Peeking into the Future: Predicting Future Person Activities and Locations in Videos

Junwei Liang, Lu Jiang, Juan Carlos Niebles, Alexander Hauptmann, Li Fei-Fei

Seminar Computer Vision

10th of July 2019

Naya Baslan

Supervisor: Osama Makansi

Outline

- Motivation
- Related Work (for trajectory and activity prediction)
- Approach: Next Model
  - Person Behavior Module
  - Person Interaction Module
  - Trajectory Generator with Focal Attention
  - Activity Prediction
- Experiments: ActEV/VIRAT and ETH & UCY
- Conclusion

Motivation

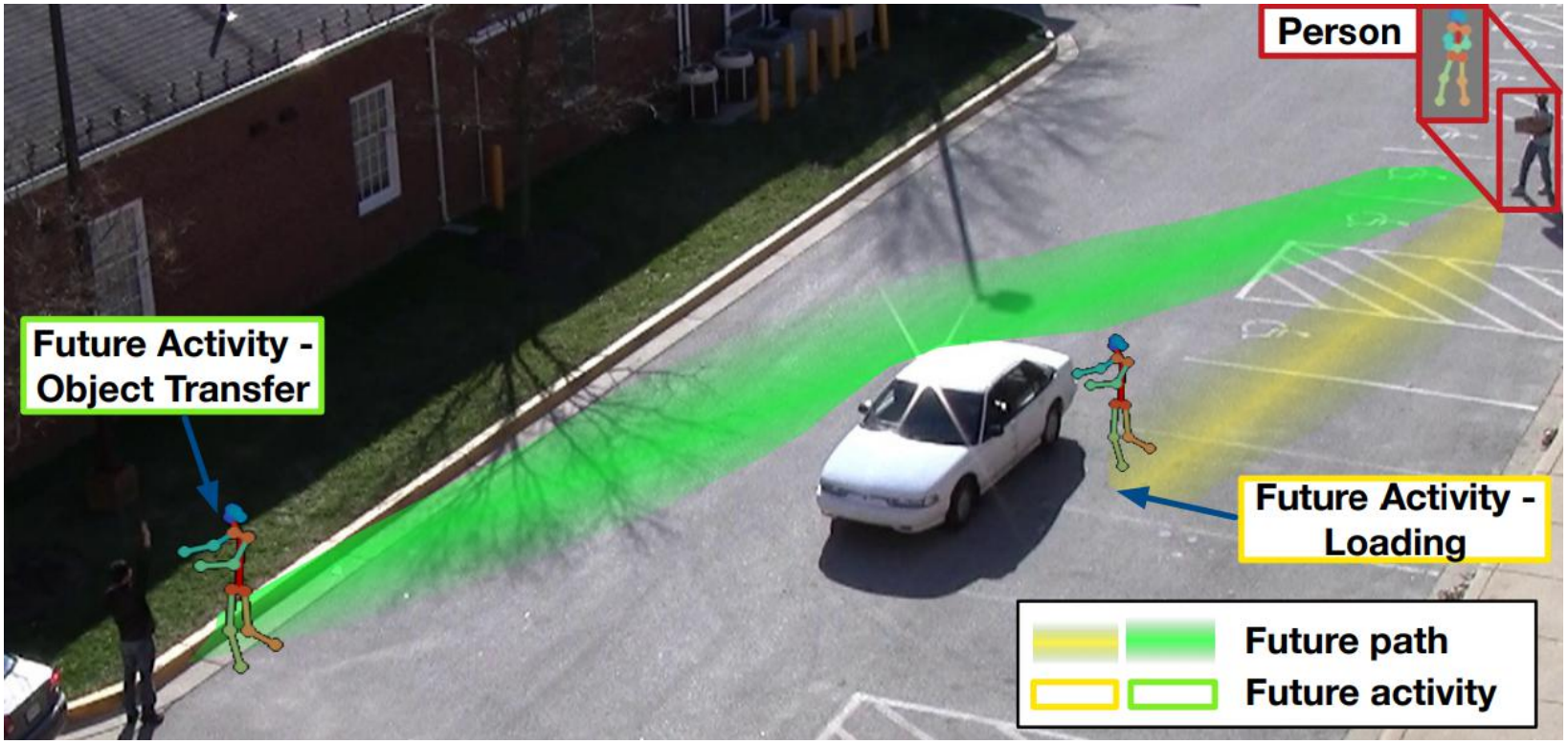


Figure 1: Introduction

Main Contribution of this Paper:

Path Prediction | Activity Prediction

joint task

Results in better prediction of future path and activity

Outline

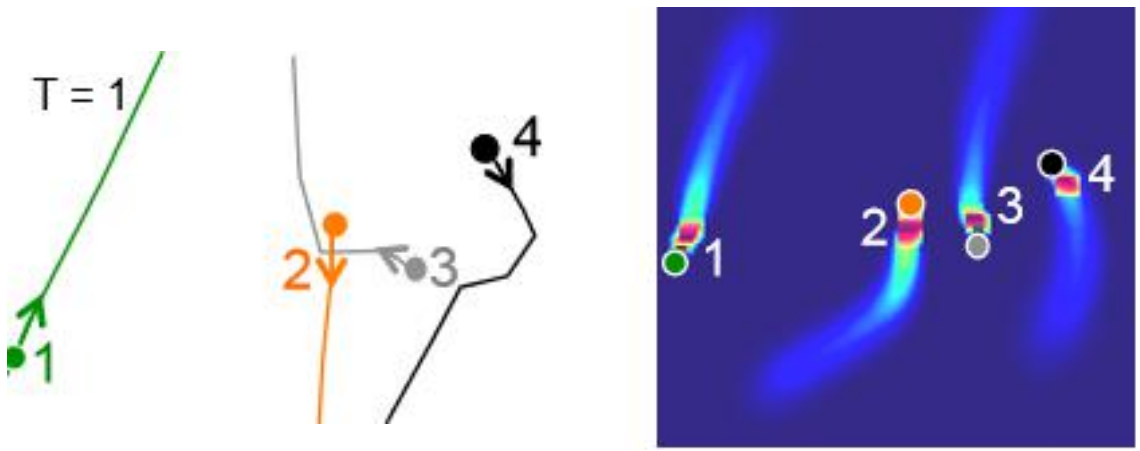
- Introduction
- Related Work (for trajectory and activity prediction)
- Approach: Next Model
  - Person Behavior Module
  - Person Interaction Module
  - Trajectory Generator with Focal Attention
  - Activity Prediction
- Experiments: ActEV/VIRAT and ETH & UCY
- Conclusion

Related Work for Trajectory Prediction

Person-person models:

Human Interactions and Behaviour → Trajectory Prediction

examples:
Social-LSTM and Social-GAN



Source: Social LSTM Human Trajectory

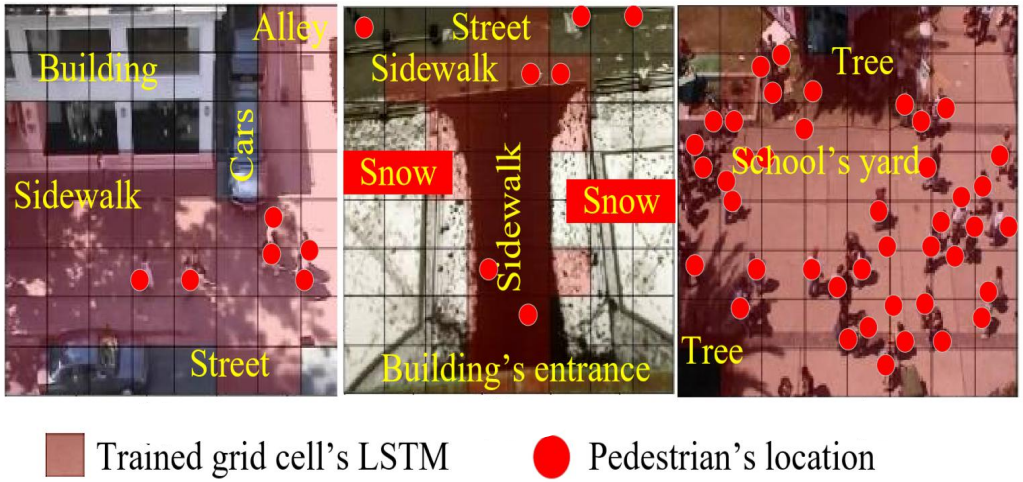
- Allows sharing of hidden states between different LSTMs

Related Work for Trajectory Prediction

Person-scene models:

