly

Model	Multi-Agent	Multi-Modal	Stochastic	Real-time Inference
CVM	✓	✗	✗	✓
Social-LSTM	✓	✗	✗	✓
LVA-LSTM	✓	✗	✗	✗
Social-GAN	✓	✗	✓	✓
DESIRE	✗	✓	✓	✗
INFER	✗	✓	✗	✗

Present

Current Progress - Crowd Scenarios

1. LVA-LSTM
    a. Implemented and evaluation ✓
    b. Interpretability ✓̸
2. Social GAN
    a. Interpretability ✓̸

╳ - not possible; ✓ - done; ✓̸ - ongoing

Current Progress - AV Scenarios

1. KITTI
   a. CVM implemented and evaluation ✓
   b. DESIRE implementation and evaluation ✓
   c. INFER: implementation and evaluation ✓
   d. DESIRE: interpretability ≠
   e. INFER: interpretability ≠
2. BAIDU
   a. CVM implementation ✓; evaluation ≠
   b. DESIRE / INFER ✗
      i. No correlation between forecasting data (only numerical i/p) and image data
      ii. Written to dataset team requesting for this

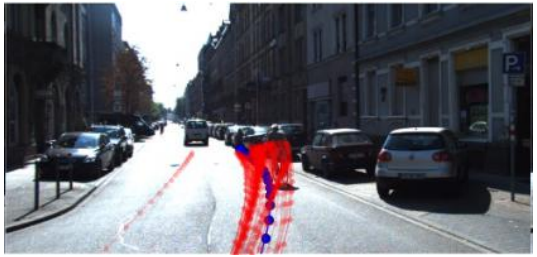
✗ - not possible; ✓ - done; ≠ - ongoing

Preliminary Analysis

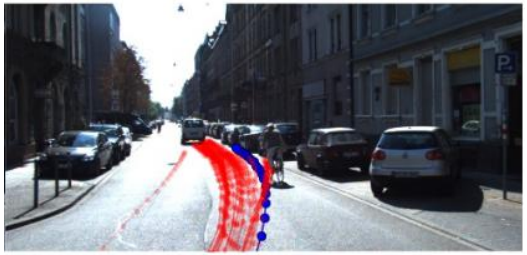
1. DESIRE
    a. Works with front-view images
    b. Ranking & Refinement step is very complex
        i. Most time consuming module in inference
        ii. Takes about 156 ms per prediction
        iii. Delta updates to the model cause delays in inference
2. INFER
    a. Auto-regressive model
    b. Works with top-down intermediate representations
    c. ConvLSTMs (64 blocks being used) are very complex
        i. Most time consuming module in inference
        ii. Take about 128 ms per prediction

