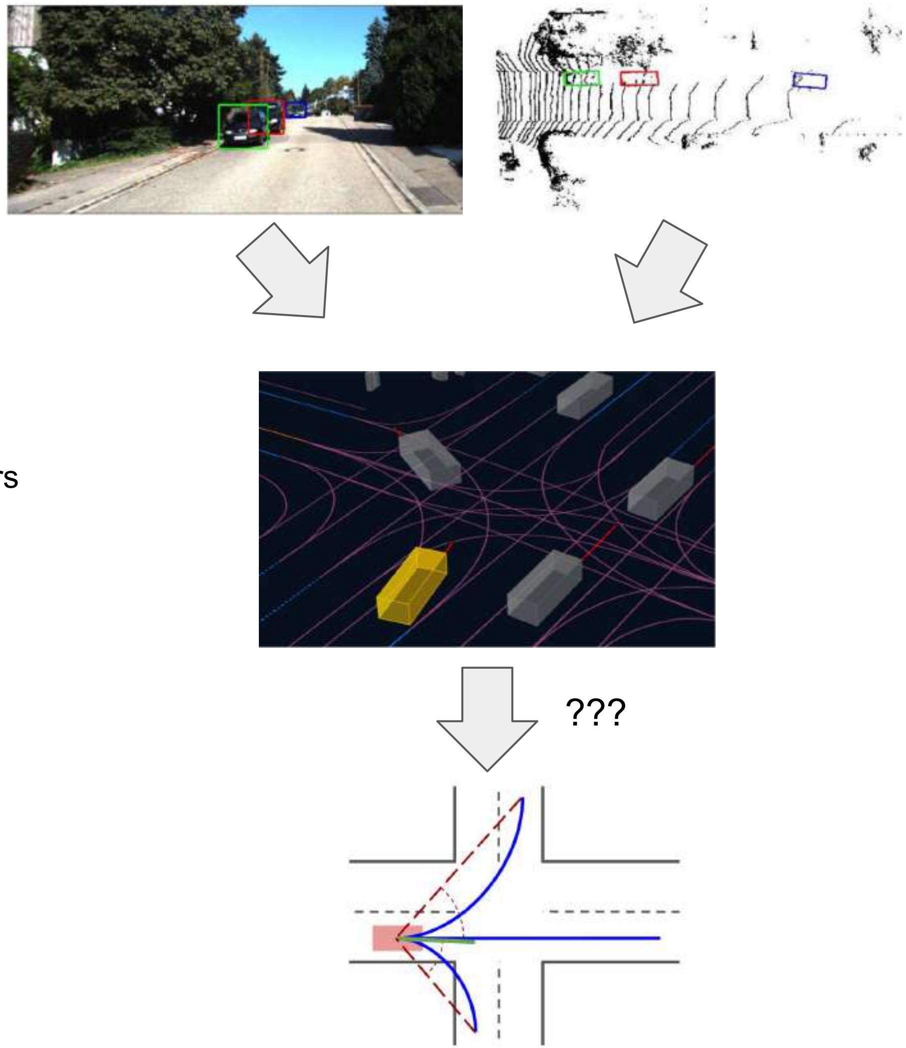
Deep Prediction for Uber Self-Driving Cars

Advisor: Prof. Jeff Schneider

Team: Abhay Gupta, Nitin Singh

Motivation

Predicting behavior of traffic actors (vehicles/pedestrians/bicyclists) to prevent accidents and aid 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