Surrounding → Trajectory

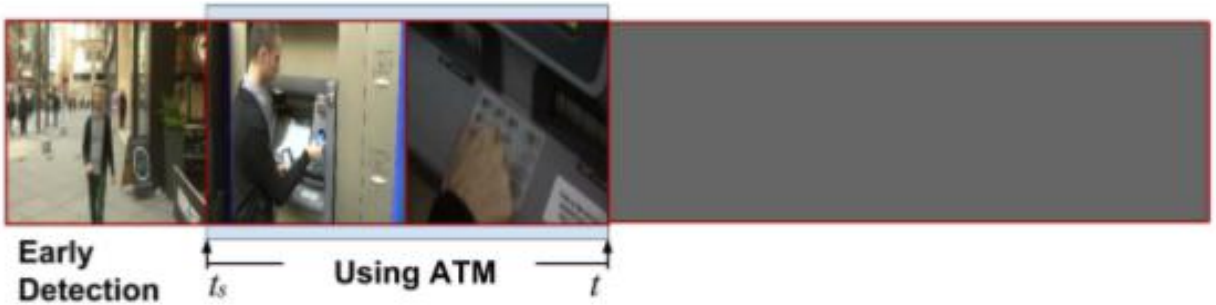
examples:
Scene-LSTM, Car-Net and SoPhie



Source: Scene-LSTM: A Model for Human Trajectory Prediction

- Models the interaction between the person and the surrounding

Related Work for Activity Prediction:



Source: Learning Activity Progression in LSTMs for Activity Detection and Early Detection

- Using LSTMs to recognize human activities
- In this paper: both activity and trajectories are jointly predicted using rich visual features and focal attention

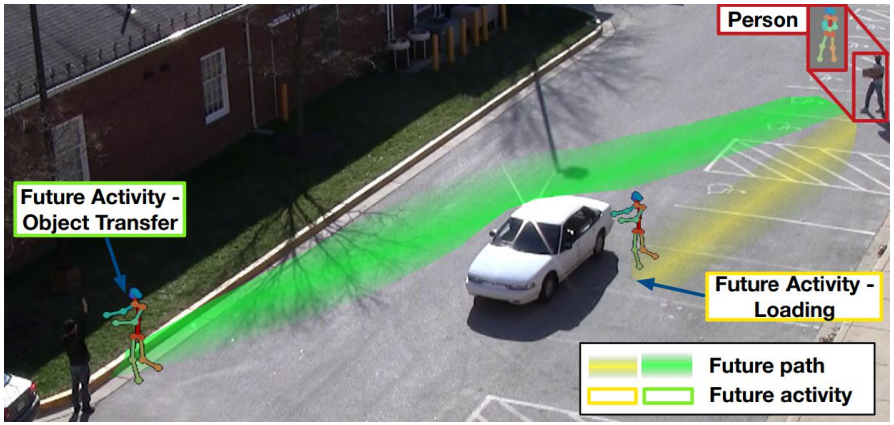
Outline

- Introduction
- Related Work (for trajectory and activity prediction)
- Approach: Next Model
  - Person Behavior Module
  - Person Interaction Module
  - Trajectory Generator with Focal Attention
  - Activity Prediction
- Experiments: ActEV/VIRAT and ETH & UCY
- Conclusion

Approach and Problem Formulation

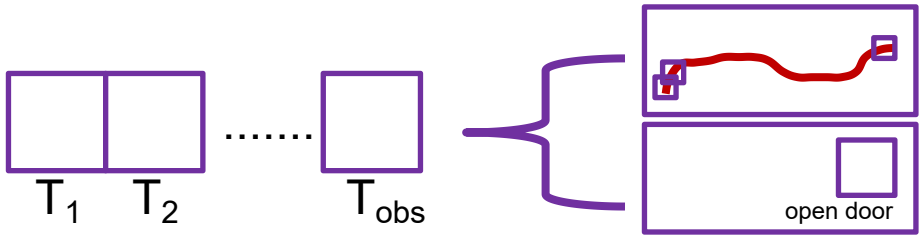
Approach:

- Next: Multi-task learning model used to predict path and activity



Problem Formulation:

1 to T_{obs}	T_{obs+1} to T_{pred}	T_{pred}
make observations	predict positions of person	predict activity



trajectory predition from T_{obs+1} to T_{pred}

activity predition at T_{pred}

Next Model: Framework

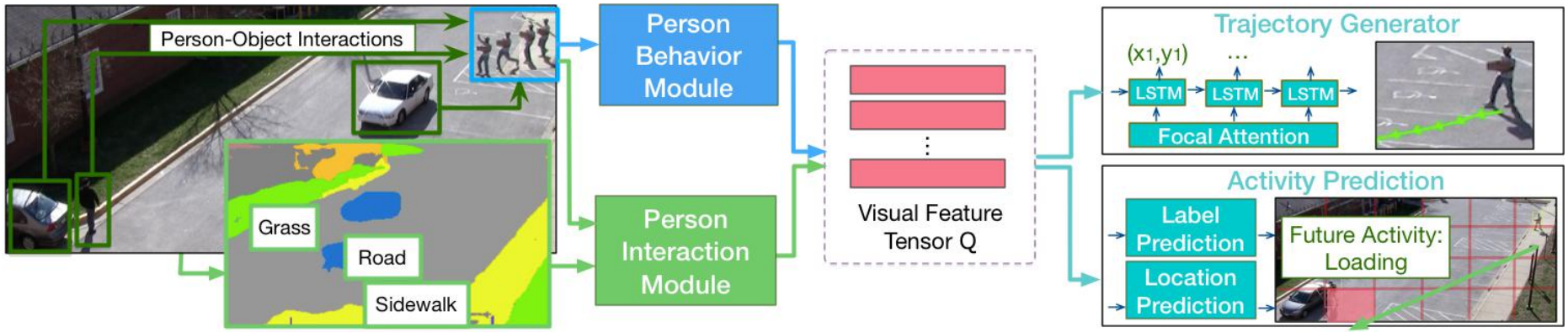
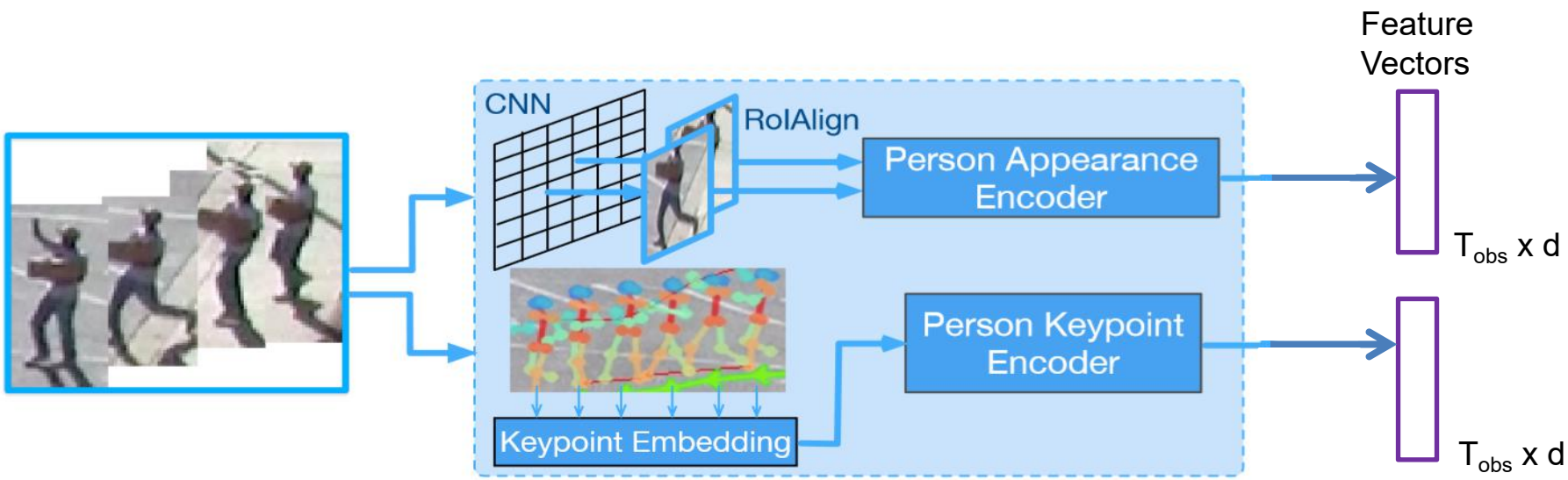


Figure 2: Model Overview

- Person Behavior Module
- Person Interaction Module
- Trajectory Generator
- Activity Prediction

