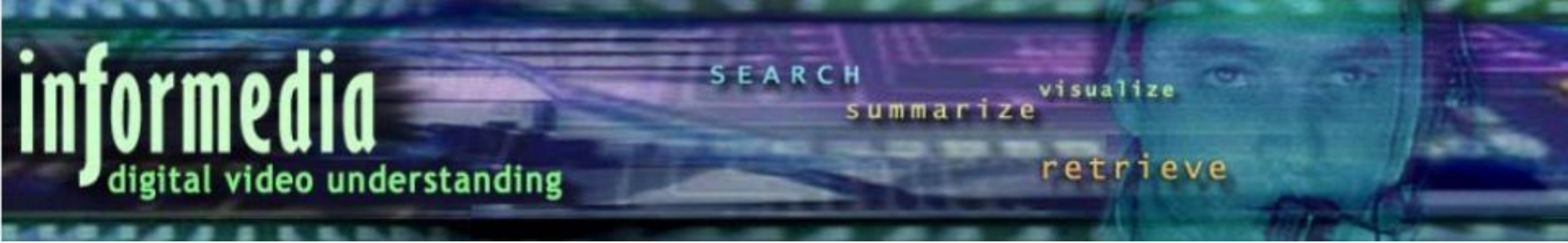


Thesis Defense

From Recognition to Prediction: Analysis of Human Action and Trajectory Prediction in Video

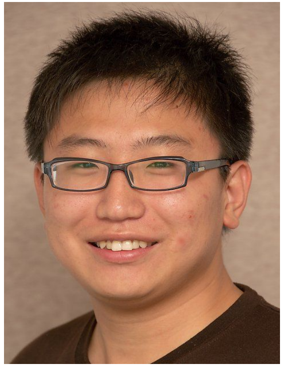
Junwei Liang
junweil@cs.cmu.edu

Carnegie Mellon University
Language Technologies Institute

Thesis Committee

- Prof. Alexander Hauptmann (Chair)
- Prof. Alan W Black
- Prof. Kris Kitani
- Dr. Lu Jiang (Google Research)



Some notes for the audience

- Please mute your mic; you can turn on video if you'd like
- Please ask only clarification questions during the presentation: unmute and ask or post them on chat

We Predict the Future Trajectory of Pedestrians

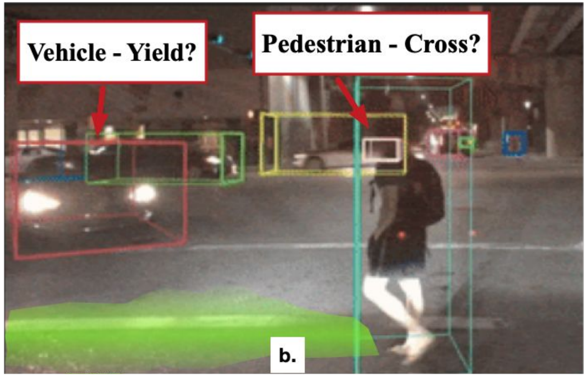
- Models observe 3~5 seconds
- Predict future 5~12 seconds

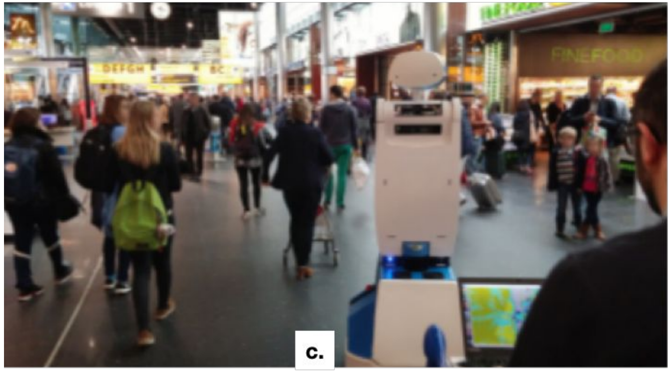
We Predict the Future Trajectory of Pedestrians

- Models observe 3~5 seconds
- Predict future 5~12 seconds
    - Human intentions (future actions) are predicted as well

Why Pedestrian Trajectory Prediction?

- Important in many real-world applications
  - Self-driving cars
  - Socially-aware robots
  - Advanced public safety monitoring - crowd dynamics estimation



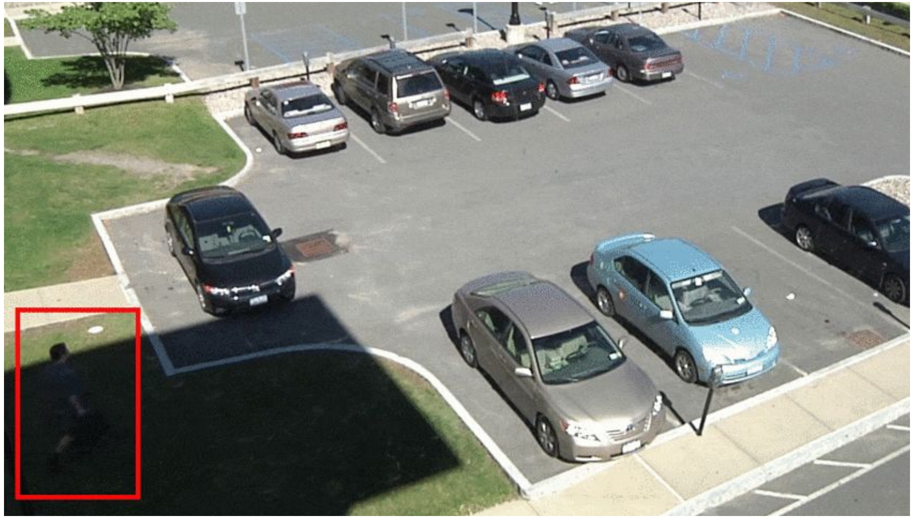


Research Challenges

- Difficulties for trajectory prediction
  - The scene constraints are complex and they are changing dynamically
    - Static scene constraints like sidewalk, crosswalk
    - Traffic actors like vehicles

Research Challenges

- Difficulties for trajectory prediction
  - The future is uncertain
  - Training data is limited for rare scenarios

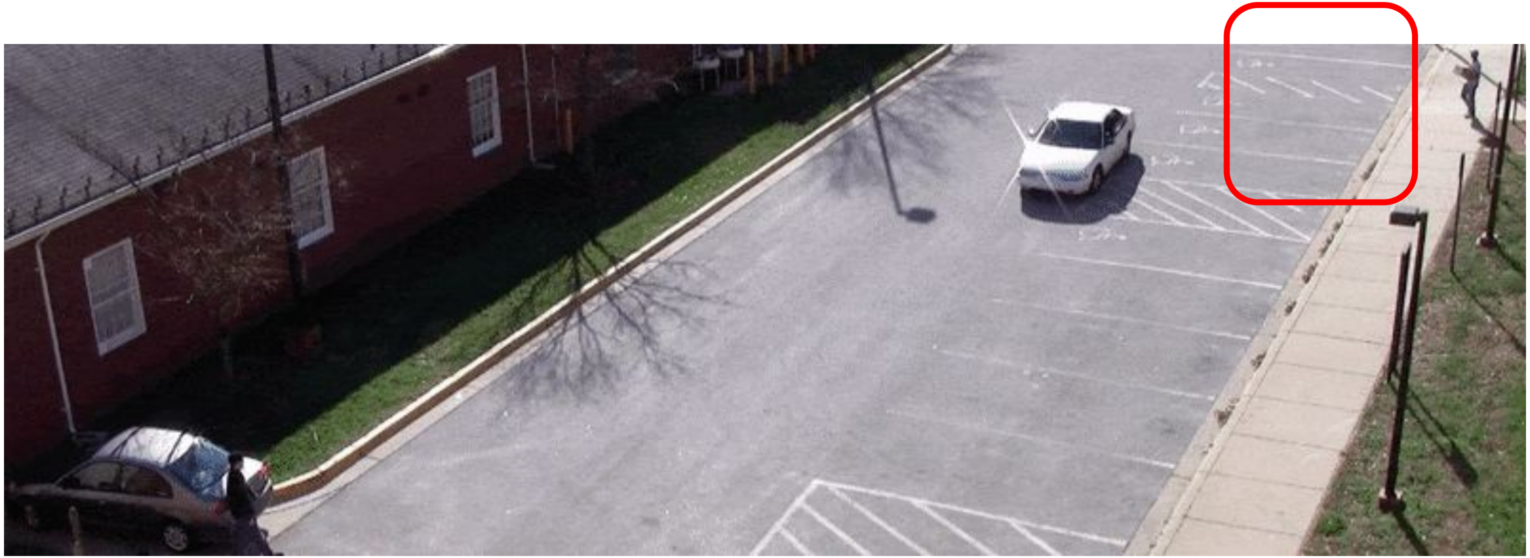


Thesis Goal and Focus

- Goal
  - To build **a robust pedestrian trajectory prediction system** by jointly analyzing **human actions** and **scene semantics**.
- Our focus
  - P1. Action Analysis
  - P2. Trajectory Prediction with Scene Semantics
  - P3. Analysis of Actions and Trajectory Prediction

Why Action Analysis?

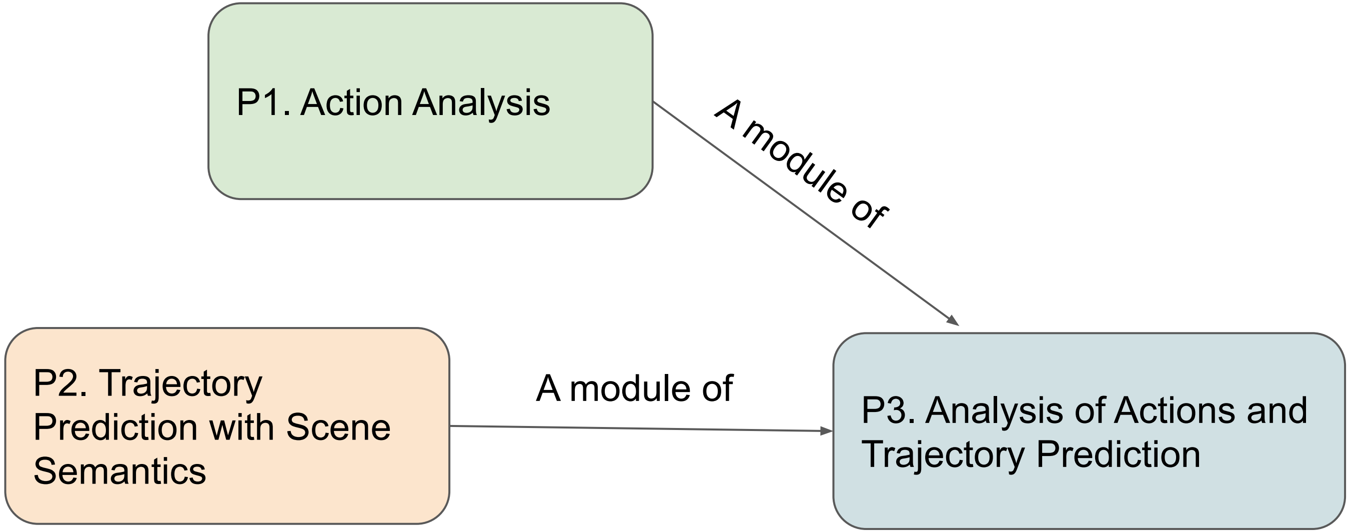
To better predict person's intent, models should detect subtle **actions** during observation.



See in the red box where the target person performs the action "wave hand".