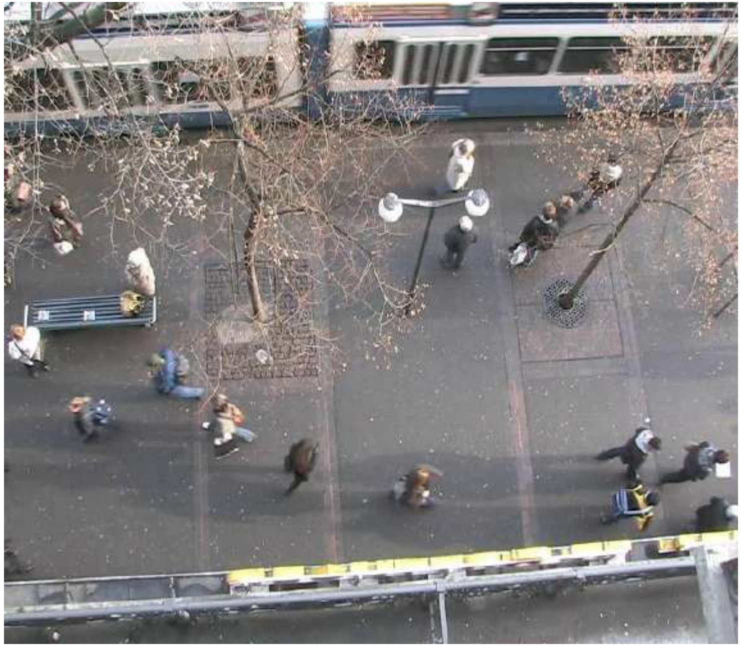
Past

Comments from Past

1) Why not handle pedestrians?
    a) This is the current focus of our research
2) Is there a combined strategy for pedestrians and cars?
    a) The spatial resolution encoded for a pedestrian vs a car is very different.
    b) It is better to use two different models than compromise on the predictions
    c) In the future, based on changes in input representation, we may come up with a strategy.
3) Why not work directly with multiple sensors which self-driving cars use?
    a) We build intermediate representations from these sensors and predict using them.
    b) It is not optimal to use raw data for the trajectory prediction problem.
    c) It makes the model heavy and real-time inference is hindered.