Person Behavior Module



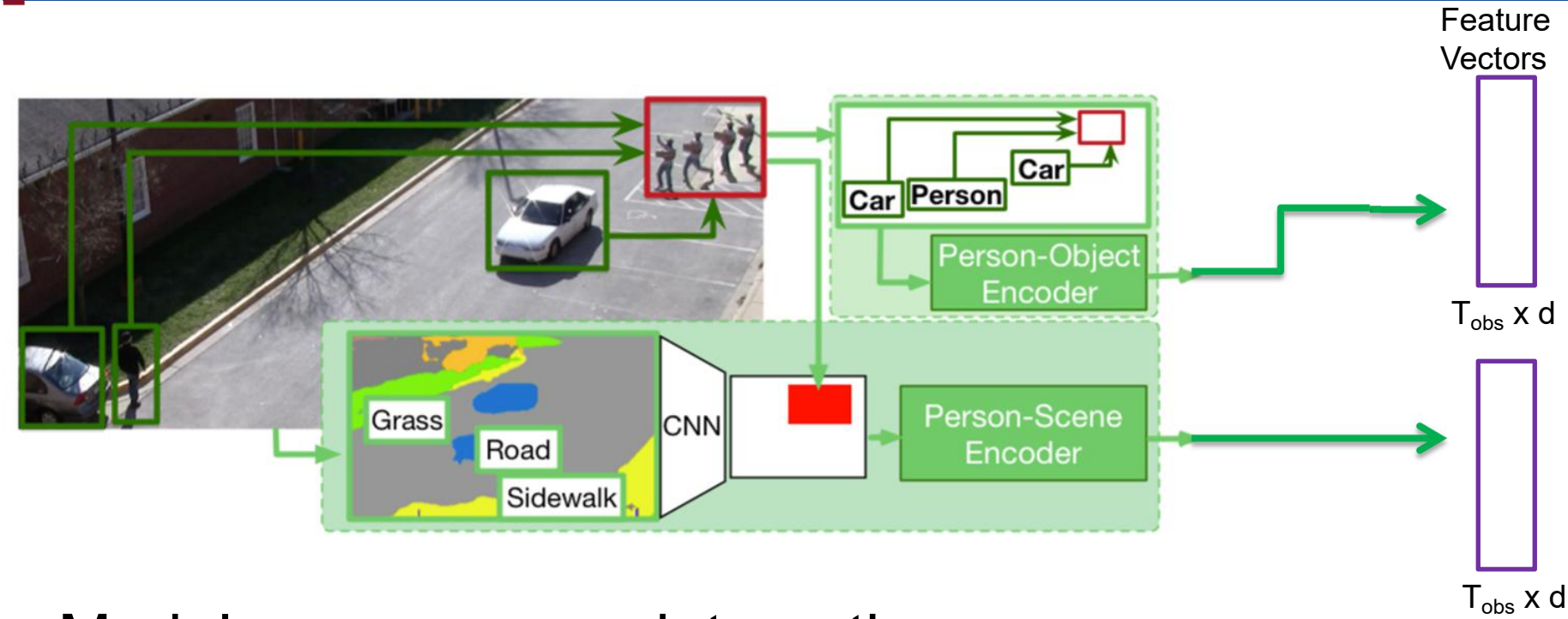
This module has 2 sub-modules:

- Person Appearance Encoder (uses a pre-trained object detection model) Source: J. Liang et al.

- Person Keypoint Encoder (uses a pre-trained person keypoint model) Source: H.-S. Fang et al.

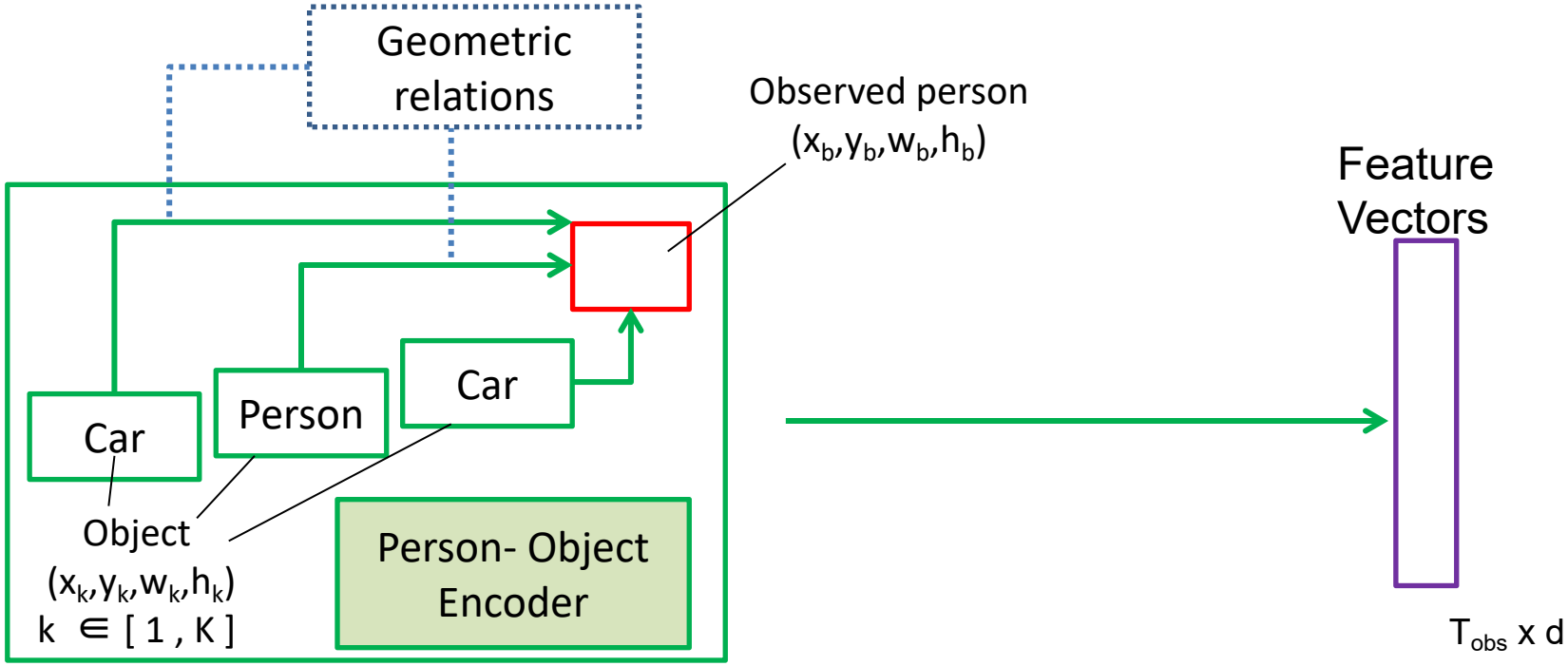
(both related to visual information about every person in the scene)

Person Interaction Module



- Model person-scene interactions
- Model person-object interactions

Person Interaction Module: Person-Object Encoder



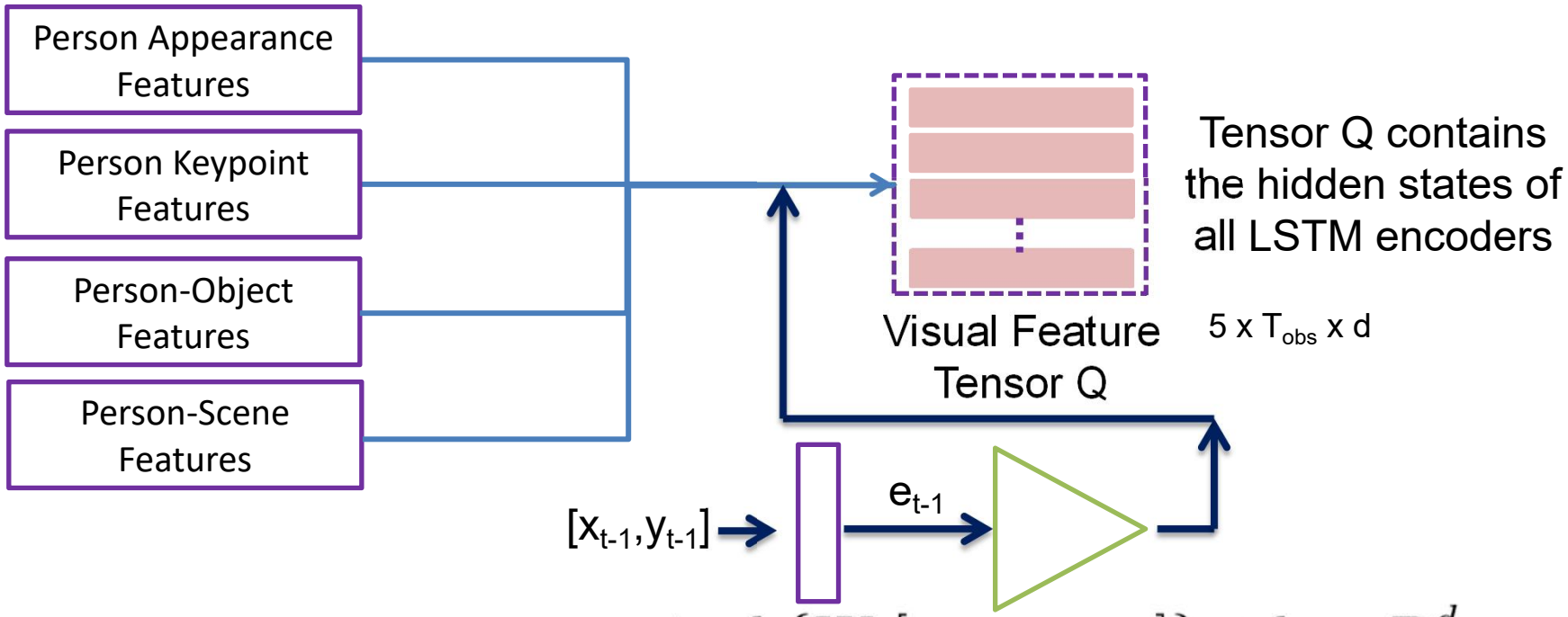
- compute the geometric relation $\mathcal{G}$ *where*

$$\mathcal{G}_k = [\log(\frac{|x_b - x_k|}{w_b}), \log(\frac{|y_b - y_k|}{h_b}), \log(\frac{w_k}{w_b}), \log(\frac{h_k}{h_b})]$$

geometic distance

fraction box size

Visual Features



Trajectory embedding: $e_{t-1} = \tanh\{W_e[x_{t-1}, y_{t-1}]\} + b_e \in \mathbb{R}^d$