Tasks and Their Relation

- Given a set of videos:



"Prediction using scene semantics and action representation"

Thesis Breakdown

	Challenges / Directions	Solutions
P1. Action Analysis	Lack of training data for rare actions	Webly Learning (C3, IJCAI, ICMR)
	Viewpoint Robustness	Spatial-temporal Alignment Network (C4, In Submission)
P2. Trajectory Prediction with Scene Semantics	Multi-modal prediction	Model and Dataset (C5, CVPR)
	Domain Robustness	Simulation-as-Augmentation (C6, ECCV)
P3. Analysis of Actions and Trajectory Prediction	Representation Robustness	Joint Action and Trajectory Prediction (C7, CVPR)
	Lack of dataset in traffic scenes	New annotated Multi-view dataset (C8)
	Long-term Trajectory Prediction	Modeling with action-enhanced contextual cues (C9)

Thesis Organization

P1. Action Analysis	P2. Trajectory Prediction with Scene Semantics	P3. Analysis of Actions and Trajectory Prediction
Efficient Object Detection and Tracking (C2) Weakly-supervised Learning (C3)	Multi-modal Future Trajectory Prediction (C5)	Joint Action and Trajectory Prediction (C7)
Viewpoint-Invariant Representation Learning (C4)	Simulation-as-Augmentation Robust Learning (C6)	Long-term Trajectory Prediction Using Scene Semantics and Action Representation (C8 & C9)

Focuses of This Presentation

P1. Action Analysis	P2. Trajectory Prediction with Scene Semantics	P3. Analysis of Actions and Trajectory Prediction
Efficient Object Detection and Tracking (C2) Weakly-supervised Learning (C3)	**Multi-modal Future Trajectory Prediction (C5)**	Joint Action and Trajectory Prediction (C7)
Viewpoint-Invariant Representation Learning (C4)	Simulation-as-Augmentation Robust Learning (C6)	**Long-term Trajectory Prediction Using Scene Semantics and Action Representation (C8 & C9)**

Roadmap

- P1. Action Analysis
- P2. Trajectory Prediction with Scene Semantics
- P3. Analysis of Actions and Trajectory Prediction
- Vision and Future Directions
- Conclusions

Roadmap

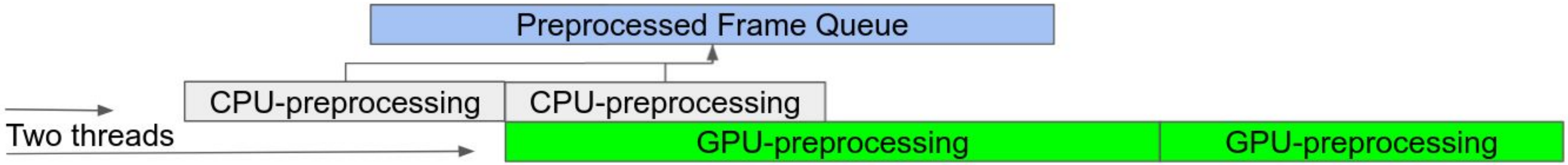
- P1. Action Analysis
  - C2. Efficient Object Detection and Tracking
  - C3. Weakly-Supervised Action Event Recognition
  - C4. Viewpoint Invariant Representation Learning
- P2. Trajectory Prediction with Scene Semantics
- P3. Analysis of Actions and Trajectory Prediction
- Vision and Future Directions
- Conclusions

C2. Efficient Object Detection and Tracking in Video

- In this chapter, our goal is to build an efficient object detection and tracking framework for extended videos
  - This usually called the “Perception” system in Self-driving systems
  - Not to beat SOTA
  - But to establish a flexible framework for any new object detection models

C2. Efficient Object Detection and Tracking in Video

- Contributions
    - **Optimized parallel processing using Tensorflow**
        - **More than 70% faster than official code**
        - **This system is part of the system that won the Activities in Extended Videos Prize Challenge (ActEV) in 2019**
        - **Github got 240+ stars and 80+ forks**



C2. Efficient Object Detection and Tracking in Video

- Visualization - Outdoor video with small person

Roadmap

- P1. Action Analysis
  - C2. Efficient Object Detection and Tracking
  - C3. Weakly-Supervised Action Event Recognition
  - C4. Viewpoint Invariant Representation Learning
- P2. Trajectory Prediction with Scene Semantics
- P3. Analysis of Actions and Trajectory Prediction
- Vision and Future Directions
- Conclusions

C3. Weakly-Supervised Action Event Recognition

- Motivation
  - Since human actions are diverse and combination of atomic actions can lead to an exponential amount of action classes, manually-annotated training data is often insufficient
  - Not enough supervised data for long-tail actions
  - To mitigate that, we propose to
    - Leveraging **webly-labeled** data
    - Utilizing multi-modal prior knowledge

“Walking with dog” video example

C3. Weakly-Supervised Action Event Recognition

- Contributions
  - We are one of the early works that study how we could better utilize weakly-supervised video data from the Internet
  - Our algorithm is able to outperform supervised training on manually-labeled data given enough noisy web data
  - Our algorithm has won several TRECVID challenges on Ad-hoc Video Search

Roadmap

- P1. Action Analysis
    - C2. Efficient Object Detection and Tracking
    - C3. Weakly-Supervised Action Event Recognition
    - C4. Viewpoint Invariant Representation Learning
- P2. Trajectory Prediction with Scene Semantics
- P3. Analysis of Actions and Trajectory Prediction
- Vision and Future Directions
- Conclusions

Why do we need viewpoint invariant models?

- Action representation should be viewpoint invariant
- Videos have camera motion and cut scene changes
    - Traditional convolution networks are not designed for viewpoint changes

Video from AVA dataset

Multi-view dataset

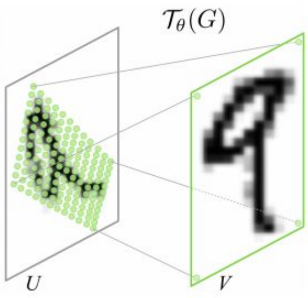
Previous Work

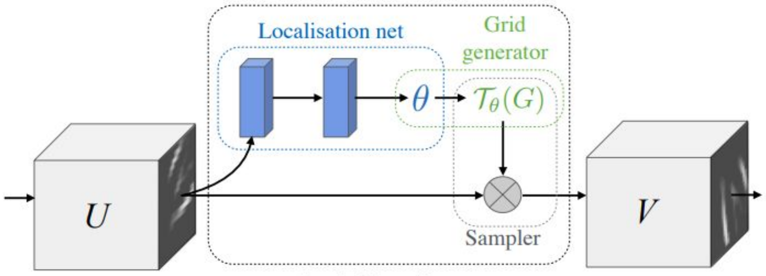
- Action recognition models - representation learning
  - Inception-3D (CVPR'17)
  - S3D (ECCV'18)
  - Non-local neural network (CVPR'18)
  - SlowFast Networks (ICCV'19)
- Viewpoint invariant models - mostly for images
  - Spatial Transformer Networks (NeurIPS'15)
  - Dynamic Routing Between Capsules (NeurIPS'17)
  - VideoCapsuleNet (NeurIPS'18)
  - Stacked Capsule Autoencoder (NeurIPS'19)

Spatial Transformer Networks (NeurIPS'15)

- Given spatial input, rearrange and get output
- A localization net to output affine transformation matrix (6 DoF) based on the input

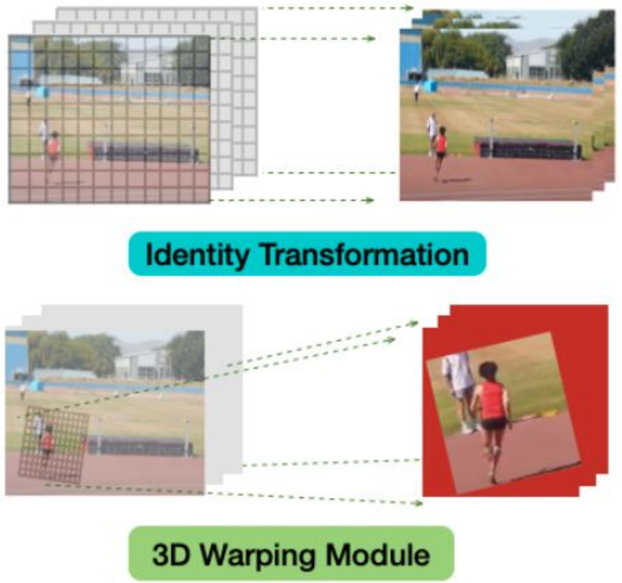
$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = \mathcal{T}_\theta(G_i) = \mathtt{A}_\theta \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix}$$





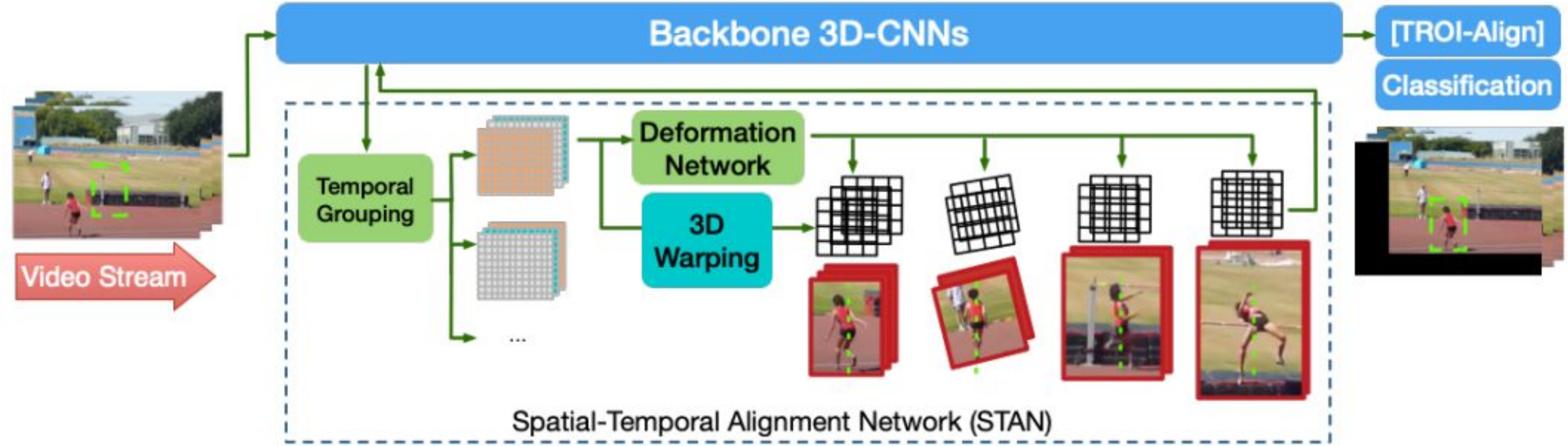
Proposed: Spatial-Temporal Alignment Network for Action Recognition