Dataset 1 - BIWI Pedestrian

ETH



HOTEL

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

Dataset 2 - UCY Crowd

ZARA

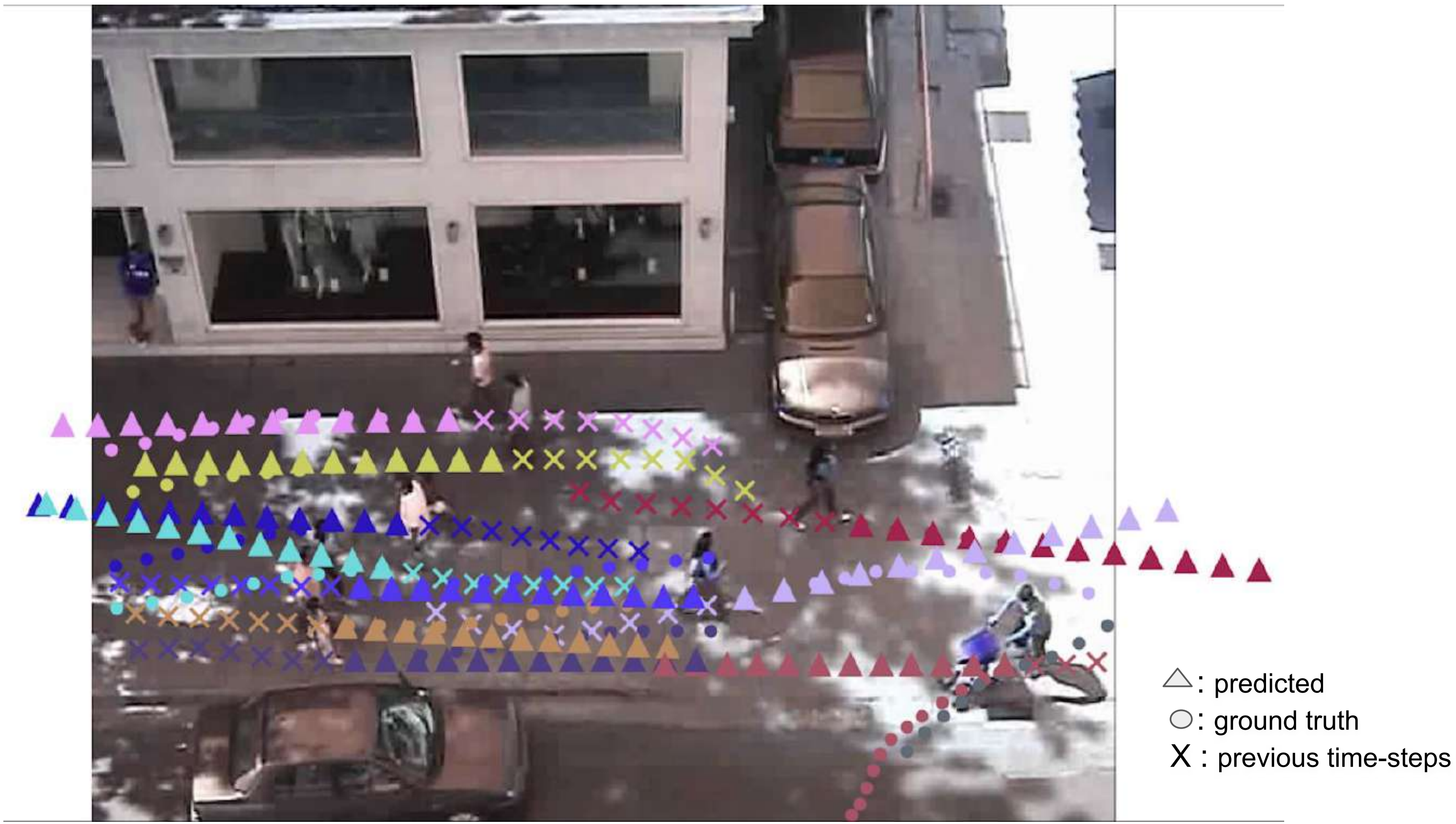
UNIVERSITY

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

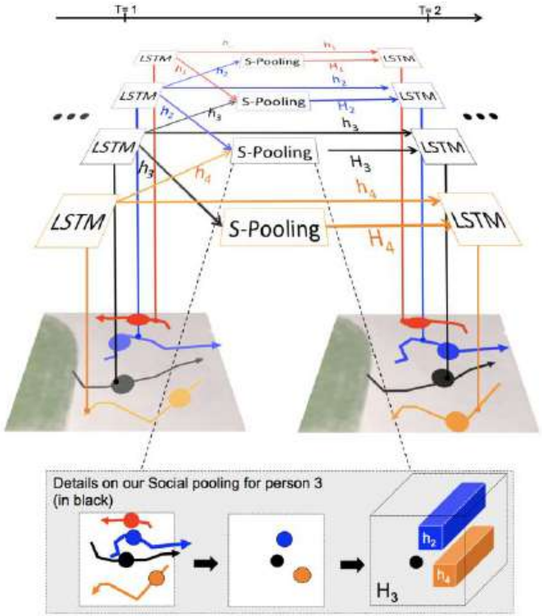
Model 1 - Constant Velocity Model (CVM)

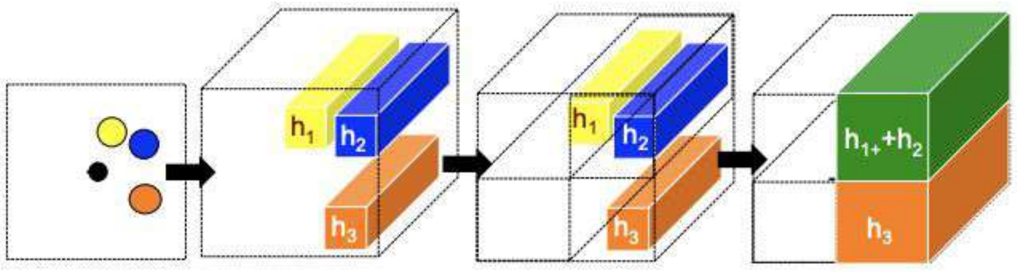
1. Assumes pedestrians walk with same velocity and in the same direction as their previous two timesteps.
2. We compute the velocity vector and propagate it for the future timesteps.