Preliminary Analysis - DESIRE

Iteration 0

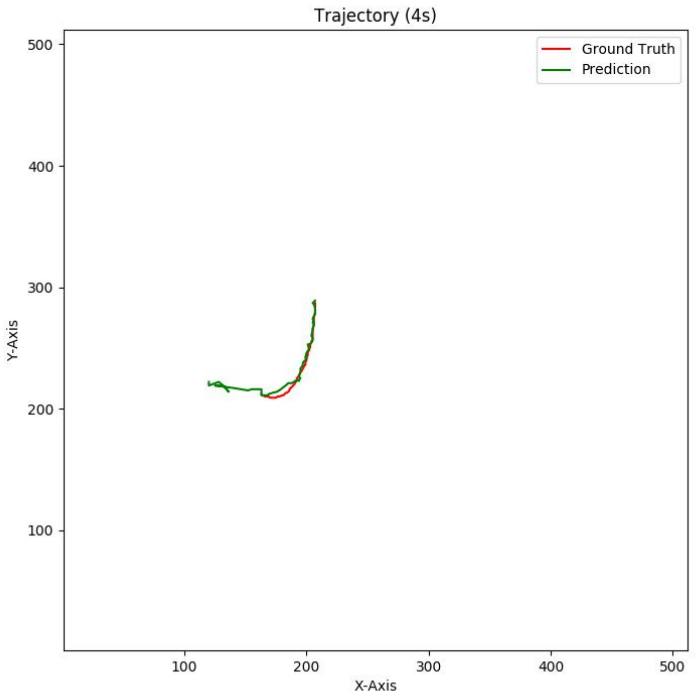


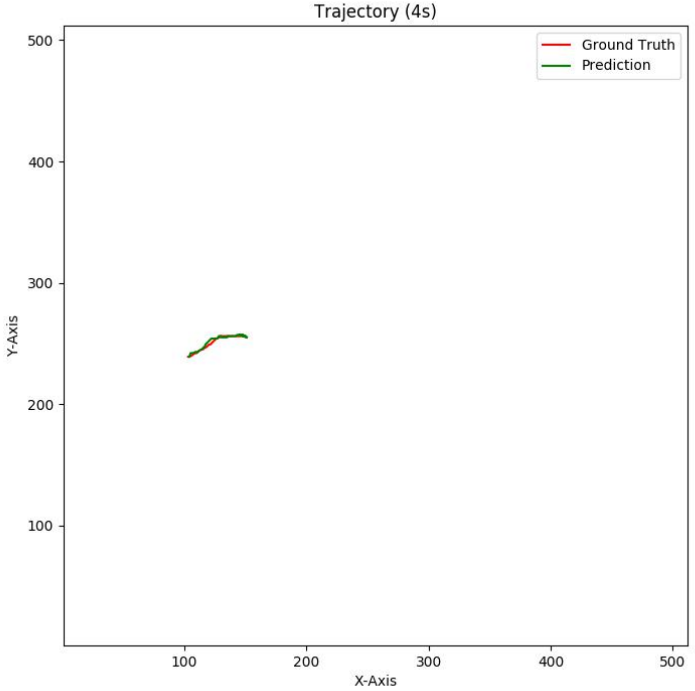
Iteration 1



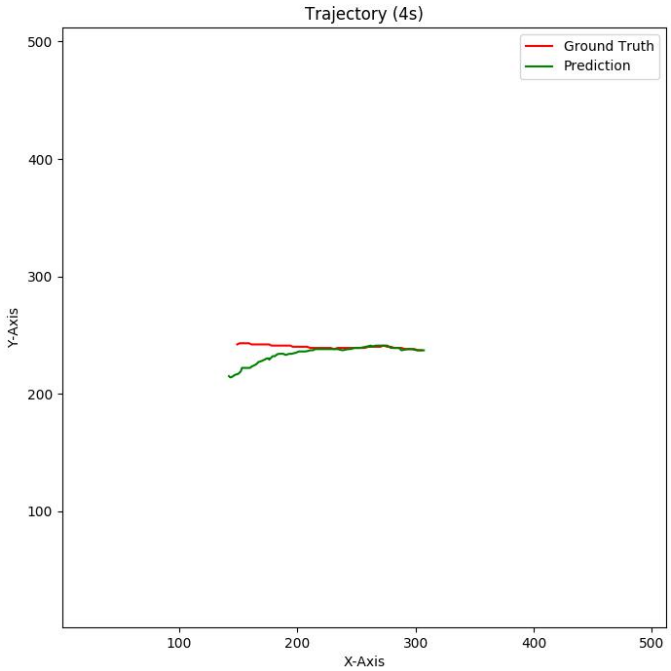
Iteration 3

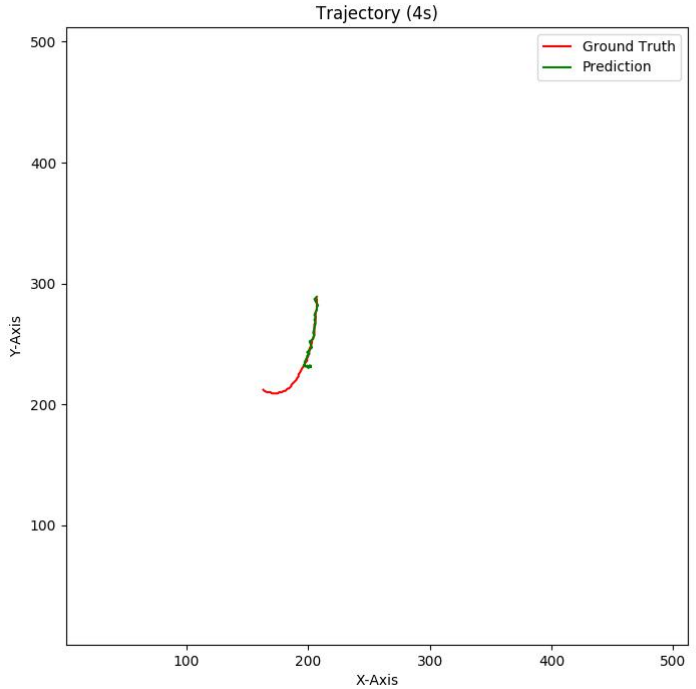
Preliminary Analysis - INFER (Success)





Preliminary Analysis - INFER (Failure)





Future

Experimenting with Loss / MultiModal Predictions

1. Currently using ADE / FDE metrics (computed using MSE Loss)
2. Add uncertainty prediction to networks similar to [1]
3. Add multi-modal prediction to networks similar to [2]
4. Add single block for multi-step prediction similar to [1]

1 - Chou, Fang-Chieh, et al. "Predicting Motion of Vulnerable Road Users using High-Definition Maps and Efficient ConvNets." (2018).
2 - Cui, Henggang, et al. "Multimodal Trajectory Predictions for Autonomous Driving using Deep Convolutional Networks." *arXiv preprint arXiv:1809.10732* (2018).