- We propose to do so for 3D video inputs



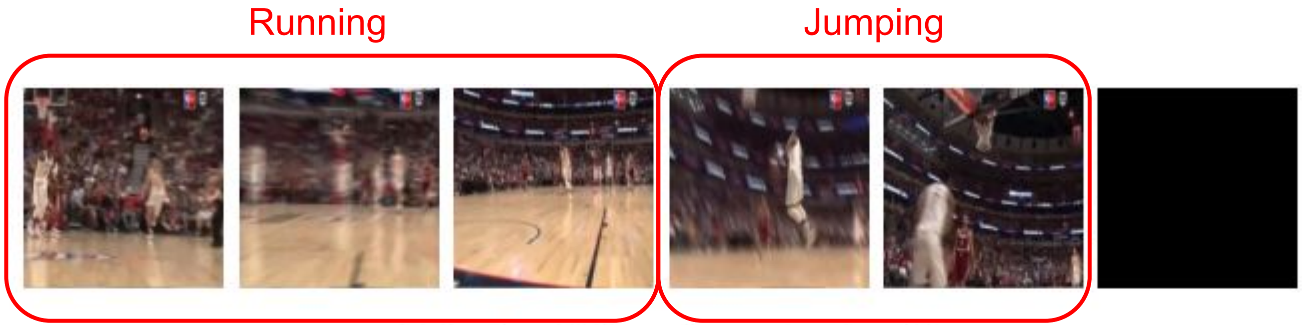
Spatial-Temporal Alignment Network for Action Recognition

- STAN Technical Details
    - The deformation network takes feature maps and outputs transformation matrix
    - Temporal grouping: different temporal slices of the feature map undergo different transformations



Technical Details: Temporal Grouping

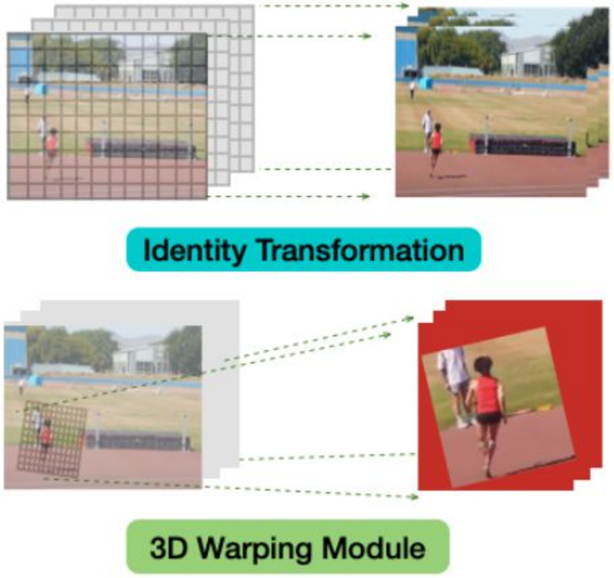
- We group video frames temporally to compute the transformation matrix
- Intuition: actions have sub-actions that would need different level of temporal scaling for better recognition



(n) dunking basketball

Transformation Visualization

- The computed transformation matrix is used to warp feature maps
    - Typical transformations include rotation, scaling and translation



Spatial-Temporal Alignment Network for Action Recognition

- Experimental Design
  - Baselines
    - ResNet3D
    - SlowFast (ICCV'19)
  - Datasets
    - Common benchmark
      - Kinetics-400
      - AVA
      - AVA-Kinetics
      - Charades
    - Multi-viewpoint dataset
      - Charades-Ego
      - MEVA

Experiments on Common Benchmark

- Kinetics-400
  - We re-implemented SlowFast and ResNet3D using Tensorflow
  - 3x10 clips inference

Models	top-1	top-5	GFLOPs
I3D [22]	0.711	0.893	-
R(2+1)D [44]	0.720	0.900	-
DynamoNet (32 frames) [7]	0.714	0.900	-
NL-R50 (32 frames) [49]	0.749	0.916	-
ResNet3D (8x8)	0.735	0.908	109.2
ResNet3D + *STAN*	**0.751**	**0.916**	113.2
SlowFast [9] (32x2)*	0.759	0.920	131.7
SlowFast + *STAN*	**0.774**	**0.931**	134.5

1.5% absolute improvement with only 2% more computation

Experiments on Common Benchmark

- AVA and Charades

Models	mAP	GFLOPs	MParams
ResNet3D (8x8)	0.234	208.0	31.75
ResNet3D + *STAN*	**0.247**	216.6	32.02
SlowFast [9] (32x2)	0.252	242.6	33.77
SlowFast + *STAN*	**0.268**	247.4	33.96

AVA Dataset

Models	mAP	GFLOPs	MParams
ResNet3D (16x8)	0.354	218.4	32.40
ResNet3D + *STAN*	**0.377**	226.4	32.47
SlowFast [9] (32x4)	0.386	131.7	34.51
SlowFast + *STAN*	**0.406**	134.5	34.53

Charades Dataset

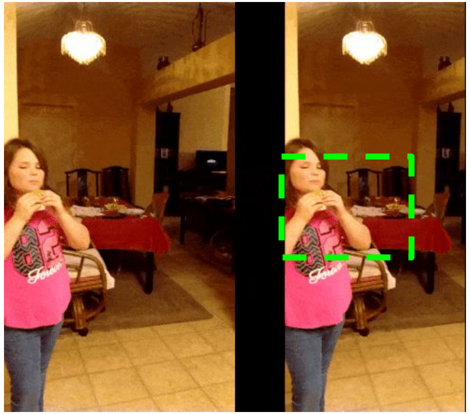
Experiments on Common Benchmark

- Ablation Experiments on AVA dataset
  - Temporal grouping
  - Domain transfer ability
    - Pretrain on K400 and fix transformation network

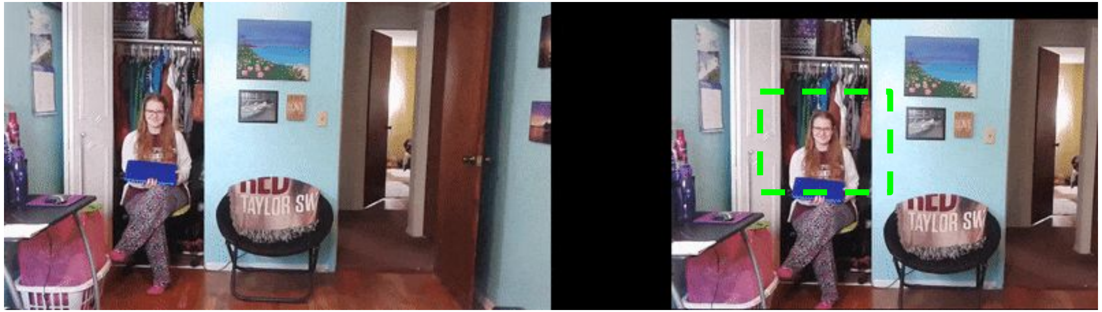
	Diff	mAP	GFLOPs
SlowFast	-	0.252	242.55
+ *STAN*	**+1.6%**	**0.268**	247.40
+ *STAN* (no tg)	+0.8%	0.260	247.40
+ *STAN* (tg=#frames)	-	0.254	246.16
+ *STAN* (fixed W_θ)	+1.2%	0.264	247.40

Qualitative Analysis

- Visualizing transformation
  - Left is original frames. Right is transformed frames
  - The transformation serves as a camera stabilization effect



Eating a sandwich

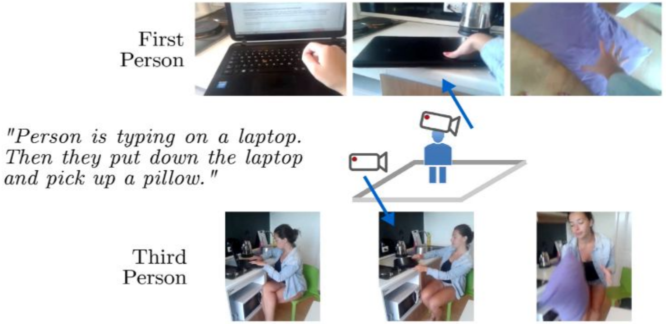


Holding a laptop

*32x4 test clips with temporal group=2, each group is about 2 seconds

Experiments on Multi-viewpoint Dataset

- Charades-Ego and MEVA
  - 3x10 clips inference for each sample
  - MEVA evaluation set: total 7082 activity instances of 35 action classes (from 257 videos)



Charades-Ego

MEVA

Experiments on Multi-viewpoint Dataset

- Charades-Ego and MEVA
  - Multiple-viewpoint for the same action samples

Models	1st-person	3rd-person
Baseline v1.0 [36]	0.282	0.232
ResNet3D (16x8)	0.298	0.361
ResNet3D + *STAN*	**0.318**	**0.366**
SlowFast [9] (32x4)	0.316	0.391
SlowFast + *STAN*	**0.326**	**0.396**

Charades-Ego

2% absolute improvement on 1st-person test; 1st-person training is scarce

Models	mAP
ResNet3D (16x8)	0.455
ResNet3D + *STAN*	**0.497**
SlowFast [9] (32x4)	0.484
SlowFast + *STAN*	**0.531**

MEVA

Summary of P1

- P1. Action Analysis
  - C2. Efficient Object Detection and Tracking
  - C3. Weakly-Supervised Action Event Recognition
  - C4. Viewpoint Invariant Representation Learning
- Summary & Contributions
  - We have presented an efficient perception system to get object tracks
  - We have tackled the problem of the lack of training data
  - We have proposed a method to learn viewpoint invariant representation
    - Better accuracy with minimal computation overhead

Focuses of This Presentation

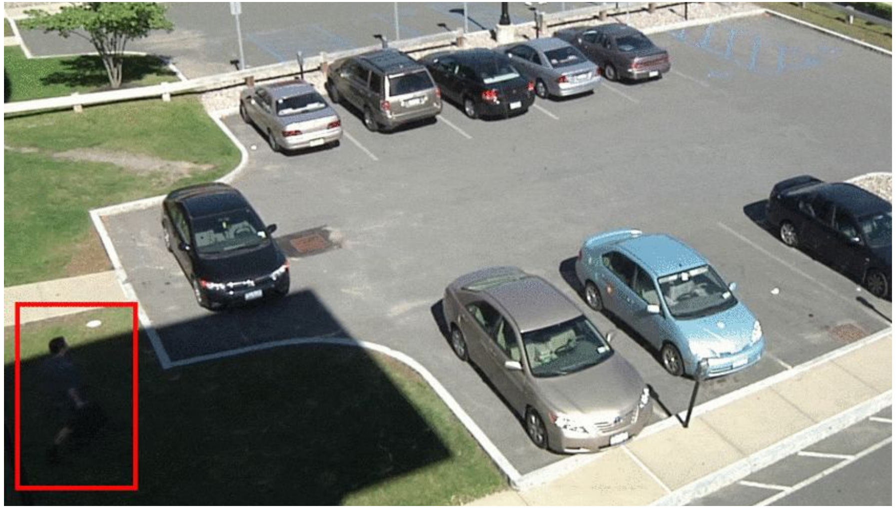
P1. Action Analysis	P2. Trajectory Prediction with Scene Semantics	P3. Analysis of Actions and Trajectory Prediction
Efficient Object Detection and Tracking (C2) Weakly-supervised Learning (C3)	**Multi-modal Future Trajectory Prediction (C5)**	Joint Action and Trajectory Prediction (C7)
Viewpoint-Invariant Representation Learning (C4)	Simulation-as-Augmentation Robust Learning (C6)	**Long-term Trajectory Prediction Using Scene Semantics and Action Representation (C8 & C9)**

Roadmap

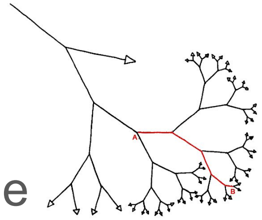
- P1. Action Analysis
- P2. Trajectory Prediction with Scene Semantics
  - C5. Multi-modal Future Trajectory Prediction
  - C6. Simulation-as-Augmentation Robust Learning
- P3. Analysis of Actions and Trajectory Prediction
- Vision and Future Directions
- Conclusions

C5. Multi-modal Future Trajectory Prediction

- Motivation
  - The future of pedestrian can be uncertain
  - As shown in this example, the person is likely to walk in multiple directions.