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]





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

i – pedestrian index

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

T_{obs} – history observed till here

T_{pred} – predictions made till here

The probabilities are modeled using Mixture Density Networks[2]

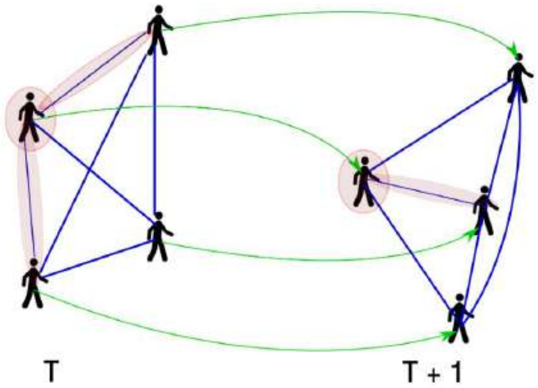
1 - Alahi, Alexandre, et al. "Social lstm: Human trajectory prediction in crowded spaces." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

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

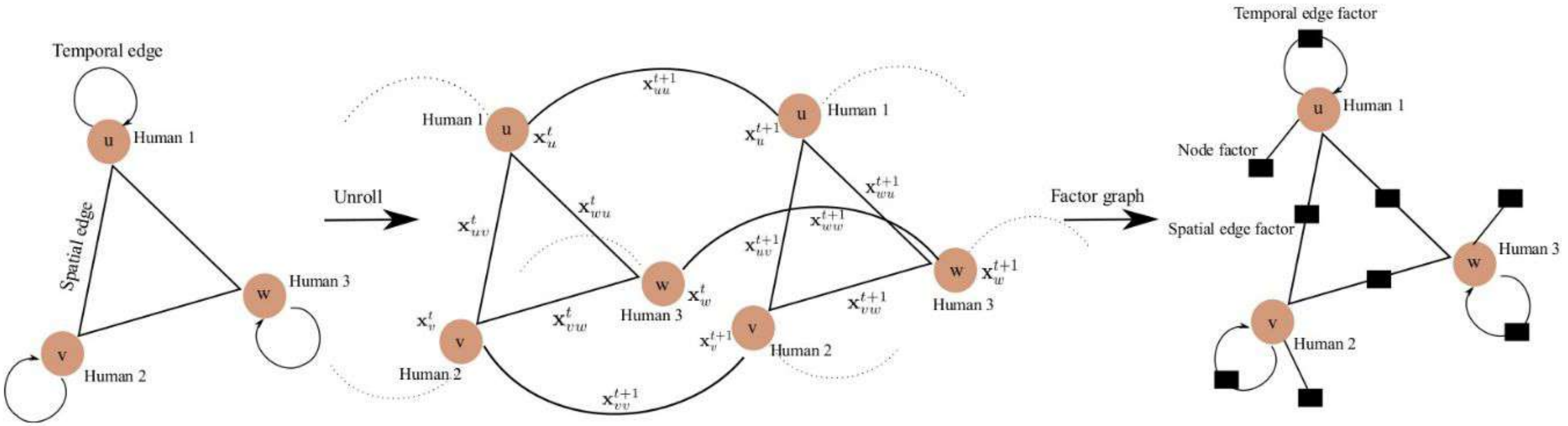
Present

Model 1 - Social Attention

A spatio-temporal graph representation for pedestrian motion.



Vemula, Anirudh, Katharina Muelling, and Jean Oh. "Social attention: Modeling attention in human crowds." *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018.



The factor graph representation of the spatio-temporal graph is trained using Structural-RNN[1]