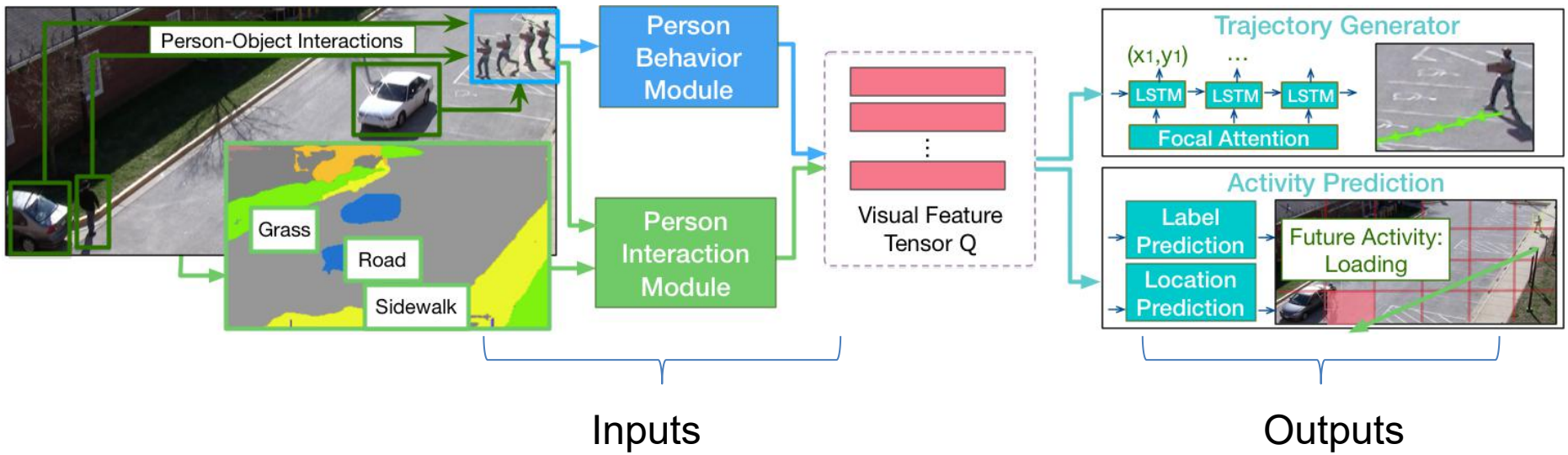
(fully connected layer wth activation function *tanh)*

The trajectory embedding function is obtained from the trajectory output at the last time instant and is fed into an LSTM encoder

The hidden states of all 5 LSTMs are packed into a tensor

Overview

Model Overview

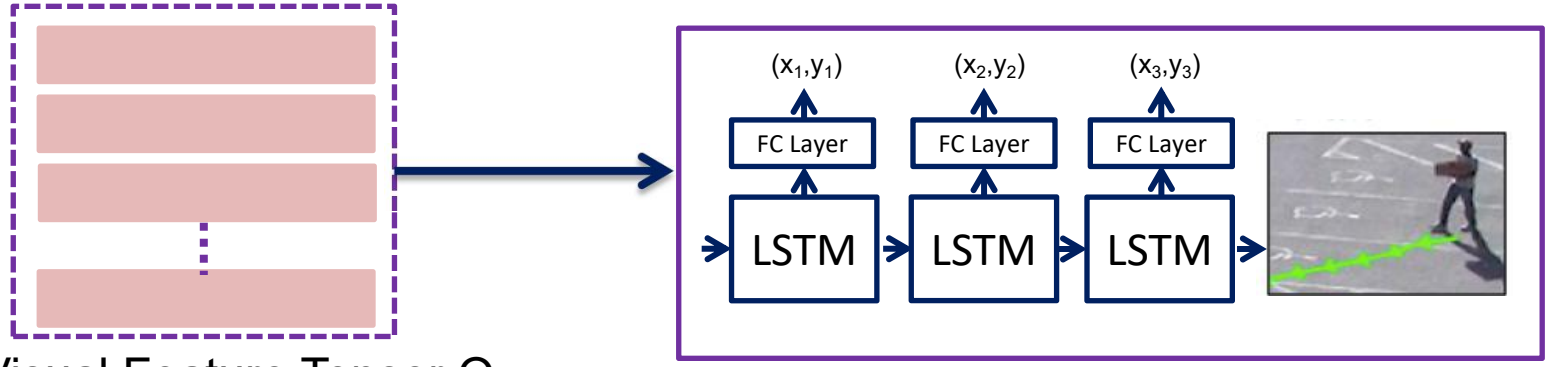


Input feature vectors are all passes into a 3D tensor

Then multi-task prediction is done (simultaneuosly predicting both the trajectory and the future activity)

Trajectory Generator

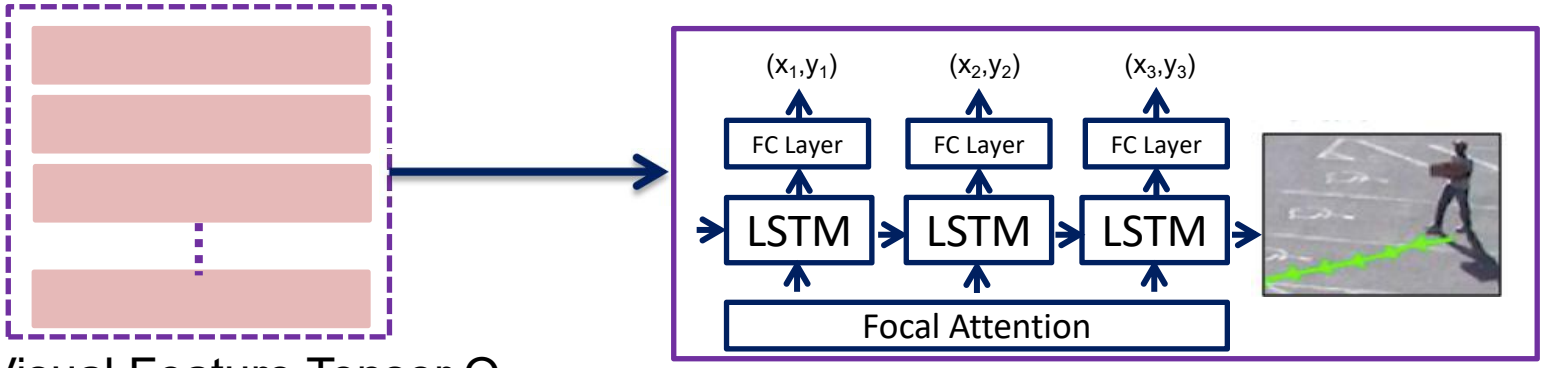
Without Focal Attention:



Visual Feature Tensor Q

LSTM decoder: $h_t = \text{LSTM}\ (\ h_{t-1}\ ,\ [e_{t-1}]\ )$

With Focal Attention:

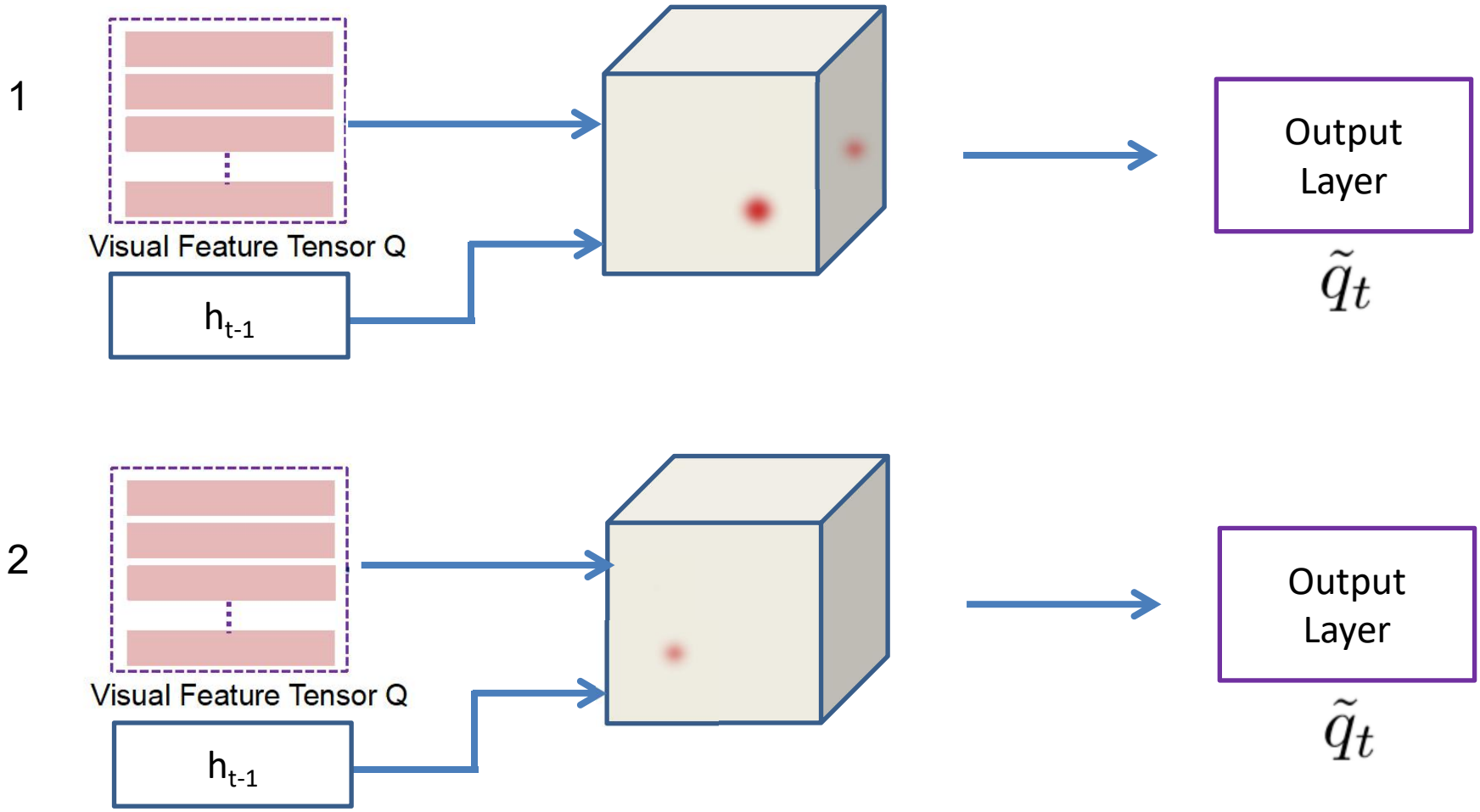


Visual Feature Tensor Q

LSTM decoder: $h_t = \text{LSTM}\ (\ h_{t-1}\ ,\ [e_{t-1}, \tilde{q}_t]\ )$

Focal attention identifies what content of the sequential data needs to be focused on to to make correct predictions

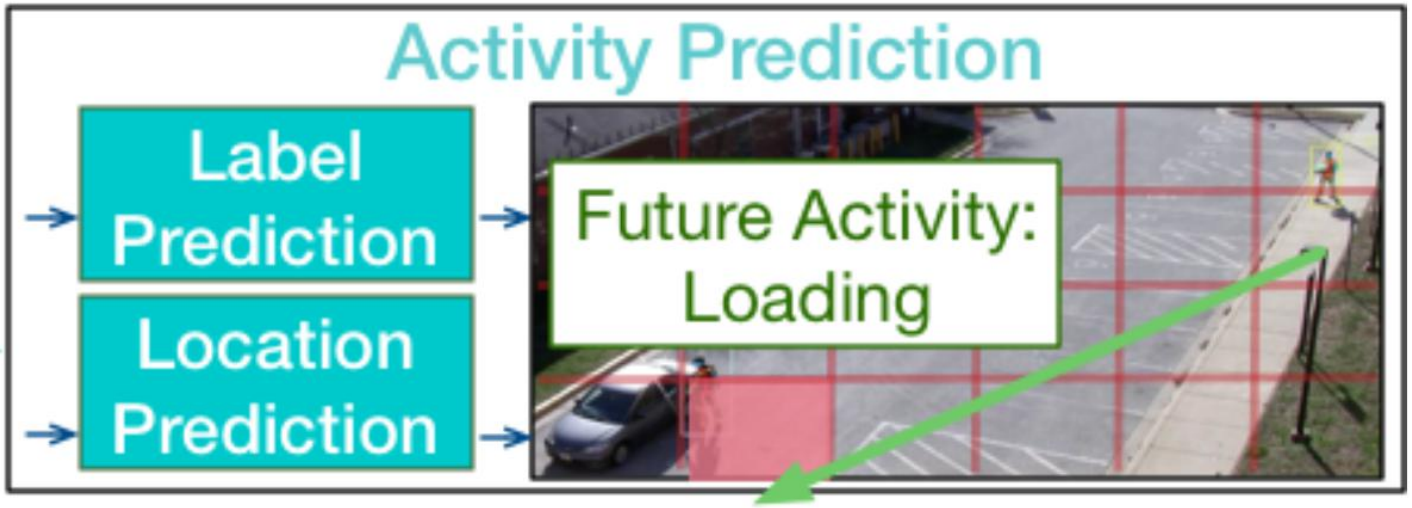
Focal Attention



Source of FVTA: J.Liang et al

In summary: focal attention models the correlation among different features and summarizes them into a lower dimensional feature vector

Activity Prediction



Main Task: activity label prediction

Auxiliary Task: activity label prediction

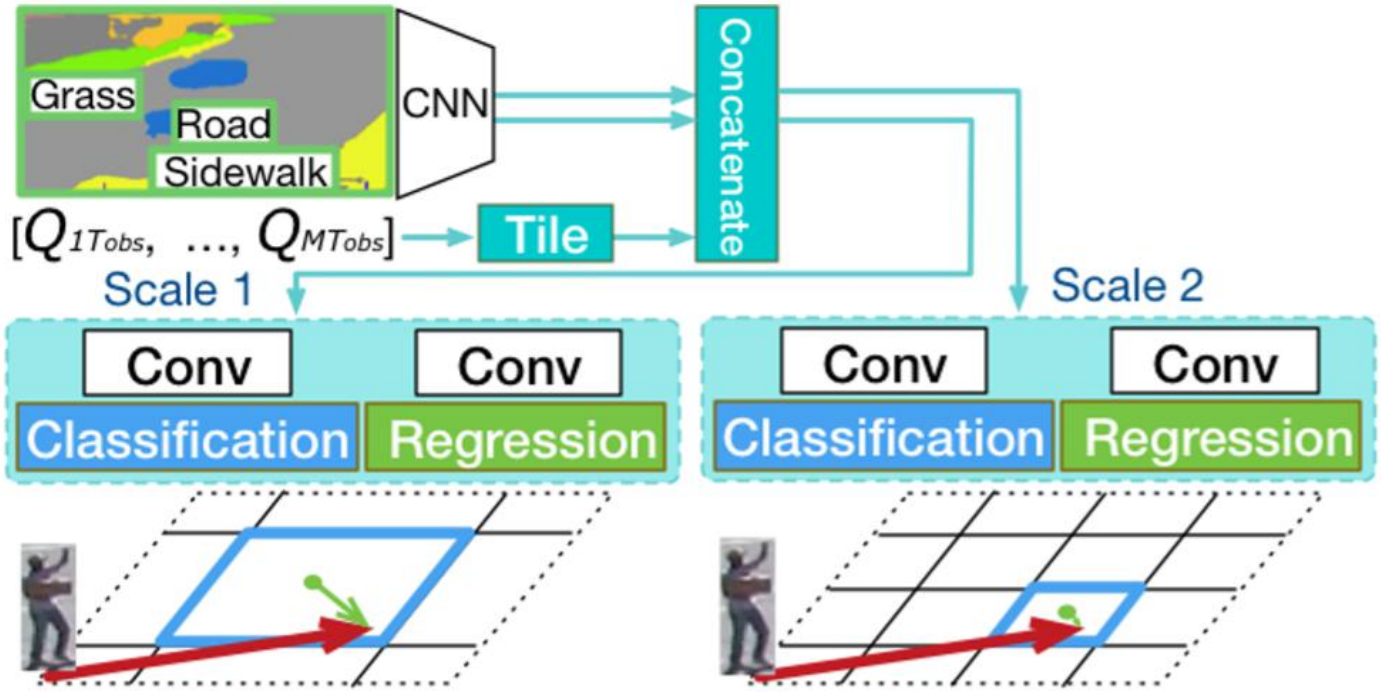
Purpose of auxiliary task:

- to bridge the gap between trajectory generation and activity label prediction

- to minimize the errors in trajectory that have been accumulating over time

Activity Location Prediction:

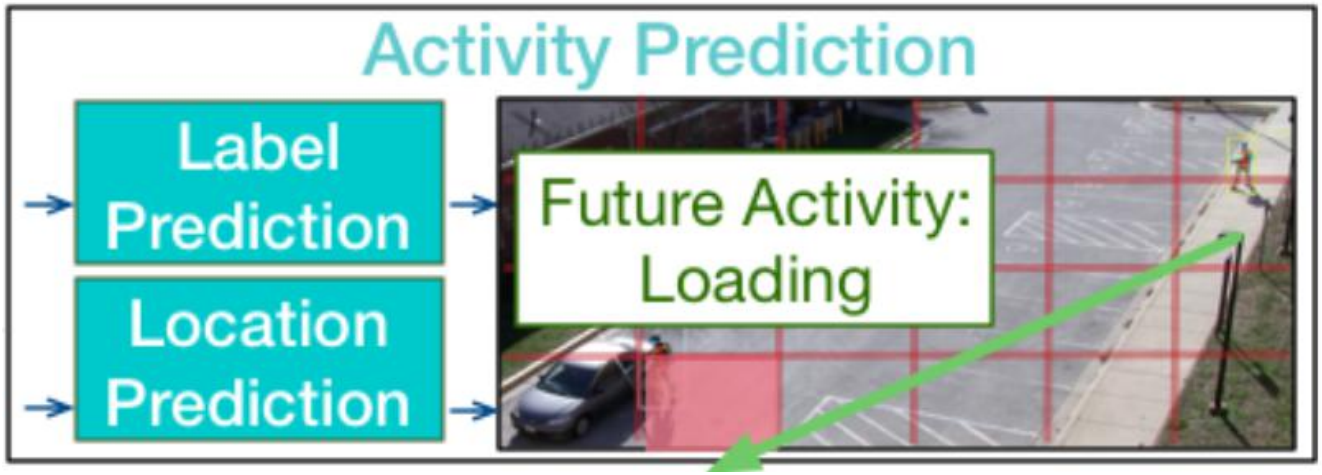
- predicts the final location of the person's activity



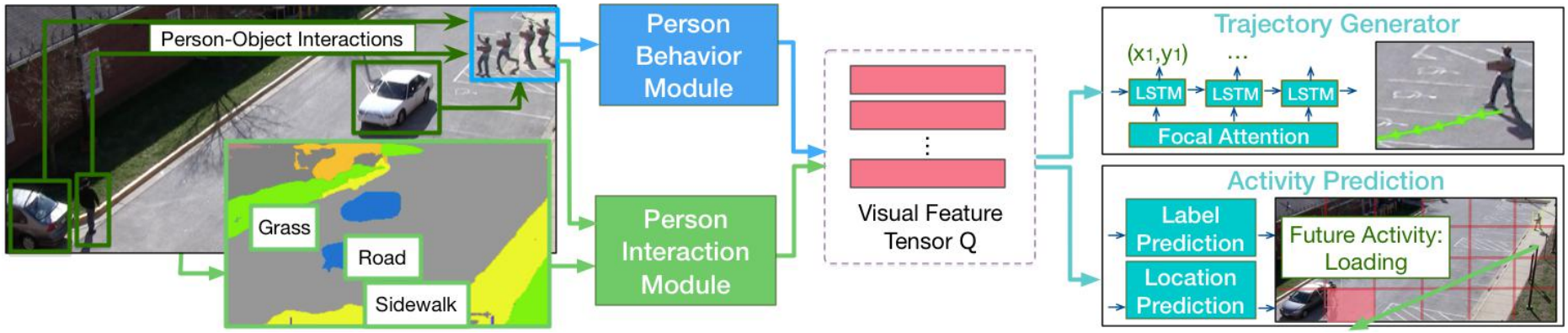
Manhattan Grid
h x w

- **<u>Activity Label Prediction:</u>**
- Predicts the person's future activity at time T_{pred}

$$\text{cls}_{act} = \text{softmax}(W_a \cdot [Q_{1T_{obs}:}, \cdots, Q_{MT_{obs}:}])$$



Training: (3 Losses)



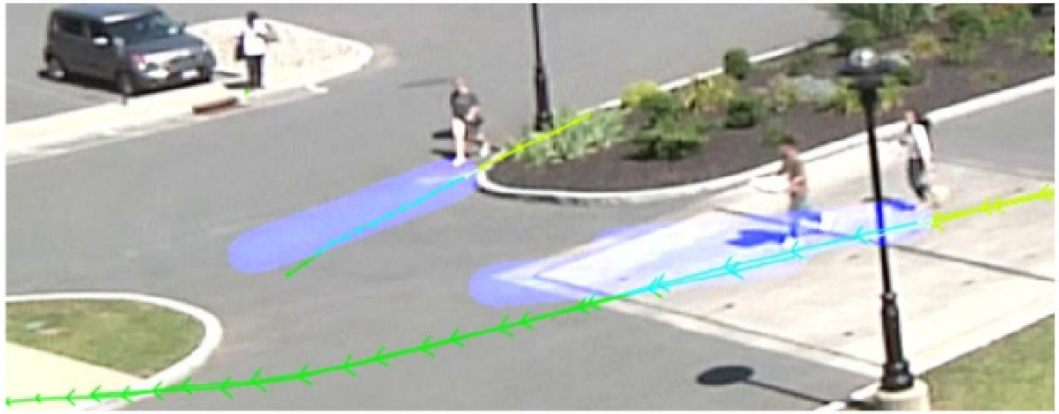
Total Loss:

$$L = \underbrace{L_{xy}}_{\text{Trajectory Prediction Loss}} + \underbrace{\lambda(L_{grid\_cls} + L_{grid\_reg})}_{\text{Activity Location Prediction Loss}} + \underbrace{L_{act}}_{\text{Activity Label Prediction Loss}}$$

Outline

- Introduction
- Related Work (for trajectory and activity prediction)
- Approach: Next Model
  - Person Behavior Module
  - Person Interaction Module
  - Trajectory Generator with Focal Attention
  - Activity Prediction
- Experiments: ActEV/VIRAT and ETH & UCY
- Conclusion

ActEV/VIRAT



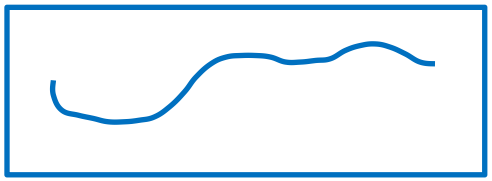
Legend:

- yellow path: observed trajectory
- green path: ground truth
- blue heatmap: predictions

- Public dataset for ativity detection released by NIST in 2018
- Metrics used for trajectory prediction:

$$\mathrm{ADE} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T_{pred}} \|\tilde{Y}_t^i - Y_t^i\|_2}{N * T_{pred}}$$

Average Displacement Error



$$\mathrm{FDE} = \frac{\sum_{i=1}^{N} \|\tilde{Y}_{T_{pred}}^i - Y_{T_{pred}}^i\|_2}{N}$$