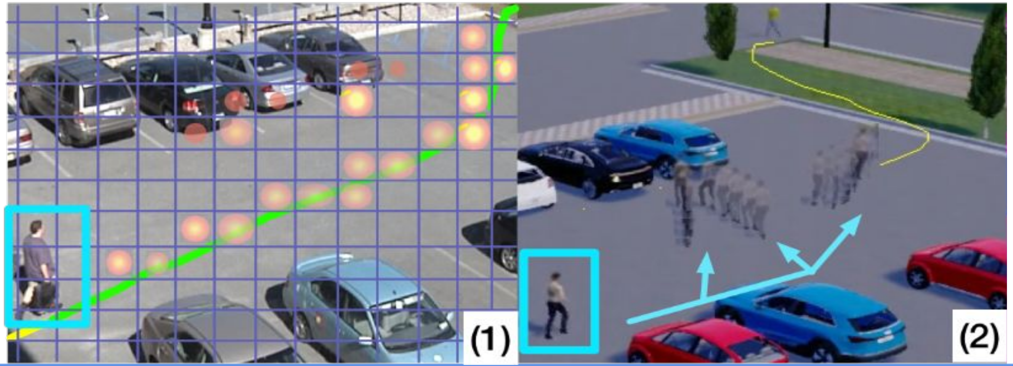
The Garden of Forking Paths
by Jorge Luis Borges



The Forking Paths Dataset

In real-world videos, only one possible trajectory is available for the same scenario.

In order to provide a quantitative evaluation of multi-future trajectory prediction, we create a trajectory dataset using a realistic simulation environment, where the agents are controlled by human annotators, to create multiple semantically plausible future paths.

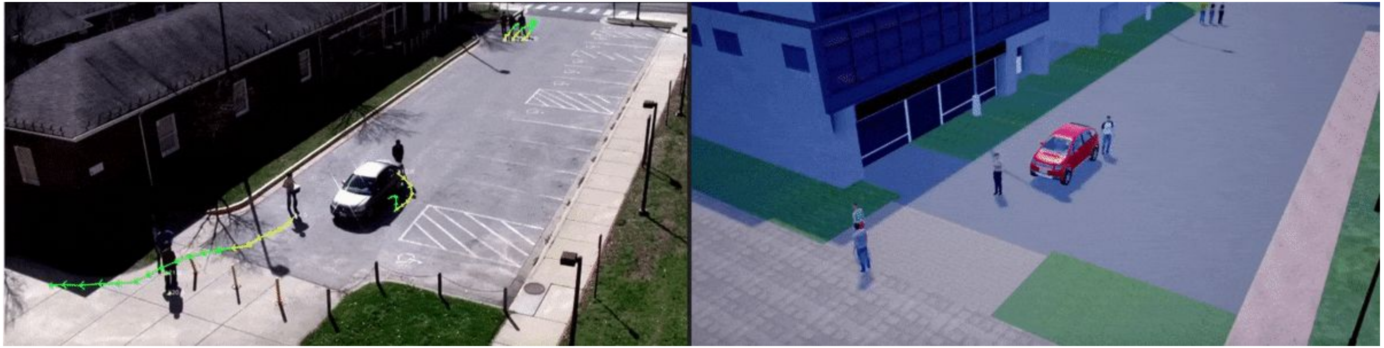


The Forking Paths Dataset

1. Scenario re-creation (~15 seconds snippet)
2. Scenario editing
3. Human annotation

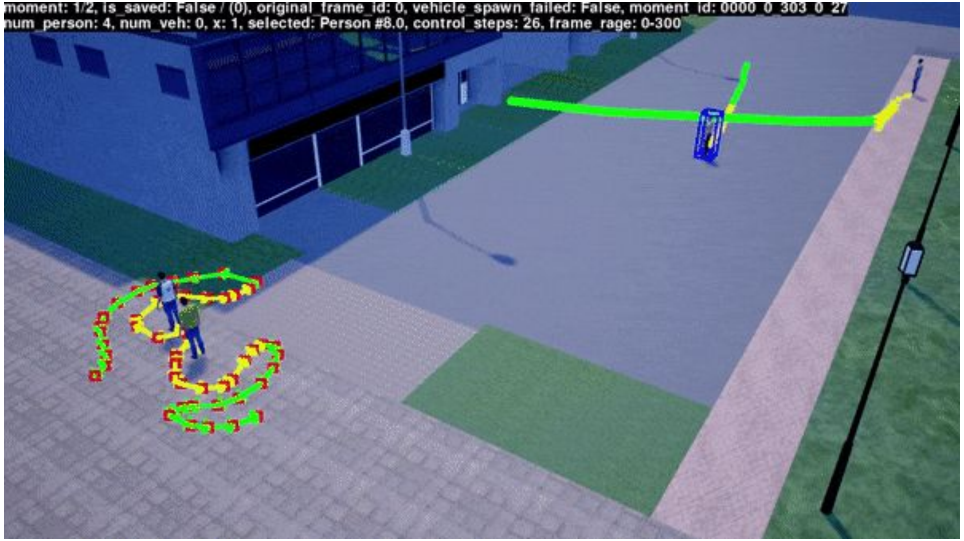
Scenario re-creation

1. Static scene reconstruction (manually through Unreal Engine 4 editor)
2. Dynamic agents (person, vehicle) reconstruction (automatically with given homography matrices)
    a. Trajectories are converted to CARLA agent control commands



Scenario Editing

- We build a GUI for scenario editing
  - Efficiently examine, add, delete person/vehicle trajectories
  - Decide which agents are plausible “multi-future” agents and their destinations



Human Annotation

- 10 annotators control the agent to reach destinations within 15 seconds and without collisions

The Forking Paths Dataset - Multi-Future Trajectory Visualization

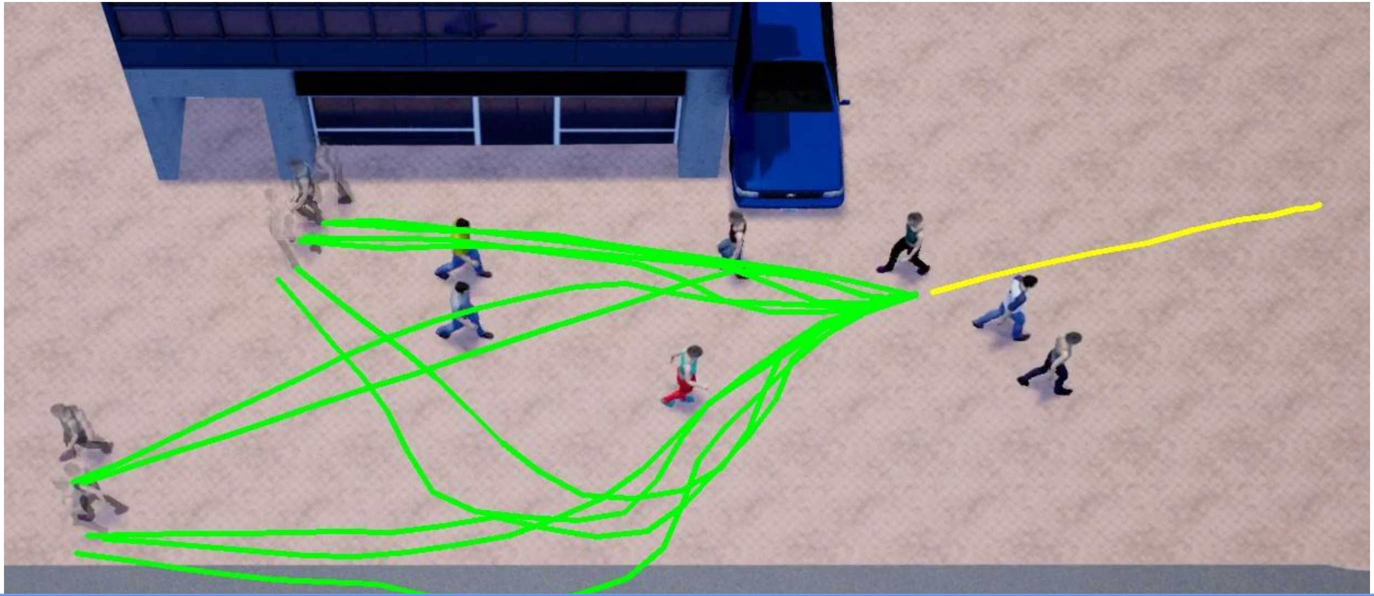
Single View Demonstration - Dataset

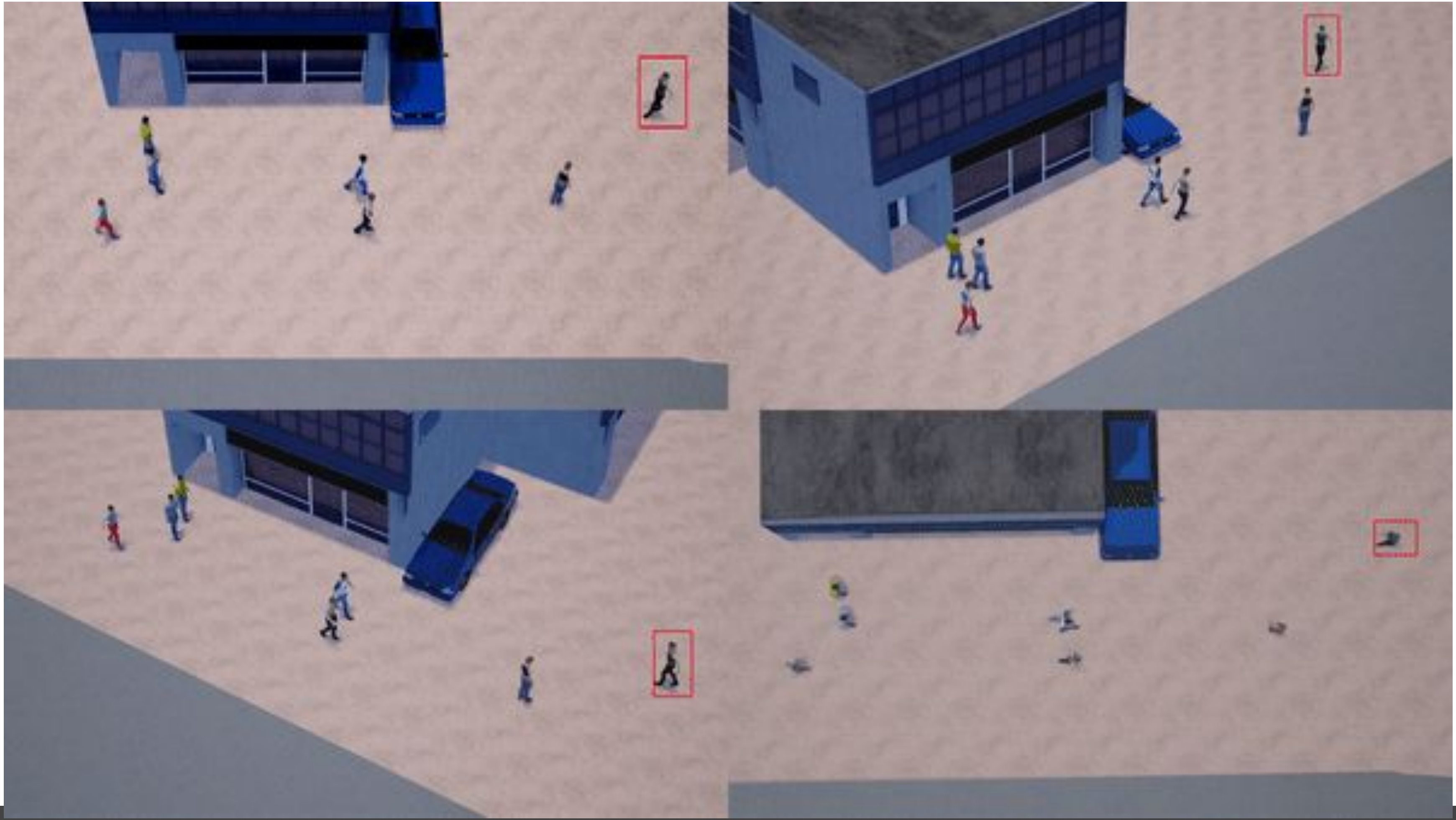
Red bounding box $\boxed{}$: Human-controlled agent