x_u^t = Node value at time t

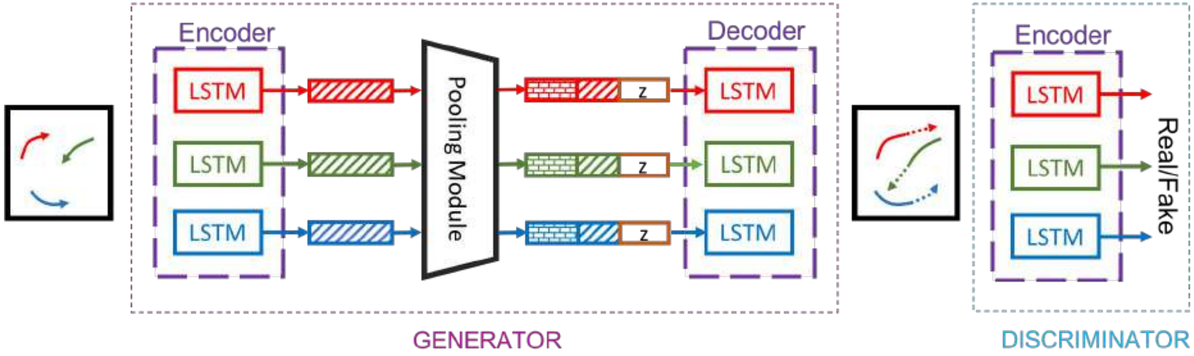
x_u^{t+1} = Node value at time t+1

x_{uv} = Spatial Edge

x_{uu} = Temporal Edge

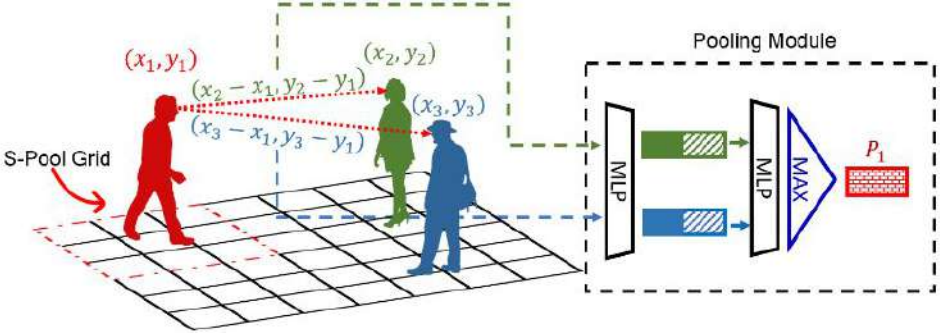
2- Jain, Ashesh, et al. "Structural-RNN: Deep learning on spatio-temporal graphs." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

Model 2 - 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.

Current Progress

1. Social GAN - implementation and evaluation - done
2. Visualizing latent space manifolds for Social GAN - ongoing.
3. Social Attention - implementation done; evaluation ongoing.
4. Developing visualization module for all results.

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)**
BIWI ETH	0.62 / 1.37	0.70 / 1.45	0.73 / 1.48	0.57 / 1.11
BIWI Hotel	0.27 / 0.54	0.55 / 1.17	0.49 / 1.01	0.36 / 0.72
UCY Zara1	0.25 / 0.56	0.25 / 0.53	0.27 / 0.56	0.21 / 0.41
UCY Zara2	0.23 / 0.49	0.31 / 0.65	0.33 / 0.70	0.21 / 0.43
UCY University	0.27 / 0.60	0.36 / 0.77	0.41 / 0.84	0.33 / 0.70

All errors are reported in meters

Prediction Length (12 timesteps) - ADE / FDE

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

All errors are reported in meters

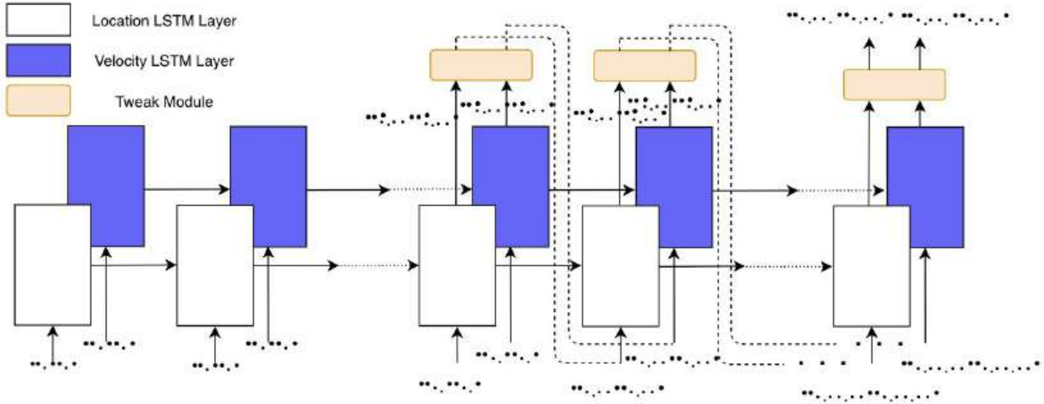
Future

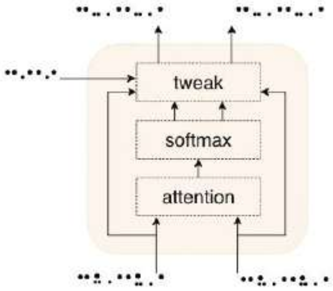
Future Work 1 - Incorporating Scene Images & Velocity of Pedestrians