Experimenting for Crowd Scenarios

1. LVA[1] uses Vanilla LSTMs
    a. Does not perform any agent interactions
    b. GANs are shown to be better
        i. Adding them to velocity models - improve results?
2. LSTMs - hard to train; short memory retention
    a. The answer - Temporal Convolutional Networks[2]
    b. Shown to be successful in NLP
    c. Will they improve results?

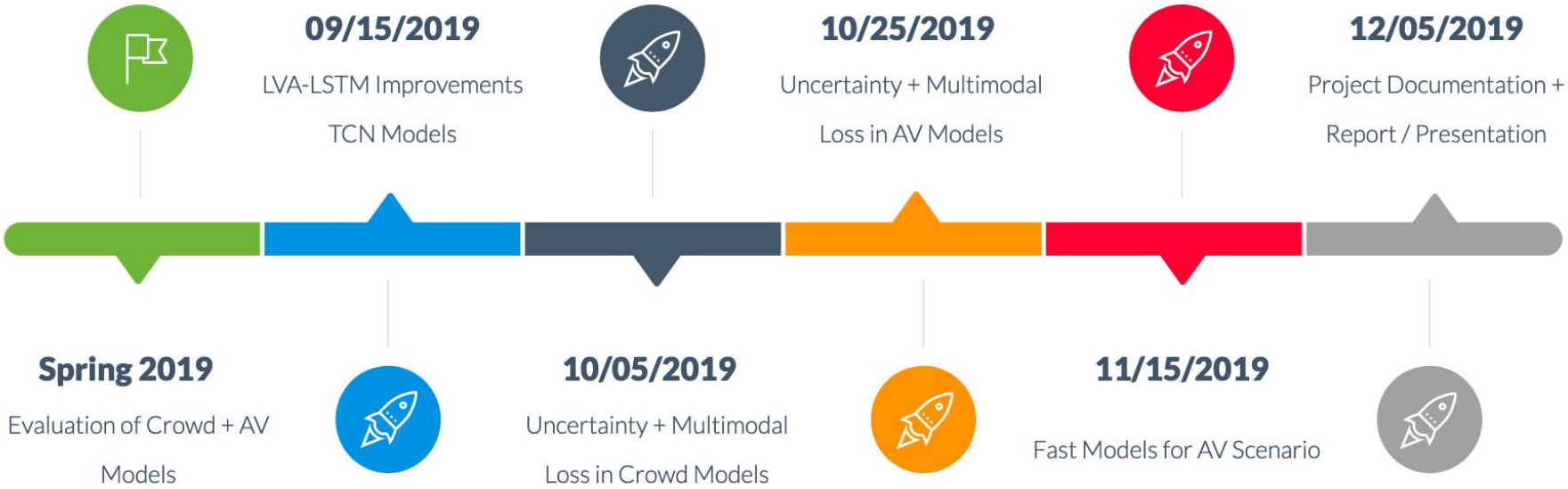
1 - Xue, Hao, Du Huynh, and Mark Reynolds. "Location-Velocity Attention for Pedestrian Trajectory Prediction." *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019.
2 - Bai, Shaojie, J. Zico Kolter, and Vladlen Koltun. "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling." *arXiv preprint arXiv:1803.01271* (2018).

Experimenting for AV Scenarios

1. Faster inference models for KITTI
    a. DESIRE / INFER are slow
    b. Use convolutional models instead for encoding essential information ?
    c. Better intermediate representations ?
2. Incorporating Oxford Robotcar[1] Dataset into our model exploration

1 - Maddern, Will, et al. "1 year, 1000 km: The Oxford RobotCar dataset." *The International Journal of Robotics Research* 36.1 (2017): 3-15.

Proposed Timeline



Acknowledgements

Author	Model
Baran Nama	Social LSTM
Agrim Gupta	Social GAN
Hao Xue	LVA LSTM
Shashank Srikanth	INFER

Gratitude to the faculty and students for their comments over the previous presentations to help concretize ideas over time and perform necessary evaluations needed for the project.

Q&A