Single View Demonstration - Dataset

Yellow trajectory: Agent past trajectory during observation

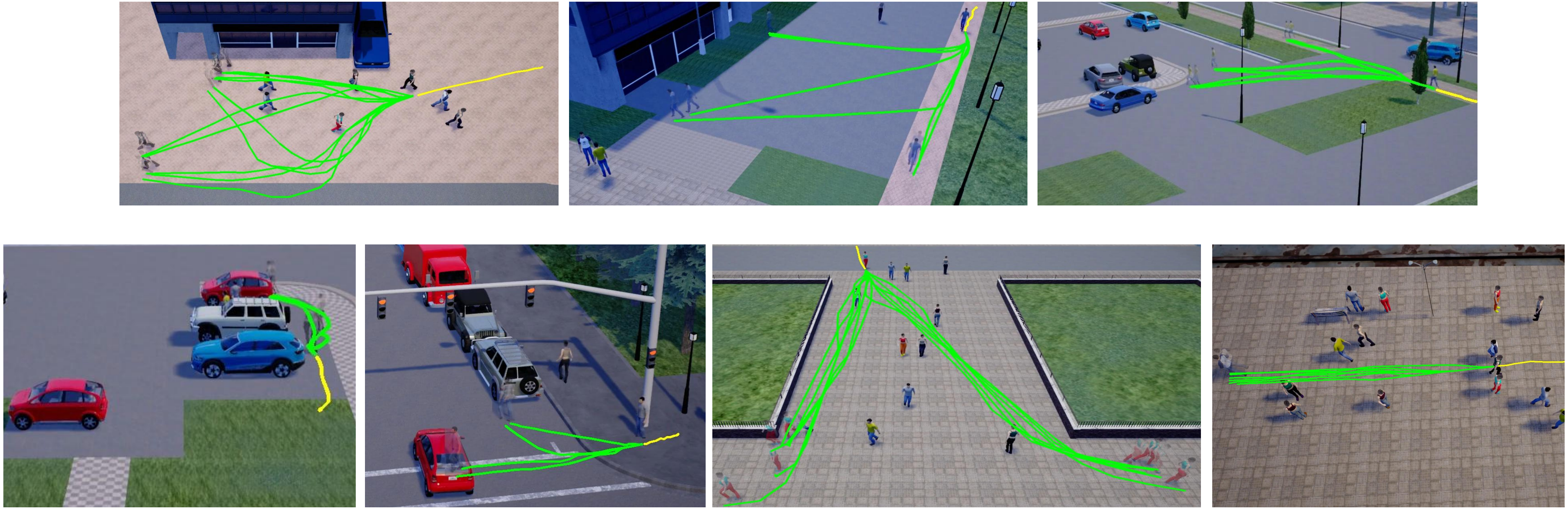
Green trajectories: Agent future trajectories from different human annotators





Single View Demonstration - Dataset

We have collected multi-future trajectories from 7 scenes.

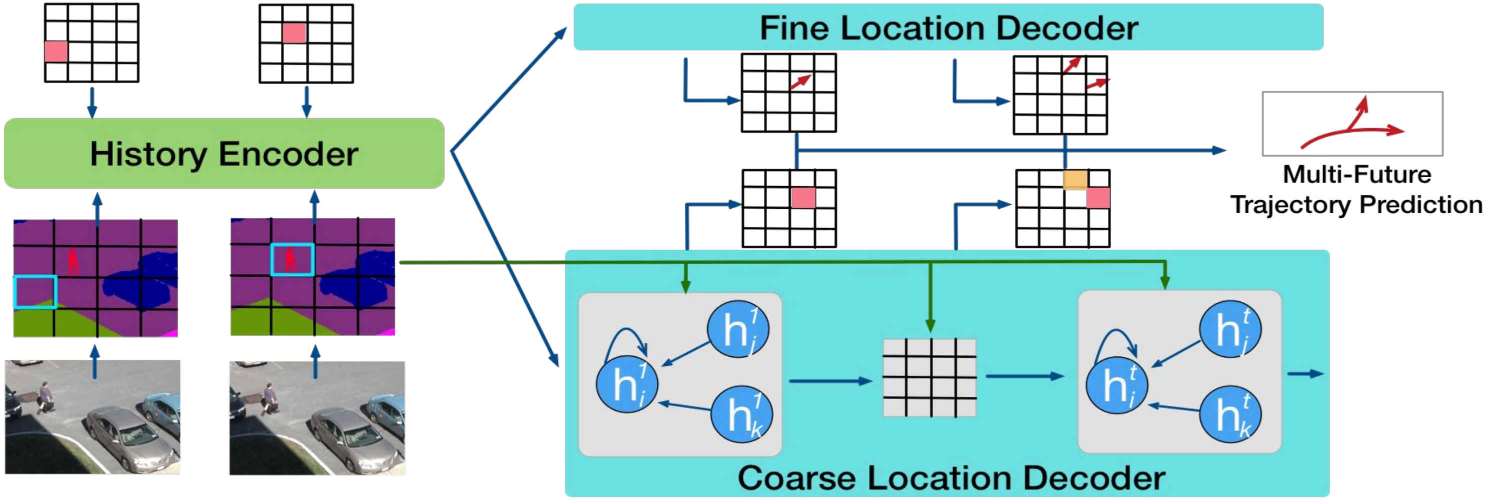


Single View Demonstration - Vehicle Scene

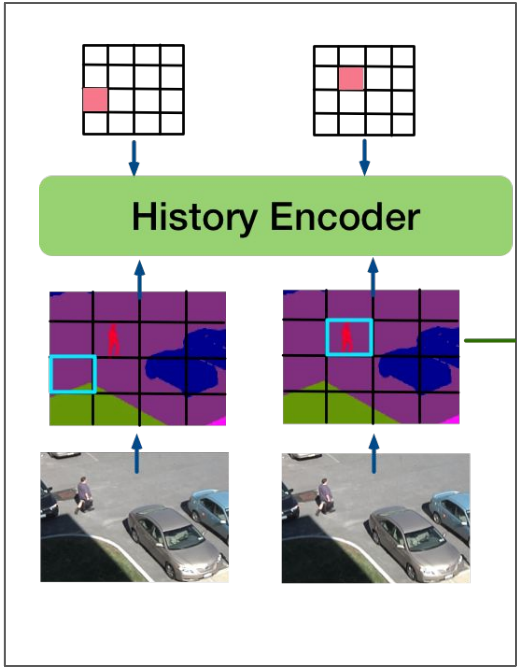
Red bounding box $\boxed{}$: Human-controlled agent

The Multiverse Model

We propose multi-decoder framework that predicts both coarse and fine locations of the person using scene semantic segmentation features.



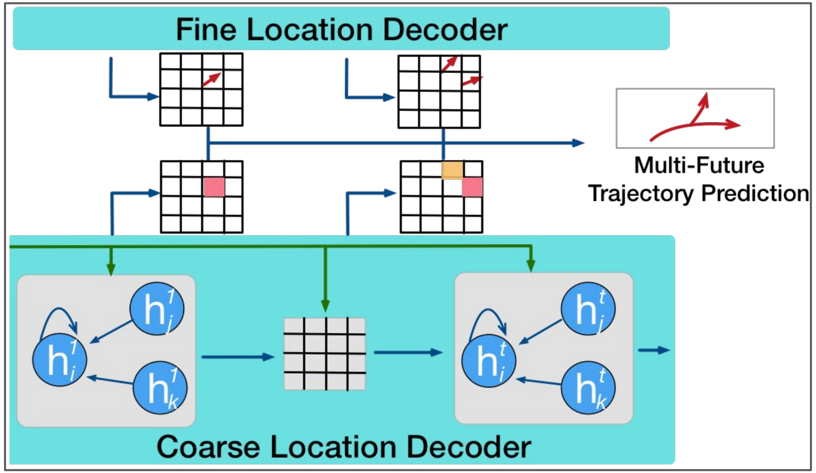
Our Model - Encoder



- Divide the scene into grids
- Multi-level History Encoder (1 ... T)
  - Pretrained scene semantic segmentation features
  - Kernel=3 convolution masked based on the person's location
  - Input into a Convolutional LSTM (tf.contrib.rnn.ConvLSTMCell)

Our Model - Decoder

- Multi-level Decoder (T+1 ... T_{pred})
  - Two levels
    - Coarse Location Decoder
    - Fine Location Decoder
  - ConvLSTM
  - At each timestep, we use graph convolution to refine the hidden states
    - Edge weights: based on neighboring scene semantics and the hidden states
  - During inferencing, use beam search for the coarse location decoder to get multiple future
  - Combining two-level outputs to get final trajectory predictions



Experiments - Evaluation Metrics

- Minimum Average/Final Displacement Error Given K Predictions (Geometric)

$$\mathrm{minADE}_K = \frac{\sum_{i=1}^{N} \sum_{j=1}^{J} \min_{k=1}^{K} \sum_{t=h+1}^{T} \|Y_t^{ij} - \hat{Y}_t^{ik}\|_2}{N \times (T-h) \times J}$$

- minADE_{20}: Minimum average error given 20 model predictions
  - 20 model predictions are compared to the ground truth at test time, and only the lowest error ones are selected to count

Experiment - Multi-Future Trajectory Prediction

Our model outperforms others on the proposed dataset for multi-future trajectory prediction. We repeat all experiments (except "linear") 5 times.

Method	Input Types	minADE_{20}		minFDE_{20}	
		45-degree	top-down	45-degree	top-down
Linear	Traj.	213.2	197.6	403.2	372.9
LSTM	Traj.	201.0 ±2.2	183.7 ±2.1	381.5 ±3.2	355.0 ±3.6
Social-LSTM [1]	Traj.	197.5 ±2.5	180.4 ±1.0	377.0 ±3.6	350.3 ±2.3
Social-GAN (PV) [14]	Traj.	191.2 ±5.4	176.5 ±5.2	351.9 ±11.4	335.0 ±9.4
Social-GAN (V) [14]	Traj.	187.1 ±4.7	172.7 ±3.9	342.1 ±10.2	326.7 ±7.7
Next [27]	Traj.+Bbox+RGB+Seg.	186.6 ±2.7	166.9 ±2.2	360.0 ±7.2	326.6 ±5.0
Ours	Traj.+Seg.	**168.9** ±2.1	**157.7** ±2.5	**333.8** ±3.7	**316.5** ±3.4

Numbers are displacement errors. Lower the better.