1. Use Social-GAN[1] base network
2. Modify to implement Velocity information to the network
    a. Adds non-contextual cues; especially when pedestrians speed-up across frames[2]
3. Modify to incorporate Images
    a. This will provide scene-level information to the model

1 - 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.
2 - 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.

Velocity Model (LVA-LSTM)

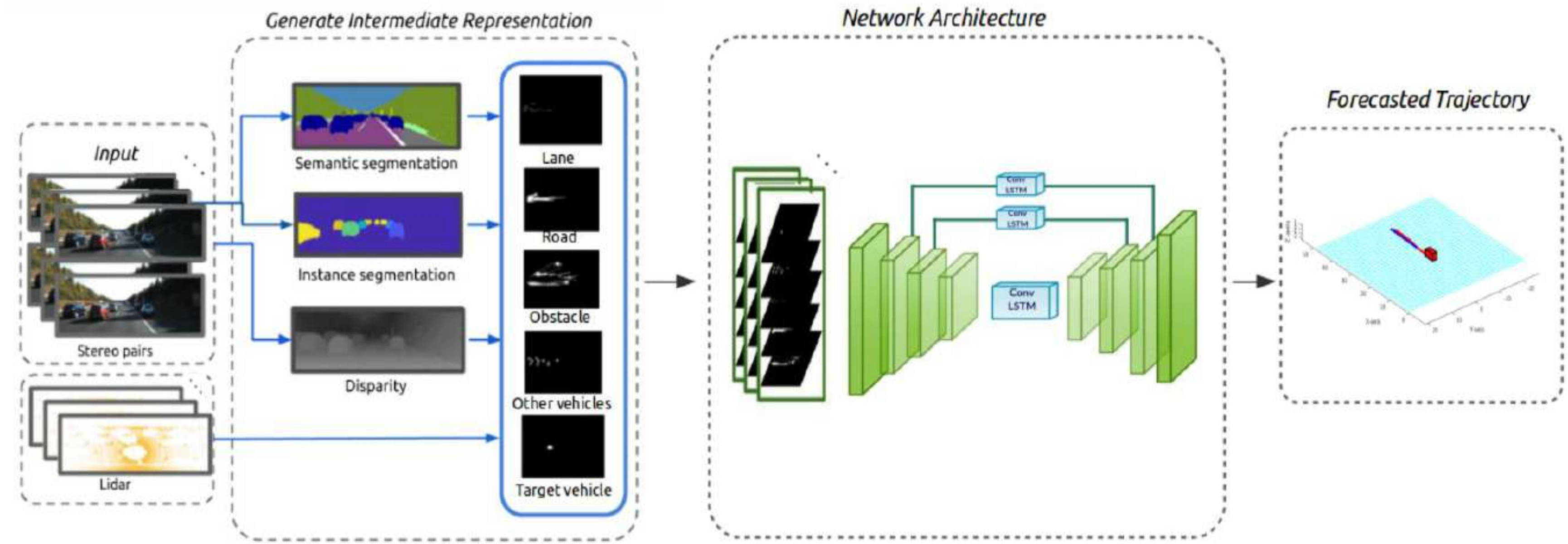




$$x_t = \alpha_t^l \hat{x}_t + \alpha_t^v \left(x_{t-1} + \hat{u}_t\right)$$
$$y_t = \alpha_t^l \hat{y}_t + \alpha_t^v \left(y_{t-1} + \hat{v}_t\right)$$
$$u_t = x_t - x_{t-1}$$
$$v_t = y_t - y_{t-1}$$
$$\alpha_t^l, \alpha_t^v - \text{output computed by softmax layer}$$

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.

Future Work 2 - Autonomous Driving



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

Proposed Timeline

Q&A