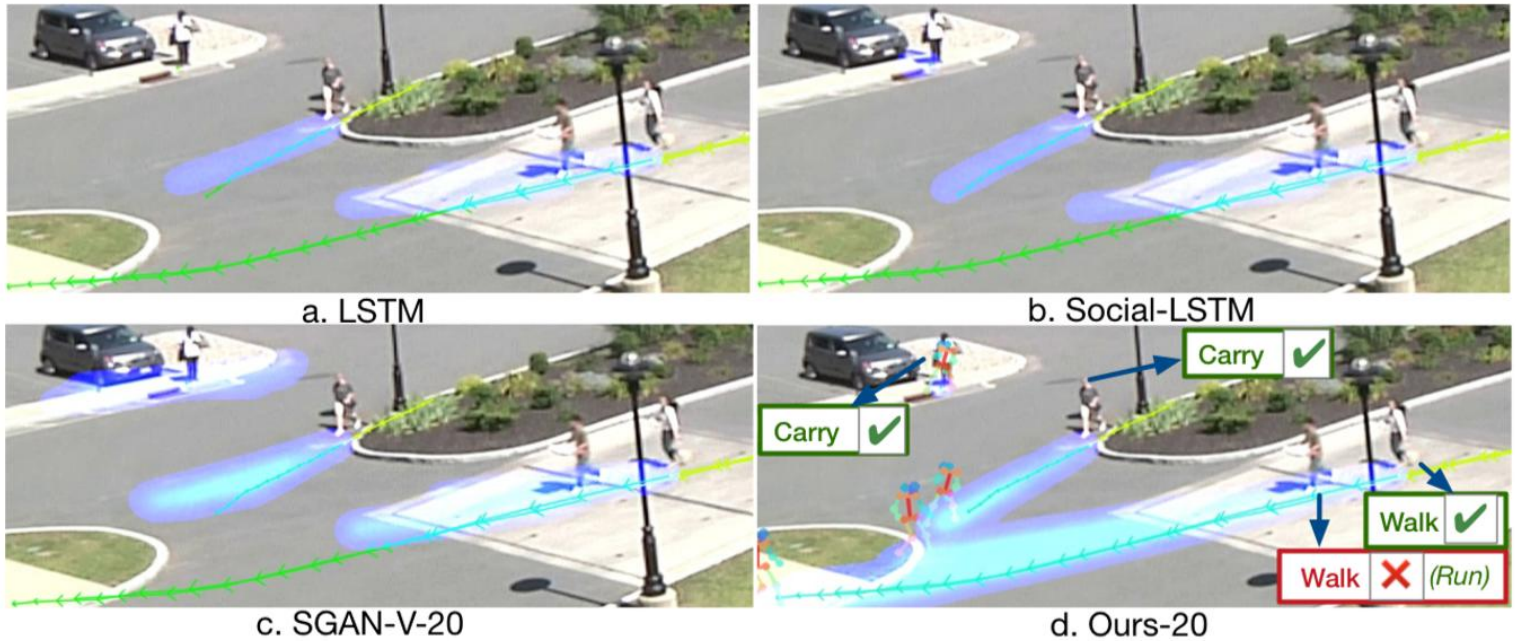
Final Displacement Error

Comparison to Baseline Methods Quantitative Results

	Method	ADE	FDE	move_ADE	move_FDE
Single Model	Linear	32.19	60.92	42.82	80.18
	LSTM	23.98	44.97	30.55	56.25
	Social LSTM	23.10	44.27	28.59	53.75
	SGAN-PV	30.51	60.90	37.65	73.01
	SGAN-V	30.48	62.17	35.41	68.77
	★ Ours	**17.99**	**37.24**	**20.34**	**42.54**
	★ Ours-Noisy	34.32	57.04	40.33	66.73
20 Outputs	SGAN-PV-20	23.11	41.81	29.80	53.04
	SGAN-V-20	21.16	38.05	26.97	47.57
	★ Ours-20	**16.00**	**32.99**	**17.97**	**37.28**

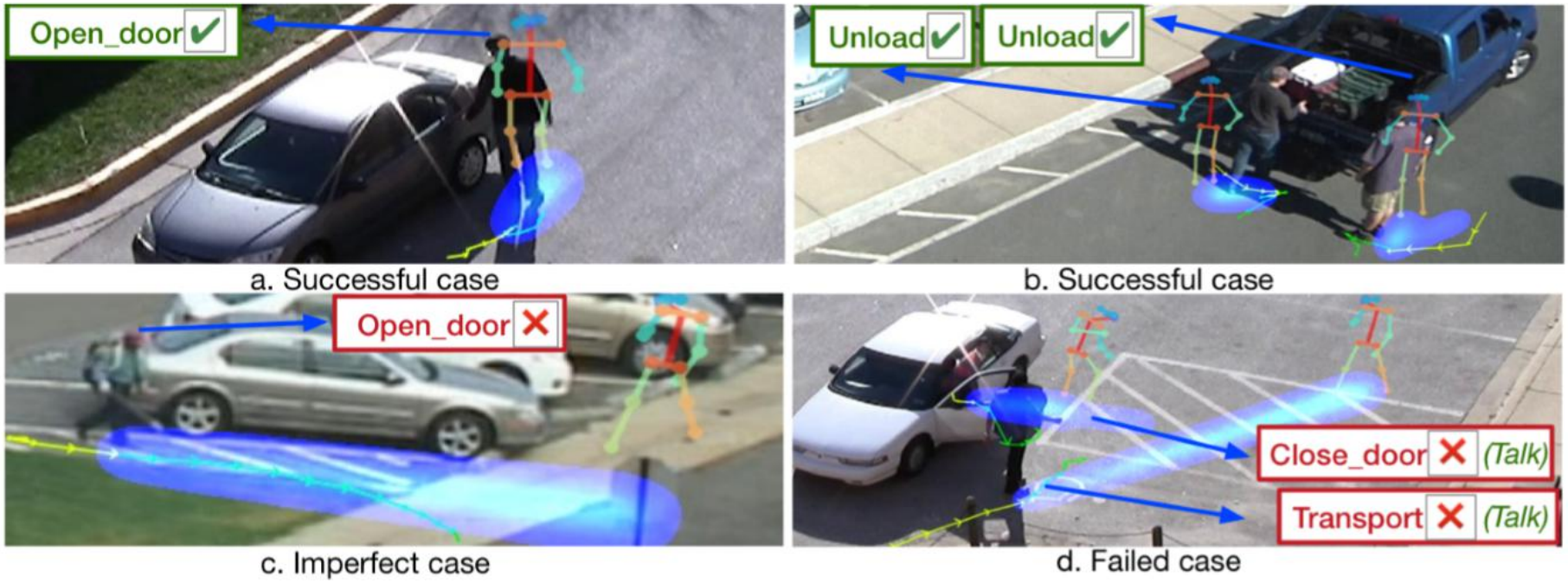
- This method outperforms all other baseline methods
- Experiments are divided into static and moving scenes
- Best performance is achieved when repeated 20 times

Comparison to Baseline Methods Qualitative Reults



- Next model results in better trajectory predictions for all 3 people.
- In addition to that, it also predicts future activities

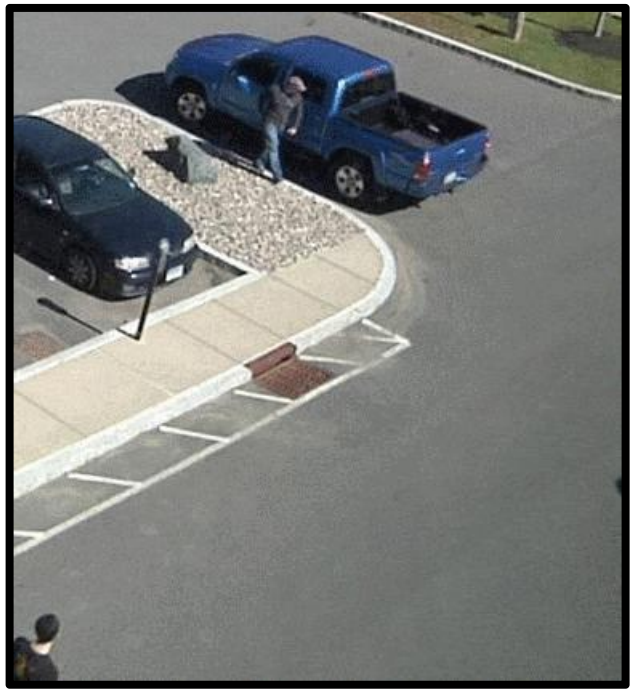
Further Qualitative Results - Next Model

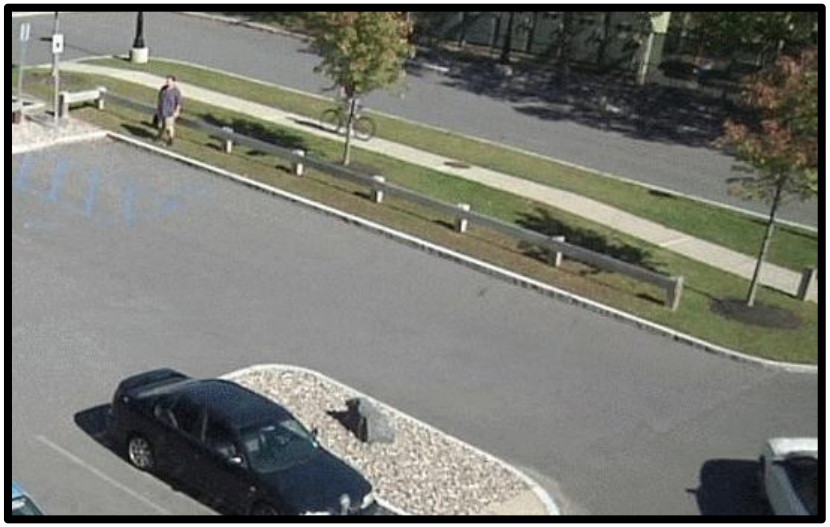


a. Successful case
b. Successful case
c. Imperfect case
d. Failed case

- Successful case is where both future trajectory and activity are predicted correctly
- Imperfect case predicts the correct trajectory but incorrect activity
- Failed case predicts both trajectory and activity incorrectly

Further Qualitative Results - Next Model





- Few frames are observed and then predictions are made on both the future activity and trajectory