Experiment - Multi-Future Trajectory Prediction

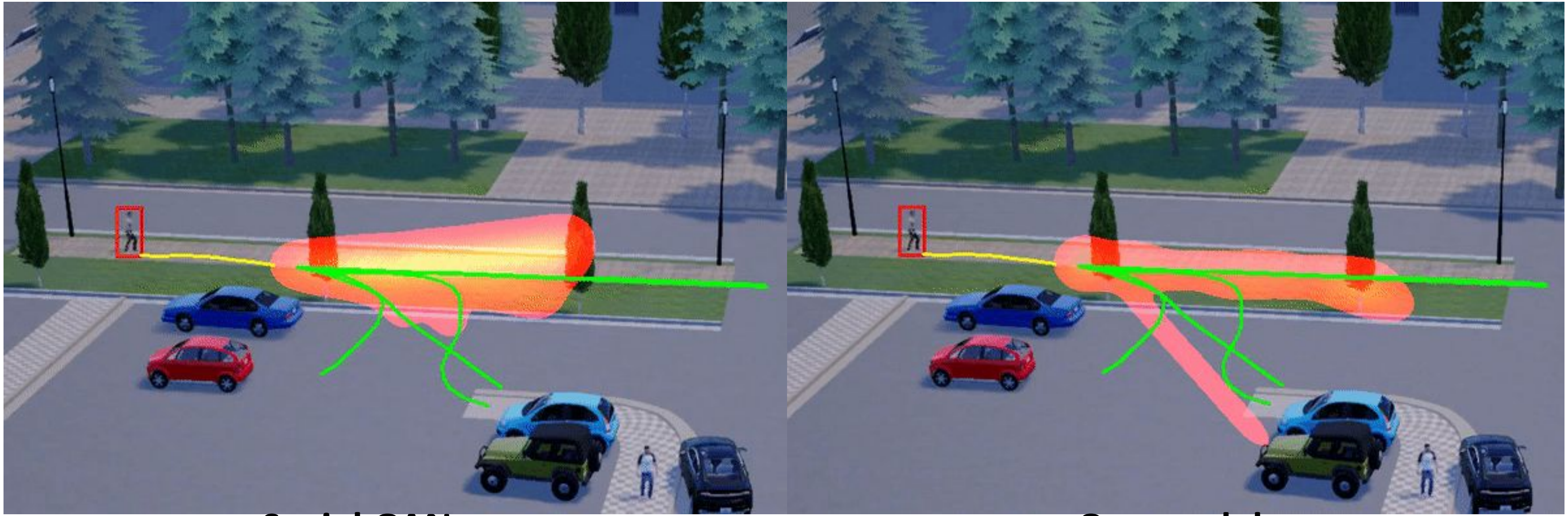
Our model outperforms others on the proposed dataset for multi-future trajectory prediction. We repeat all experiments (except "linear") 5 times.

10% less average errors than Social-GAN

Method	Input Types	minADE_{20}			
		45-degree	top-down	45-degree	top-down
Linear	Traj.	213.2	197.6	403.2	372.9
LSTM	Traj.	201.0 ±2.2	183.7 ±2.1	381.5 ±3.2	355.0 ±3.6
Social-LSTM [1]	Traj.	197.5 ±2.5	180.4 ±1.0	377.0 ±3.6	350.3 ±2.3
Social-GAN (PV) [14]	Traj.	191.2 ±5.4	176.5 ±5.2	351.9 ±11.4	335.0 ±9.4
Social-GAN (V) [14]	Traj.	187.1 ±4.7	172.7 ±3.9	342.1 ±10.2	326.7 ±7.7
Next [27]	Traj.+Bbox+RGB+Seg.	186.6 ±2.7	166.9 ±2.2	360.0 ±7.2	326.6 ±5.0
Ours	Traj.+Seg.	**168.9** ±2.1	**157.7** ±2.5	**333.8** ±3.7	**316.5** ±3.4

Numbers are displacement errors. Lower the better.

Qualitative Comparison



Social GAN

Our model

Redbox: Human-controlled agent

Green trajectories: trajectories during prediction period

Yellow trajectory: trajectory during observation period

Yellow-orange heatmap: Multi-future model predictions

C5. Multi-modal Future Trajectory Prediction - Contributions

- Introduced the first dataset that allows us to compare models in a quantitative way in terms of their ability to predict multiple plausible futures.
- Proposed a new effective model for multi-future trajectory prediction.
- Established a new state-of-the-art result on the challenging VIRAT/ActEV benchmark, and compared various methods on our multi-future trajectory prediction datasets.

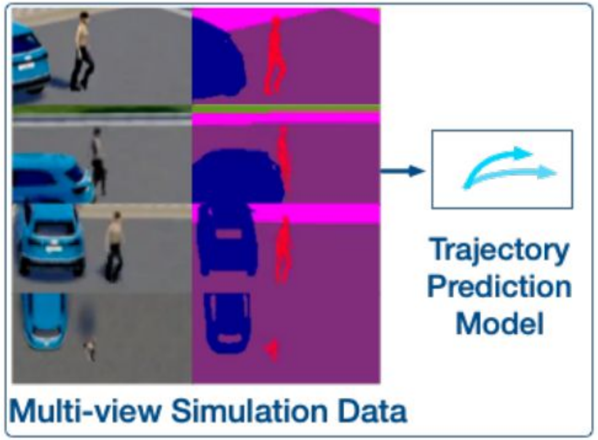
Roadmap

- P1. Action Analysis
- P2. Trajectory Prediction with Scene Semantics
  - C5. Multi-modal Future Trajectory Prediction
  - C6. Simulation-as-Augmentation Robust Learning
- P3. Analysis of Actions and Trajectory Prediction
- Vision and Future Directions
- Conclusions

C6. Learning from 3D Simulation for Trajectory Prediction

In this chapter, we study the problem of trajectory prediction in unseen cameras.

We propose a method, SimAug, to train robust models using simulation data that could generalize to unseen camera viewpoints and scenes (see below).



Stanford Drone Argoverse

Time mark: 22:30

Summary of P2

- P2. Trajectory Prediction with Scene Semantics
  - C5. Multi-modal Future Trajectory Prediction
  - C6. Simulation-as-Augmentation Robust Learning
- Summary & Contributions
  - In this part, we study trajectory prediction models with scene semantic cues
  - We study multimodal future prediction and propose the first manually-annotated quantitative benchmark
  - We also develop a robust learning method for better generalization of prediction model using 3D simulation

Focuses of This Presentation

P1. Action Analysis	P2. Trajectory Prediction with Scene Semantics	P3. Analysis of Actions and Trajectory Prediction
Efficient Object Detection and Tracking (C2) Weakly-supervised Learning (C3)	**Multi-modal Future Trajectory Prediction (C5)**	Joint Action and Trajectory Prediction (C7)
Viewpoint-Invariant Representation Learning (C4)	Simulation-as-Augmentation Robust Learning (C6)	**Long-term Trajectory Prediction Using Scene Semantics and Action Representation (C8 & C9)**

Roadmap

- P1. Action Analysis
- P2. Trajectory Prediction with Scene Semantics
- P3. Analysis of Actions and Trajectory Prediction
    - C7. Joint Action and Trajectory Prediction
    - C8 & C9. Long-term Trajectory Prediction Using Scene Semantics and Action Representation
- Vision and Future Directions
- Conclusions

C7. Joint Action and Trajectory Prediction

In this chapter, our goal is to jointly predict a person's future trajectory and action on common benchmarks (short-term prediction)

Intuition

- People navigate in the scene with a specific purpose in mind.
- People's purpose can be inferred from **their appearance, body language** as well as nearby **environment**.

Our Model - Next

1. We design a *Person Behavior Module* and *Person Interaction Module* to model the target person as well as their interaction with the scene and other objects.
2. We utilize multi-task learning for joint trajectory and action prediction

Experiments

Single Future: only 1 prediction allowed
Multi-Future: 20 model outputs; Find the best one using ground truth

Setup:

- Predict 4.8 seconds

Baselines:

1. Linear Regressor
2. LSTM
3. Social LSTM
4. Social GAN
5. **Social GAN + Scene** (SoPhie)

	Method	AVG
Single Future	Linear	0.79 / 1.59
	LSTM	0.70 / 1.52
	Social LSTM	0.72 / 1.54
	Ours-single-model	**0.52 / 1.14**
Multi-Future	Social GAN (P)	0.58 / 1.18
	Social GAN (PV)	0.61 / 1.21
	SoPhie	0.54 / 1.15
	Ours-20	**0.46 / 1.00**

Table 2. ETH & UCY Experiment

Single output is better than SoPhie with 20 outputs

Metrics:

- Single Future: $minADE_1$ / $minFDE_1$
- Multi-Future: $minADE_{20}$ / $minFDE_{20}$

C7. Joint Action and Trajectory Prediction - Contributions

- We have presented the first model that predict human trajectory and future activity simultaneously
- We are one of the early works that utilize rich visual features including person appearance, person keypoints and scene semantics for short-term trajectory prediction
- We achieve SOTA performance on ETH/UCY dataset

Roadmap

- P1. Action Analysis
- P2. Trajectory Prediction with Scene Semantics
- P3. Analysis of Actions and Trajectory Prediction
  - C7. Joint Action and Trajectory Prediction
  - C8 & C9. Long-term Trajectory Prediction Using Scene Semantics and Action Representation
- Vision and Future Directions
- Conclusions

C8. Long-term Trajectory Prediction Using Scene Semantics and Action Representation

- We propose a new long-term trajectory prediction dataset with multi-viewpoint video data and a new model that incorporates action representations and scene understanding
  - Short-term: predict ~5 seconds (8 time-steps), long-term*: predict 12 seconds (30 time-steps)
- Why long-term?
  - Short-term future prediction is not enough to ensure safe operations
- Motivation of collecting a new dataset
  - Common trajectory benchmark's (ETH/UCY/SDD) trajectory length is short in general
  - They also lack action annotation and multi-viewpoint video data in **traffic scenes**

* the "long-term" definition is consistent with recent published work [99, 185, 224]

A Multi-view Long-term Trajectory Prediction Benchmark

- We utilize the MEVA dataset
  - Activity annotation is provided without full person/vehicle tracks
  - We need to run object tracking across cameras to get them

A Multi-view Long-term Trajectory Prediction Benchmark

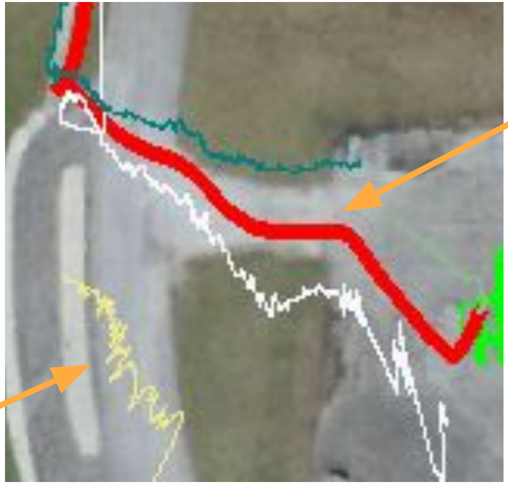
- The MEVA-Trajectory Dataset
  - Human annotation - rejecting wrong global tracks
    - Automatic global track: 2549, annotated down to 864
  - Please refer to the thesis write-up for details of dataset collection process

Accepted global track

Rejected global track (many ID switches)

A Multi-view Long-term Trajectory Prediction Benchmark

- The MEVA-Trajectory Dataset
  - Trajectory smoothing with moving averages
  - Please refer to the thesis write-up for details of dataset collection process



Smoothed global trajectory

Rejected local trajectory

(Track length 2:50)

A Multi-view Long-term Trajectory Prediction Benchmark

- The MEVA-Trajectory Dataset
  - Comparison with common benchmarks

Datasets	ETH,UCY [118, 162]	SDD [183]	KITTI [59]	ActEV [158]	Ours
HD Resolution	-	-	✓	partial	✓
Multi-View	-	-	-	-	✓
Extended Length	-	-	✓	✓	✓
Event/Goal-Driven	-	-	-	partial	✓
Traffic Scene	-	partial	✓	✓	✓
Activity Annotation	-	-	-	✓	✓

A Multi-view Long-term Trajectory Prediction Benchmark

- The MEVA-Trajectory Dataset
  - Comparison with common benchmarks

	ETH, UCY	ActEV	Ours
#Cameras	4	5	10
Total Traj. Length	4:59:05	12:14:44	15:36:17
#Traj.	1535	1073	2060 / 864*
Median Traj. Length	8.8	28.8	48.3
Median #Camera	1	1	2
Annotations	Person coordinates	Person+object bounding boxes,activities	Person+object bounding boxes,activities

A Multi-view Long-term Trajectory Prediction Benchmark

- The MEVA-Trajectory Dataset
  - Visualization of the facility

Camera view

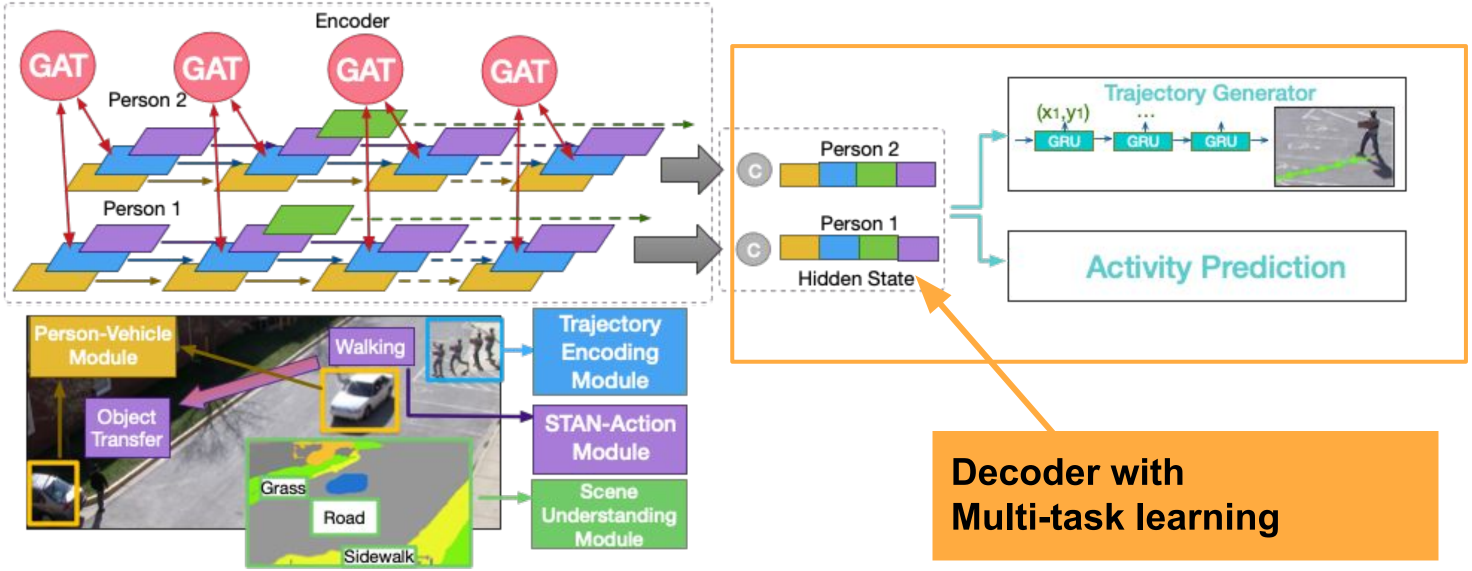
Top-down view

C9. Long-term Trajectory Prediction with Scene and Action Understanding

- Goal
  - Expand the common trajectory prediction horizon into long-term setting
    - Predict 12 seconds into the future (previously is ~5 seconds)
  - With the aid of graph attention, scene semantic understanding and action analysis representations

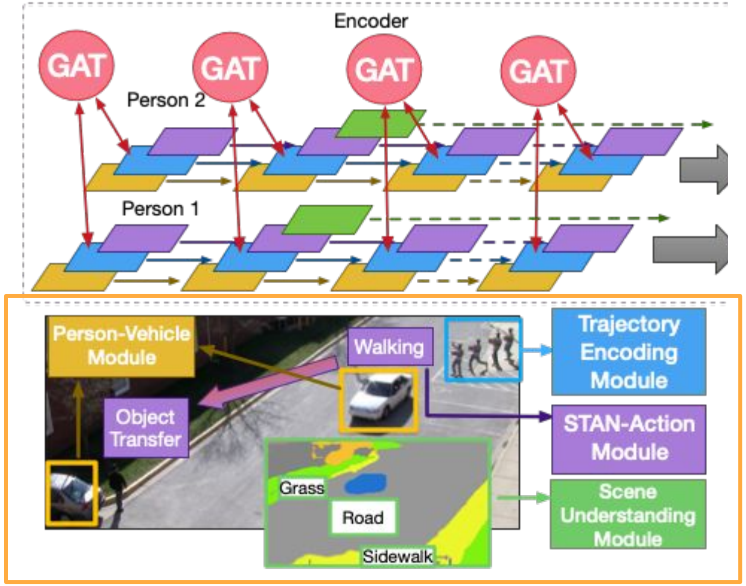
The Next-GAT Model

- We utilize enhanced contextual understanding for trajectory and activity prediction



The Next-GAT Model

- We utilize enhanced contextual understanding for trajectory and activity prediction



***Trajectory Encoding with GAT**
***STAN-Action**
Scene Understanding
Person-Vehicle

Experiments

- Previous Work
  - Social-GAN: Representative earlier work on multimodal trajectory prediction
  - ST-GAT: Representative method using graph attention network
  - STGCNN: Recent highly-cited method using convolution network
  - Next (Chapter 7)
- Datasets
  - ActEV and MEVA-Trajectory
- Tasks
  - Short-term and Long-term
  - Single-Future: One model output and use $minADE_1$ / $minFDE_1$ as metrics
  - Multi-Future: 20 model output and use $minADE_{20}$ / $minFDE_{20}$ as metrics

Results - ActEV

- We compare with representative recent methods
    - Significant improvement especially on long-term prediction

NN: Nearest Neighbor

	Short-term Trajectory Prediction			Long-term Trajectory Prediction		
	Act	Single-Future	Multi-Future	Act	Single-Future	Multi-Future
NN	-	1.79/3.12	-	-	3.47/6.5	-
Const. Vel.	-	1.17/2.25	-	-	2.78/5.74	-
SGAN	-	1.21/2.25	0.88/1.63	-	3.37/6.66	2.69/5.29
STGAT	-	1.43/2.75	0.88/1.68	-	4.05/7.78	2.27/4.63
STGCNN	-	1.48/2.57	1.08/1.93	-	3.46/6.51	2.78/5.46
Next	0.192	1.06/2.03	0.87/1.79	0.211	2.22/4.56	1.97/4.05
Next-GAT	**0.236**	**0.84/1.57**	**0.76/1.42**	**0.267**	**1.94/4.05**	**1.63/3.36**

Constant Velocity already good

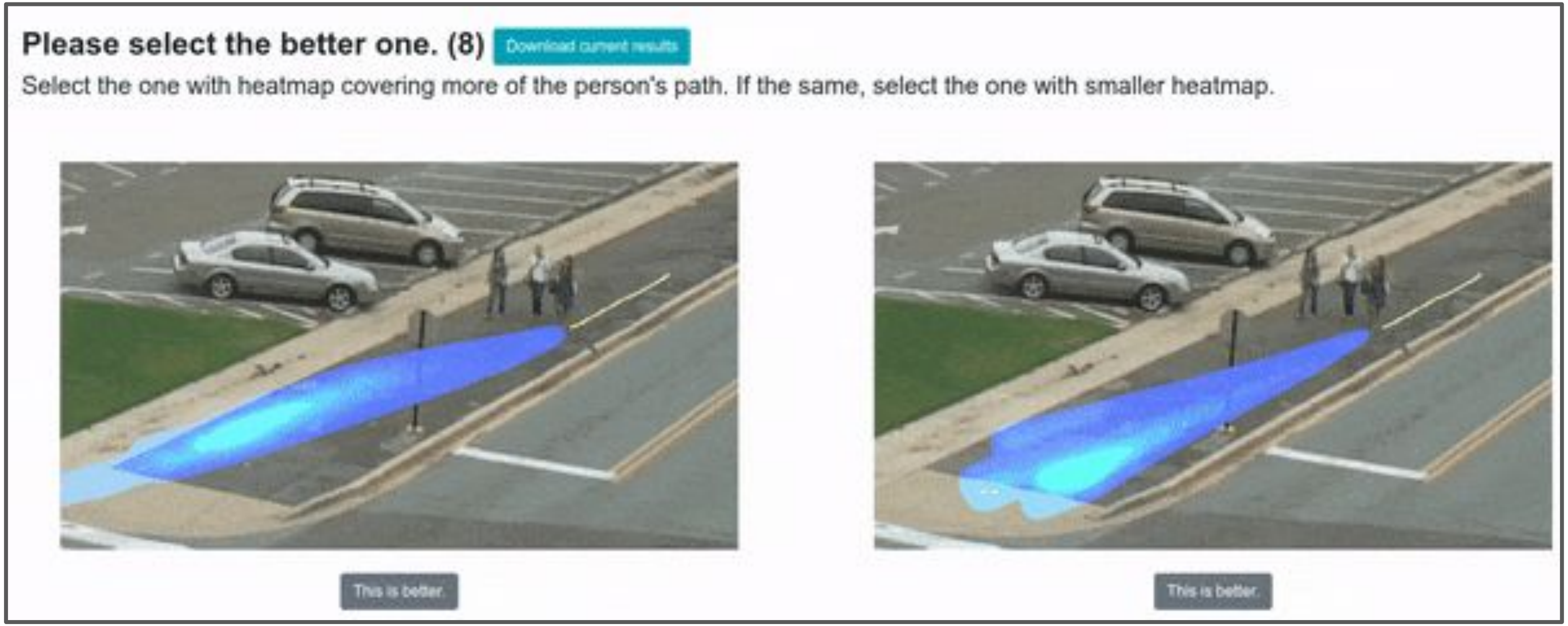
STGCNN better at single but worse at multi

Ours 28% better than STGAT

The numbers are in meters (except mAP)

Results - ActEV

- Human interpretation of the error gap
    - We conduct a user study with randomized paired example comparison