Ablation Studies

Method	ADE ↓	FDE ↓	Act mAP ↑
Our full model	17.91	37.11	0.192
No p-behavior	18.99	39.82	0.139
No p-interaction	18.83	39.35	0.163
No focal attention	19.93	42.08	0.144
No act label loss	19.48	41.45	-
No act location loss	19.07	39.91	0.152
No multi-task	20.37	42.79	-

- Person behavior features are more essential
- Worse results if focal attention is removed
- Worse results if multi-task learning is removed

ETH & UCY

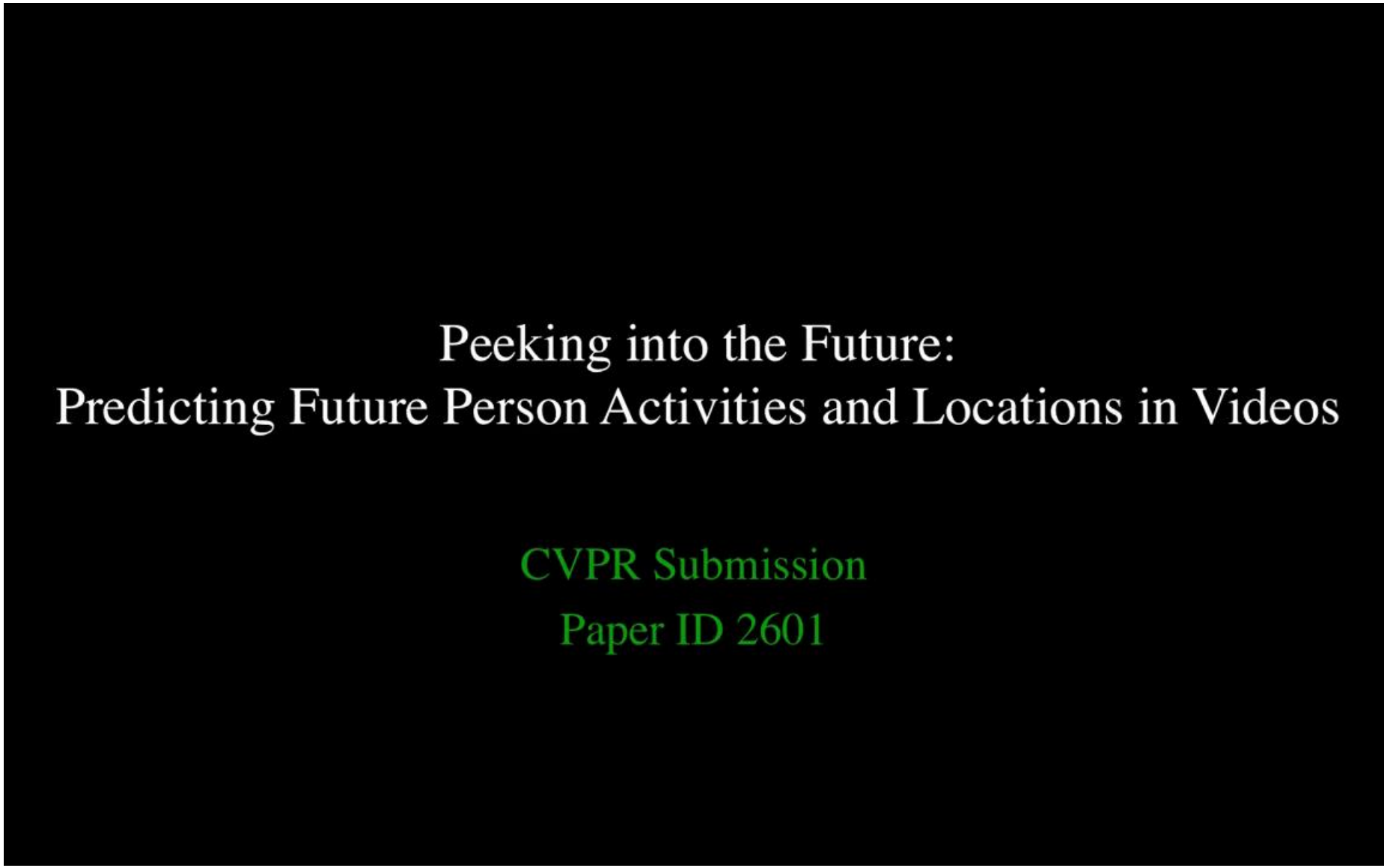
- These datasets are for predicting person trajectory

	Method	ETH	HOTEL	UNIV *	ZARA1	ZARA2	AVG
Single Model	Linear	1.33 / 2.94	0.39 / 0.72	0.82 / 1.59	0.62 / 1.21	0.77 / 1.48	0.79 / 1.59
	LSTM	1.09 / 2.41	0.86 / 1.91	**0.61 / 1.31**	**0.41 / 0.88**	0.52 / 1.11	0.70 / 1.52
	Alahi *et al.* [1]	1.09 / 2.35	0.79 / 1.76	0.67 / 1.40	0.47 / 1.00	0.56 / 1.17	0.72 / 1.54
	Ours-single-model	**0.88 / 1.98**	**0.36 / 0.74**	0.62 / 1.32	0.42 / 0.90	**0.34 / 0.75**	**0.52 / 1.14**
20 Outputs	Gupta *et al.* [7](V)	0.81 / 1.52	0.72 / 1.61	0.60 / 1.26	0.34 / 0.69	0.42 / 0.84	0.58 / 1.18
	Gupta *et al.* [7](PV)	0.87 / 1.62	0.67 / 1.37	0.76 / 1.52	0.35 / 0.68	0.42 / 0.84	0.61 / 1.21
	Sadeghian *et al.* [26]	**0.70 / 1.43**	0.76 / 1.67	**0.54 / 1.24**	**0.30 / 0.63**	0.38 / 0.78	0.54 / 1.15
	Ours-20	0.73 / 1.65	**0.30 / 0.59**	0.60 / 1.27	0.38 / 0.81	**0.31 / 0.68**	**0.46 / 1.00**

ADE FDE

DEMO

DEMO


Peeking into the Future:
Predicting Future Person Activities and Locations in Videos
CVPR Submission
Paper ID 2601

Outline

- Introduction
- Related Work (for trajectory and activity prediction)
- Approach: Next Model
  - Person Behavior Module
  - Person Interaction Module
  - Trajectory Generator with Focal Attention
  - Activity Prediction
- Experiments: ActEV/VIRAT and ETH & UCY
- Conclusion

Conclusion

- The proposed model in this paper is used to predict future activities of pedestrians as well as future activities
- A person is modeled using visual semantics such as appearance and behavior
- Experiments were done on two common benchmarks and results show that it outperforms state-of-the-art methods
- Applications of this model range from autonomous driving to human assisting robots

References

- [1] J. Liang, L. Jiang, J. Niebles, A. Hauptmann, L. Fei Peeking into the Future: Predicting Future Person Activities and Locations in Videos
- [2] J.Liang,L.Jiang,L.Cao,L.-J.Li,andA.Hauptmann. Focal visual-text attention for visual question answering. In CVPR, 2018. 3, 4
- [3] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. Social lstm: Human trajectory prediction in crowded spaces
- [4]S. Pellegrini, A. Ess, and L. Van Gool. Improving data association by joint modeling of pedestrian trajectories and groupings
- [5] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu. RMPE: Regional multi-person pose estimation

Thank you!

Questions?

Extras

- Focal Attention

$$S^t = h_{t-1}^{\mathsf{T}} \cdot Q_{ij}$$

$$A^t = \mathrm{softmax}(\max_{i=1}^{M} S_{i:}^t) \in \mathbb{R}^M$$

$$B^t = [\mathrm{softmax}(S_{1:}^t), \cdots, \mathrm{softmax}(S_{M:}^t)] \in \mathbb{R}^{M \times T_{obs}}$$

$$\tilde{q}_t = \sum_{j=1}^{M} A_j^t \sum_{k=1}^{T_{obs}} B_{jk}^t Q_{jk:} \in \mathbb{R}^d$$

Extras

- Object and Activity Classes

	Classes
Object	Bike, Construction_Barrier, Construction_Vehicle, Door, Dumpster, Parking_Meter, Person, Prop, Push_Pulled_Object, Vehicle
Activity	Carry, Close_Door, Close_Trunk, Crouch, Enter, Exit, Gesture, Interaction, Load, Object_Transfer, Open_Door, Open_Trunk, PickUp, PickUp_Person, Pull, Push, Ride_Bike, Run, SetDown, Sit, Stand, Talk, Talk_phone, Texting, Touch, Transport, Unload, Use_tool, Walk

Extras

- Single Feature Ablation and Activity Detection Experiments

Method	ADE ↓	FDE ↓	Act mAP ↑
Our full model	17.91	37.11	0.192
No p-object	18.17	37.13	0.198
No p-scene	18.18	37.75	0.206
No p-keypoint	18.25	37.96	0.190
No p-appearance	18.20	37.79	0.154
Act Detect	18.27	37.68	0.155