Results - ActEV

- Human interpretation of the error gap
  - We conduct a user study

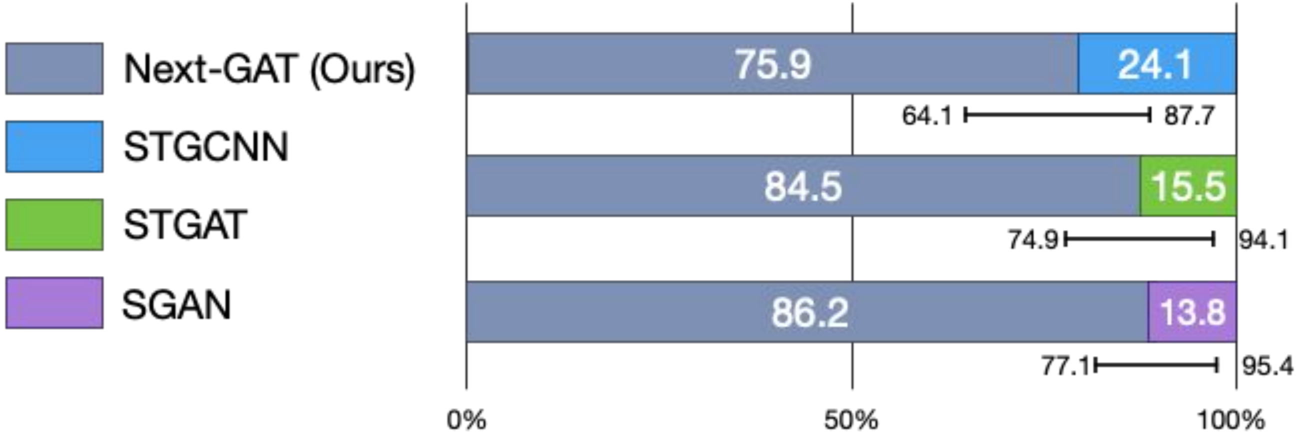
ADE

1.63

2.78

2.27

2.69



Qualitative Analysis

SGAN

STGAT

STGCNN

Ours

Error Analysis

SGAN

STGAT

Sudden right turn

STGCNN

Ours

Results - MEVA-Trajectory

- We compare with representative recent methods
    - Significant improvement especially on long-term prediction

STGCNN a lot worse than ActEV

	Short-term Trajectory Prediction			long-term Trajectory Prediction		
	Act	Single-Future	Multi-Future	Act	Single-Future	Multi-Future
NN	-	7.32/13.54	-	-	15.29/30.00	-
Const. Vel.	-	2.76/5.76	-	-	8.35/17.89	-
SGAN	-	3.41/7.21	1.92/4.04	-	8.77/18.11	7.24/14.98
STGAT	-	5.05/10.43	2.00/4.15	-	14.75/29.51	7.71/15.68
STGCNN	-	4.79/8.56	3.36/6.33	-	14.60/27.42	11.54/22.63
Next	0.257	2.14/5.04	1.95/4.55	0.176	7.62/18.20	6.98/16.60
Next-GAT	**0.328**	**1.91/4.33**	**1.63/3.75**	**0.299**	**6.51/14.67**	**5.60/12.82**

Ours' single output is better than baselines' 20 outputs

Action prediction is significantly better

The numbers are in feet (except mAP)

Results - MEVA-Trajectory

- Ablation study
  - Single Trajectory

	long-term Trajectory Prediction		
	Activity	minADE_1	minFDE_1
Next-GAT	0.299	6.51	14.67
Next	0.176	7.62	18.2
Next-GAT-ResNet	0.253	7.02	15.55
Next-GAT-noScene	0.280	6.88	15.78
GRU-EncodeDecode	-	9.69	20.97

Graph attention is important

STAN-Action improves activity prediction

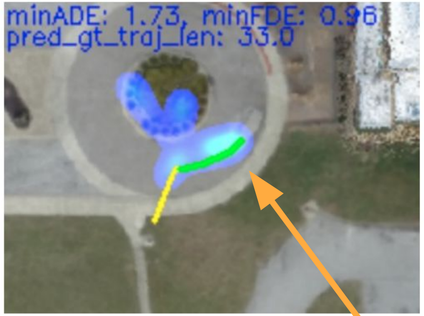
Scene semantic segmentation helps a bit

Visual feature is crucial

Results - MEVA-Trajectory

- Qualitative analysis

SGAN



STGAT

STGCNN

Ours

Video frames (two cameras)

Predicted correct turn

Time mark: 40:00

Summary of P3

- P3. Analysis of Actions and Trajectory Prediction
  - C7. Joint Action and Trajectory Prediction
  - C8 & C9. Long-term Trajectory Prediction using scene semantics and action representation
- Summary & Contributions
  - In this part, we focus on joint modeling methods and develop a trajectory and action prediction model that takes into account contextual cues of both the target agent's behavior cues and scene semantics
  - We propose a new multi-view long-term trajectory prediction benchmark in traffic scenes, MEVA-Trajectory
  - We achieve state-of-the-art performance on MEVA-Trajectory

Roadmap

- P1. Action Analysis
- P2. Trajectory Prediction with Scene Semantics
- P3. Analysis of Actions and Trajectory Prediction
- Vision and Future Directions
- Conclusions

Vision and Future Directions

- Applications (Short-term Directions)
  - First-person view prediction
  - Long-tail action/trajectory prediction
    - Accidents, disaster events
  - Computation-accuracy trade-off
  - Trajectory prediction in sports
  - **Crowd dynamics estimation for public safety monitoring**

Vision and Future Directions

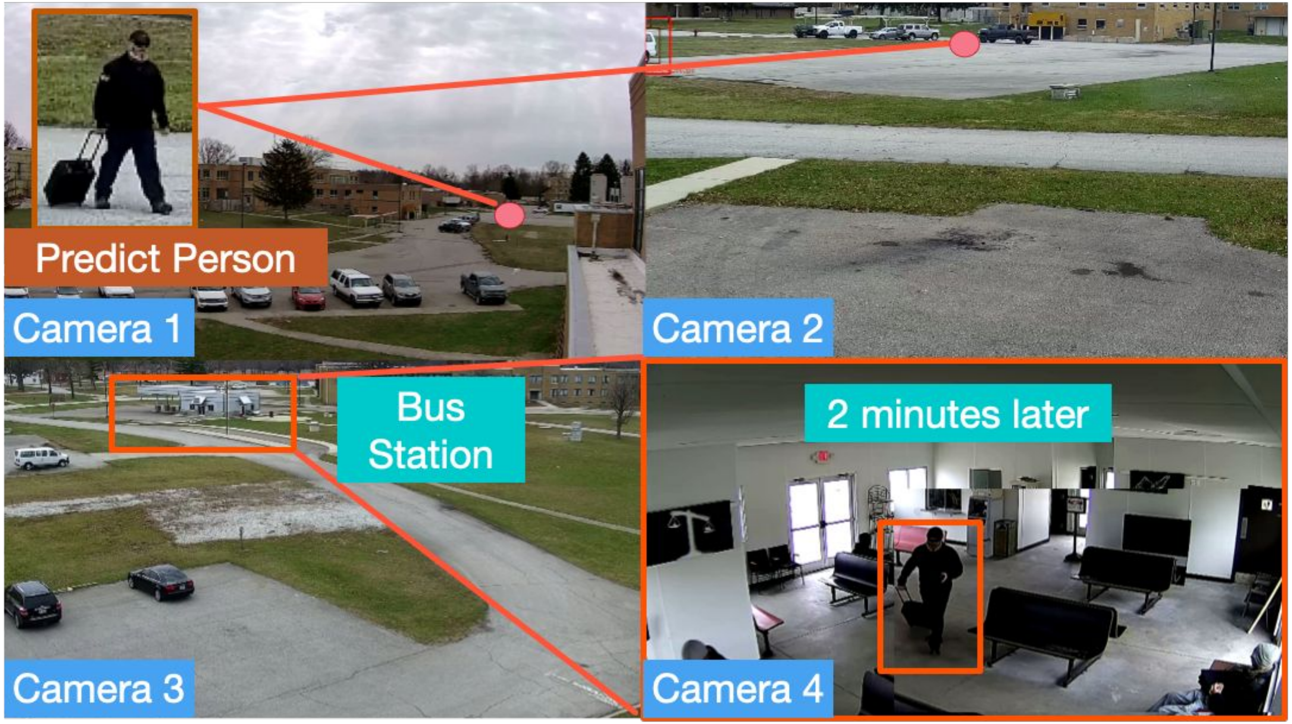
- Crowd Dynamics Estimation for Public Safety Monitoring
  - Crowd counting for the Washington Post leads to a front-page news
  - Future prediction of crowd dynamics could avoid mass casualty events

Vision and Future Directions

- Model & Algorithm (Long-term Directions)
  - Modeling different populations
  - Unifying vehicle trajectory prediction and pedestrian prediction
  - **Common sense reasoning for long-term future prediction**

Vision and Future Directions

- Common sense reasoning for long-term future prediction
  - A person with a luggage is likely to travel -> bus station is for travelers



Conclusion

- Key Research Question
  - How to build a robust trajectory prediction system with enhanced semantic context understanding for urban traffic scenes
- Tackled Three Tasks
  - P1. Action Analysis
  - P2. Trajectory Prediction with Scene Semantics
  - P3. Analysis of Actions and Trajectory Prediction
- Proposed Two New Datasets
  - The Forking Path Dataset: the first multimodal human-annotated benchmark
  - The MEVA-Trajectory Dataset: a multi-viewpoint long-term trajectory benchmark

Academic Impact

- Chapter 7 of our work has received 140+ citations and it is one of the top-cited paper at CVPR’19 on this topic. Notably, researchers have extended our work on:
    - Multi-task learning for trajectory prediction [15, 174]
    - Action prediction [28, 108]
    - Ego-centric view trajectory prediction [19, 165, 172]
    - Efficiency [231, 239]
    - Graph models [28, 211, 251]
- Chapter 5’s new dataset has been used by [144, 169] and more
- Most of our research work has been open-sourced and our Github repositories have a total of 800+ stars and 300+ forks as of June 2021.

Thank you

- Projects: https://www.cs.cmu.edu/~junweil/#projects
- Code: https://github.com/JunweiLiang
- Youtube: https://www.youtube.com/channel/UC-z7ZWp8Rbu2xhxnbAL_bRQ
- 知乎: https://www.zhihu.com/people/junwei-liang-50
- Blog: https://medium.com/@junweil
- Email: junweil@cs.cmu.edu
- Thanks to:
  - Alex, Lu, Kris, Alan
  - Sponsors: NSF, NIST, IARPA, Yahoo!, Google Cloud, Baidu Scholarship
  - Admin: Stacey Young
  - Mentors & Collaborators:
    - Liangliang Cao, Xuehan Xiong, Ting Yu, Kevin Murphy, Juan Carlos Niebles, Fei-Fei Li, Jia Li

My Journey So Far...

Aug. 2015

June. 2021

ICMR 2017
June 6-9, Bucharest, Romania

IJCAI-16

ICASSP 2017

AAAI-17 SAN FRANCISCO

2018 CVPR
SALT LAKE CITY • JUNE 18-22

ICMR 2018

ACMmultimedia
Nice, France 21-25 October 2019

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE

CVPR LONG BEACH CALIFORNIA
June 16-20, 2019

CVPR SEATTLE WASHINGTON
JUNE 16-18 2020

ECCV'20 ONLINE
23-28 AUGUST 2020

Reference

Liang, Junwei, et al. "SimAug: Learning Robust Representations from Simulation for Trajectory Prediction." ECCV 2020.

Liang, Junwei, et al. "The garden of forking paths: Towards multi-future trajectory prediction." CVPR 2020.

Liang, Junwei, et al. "Peeking into the future: Predicting future person activities and locations in videos." CVPR 2019.

Liang, Junwei, et al. “Leveraging Multi-modal Prior Knowledge for Large-scale Concept Learning in Noisy Web Data.” ICMR 2017

Liang, Junwei, et al. "Focal visual-text attention for memex question answering." TPAMI 2019.

Liang, Junwei, et al. "Focal visual-text attention for visual question answering." CVPR 2018.

Liang, Junwei, et al. "Learning to Detect Concepts from Webly-Labeled Video Data." IJCAI 2016.

Liang, Junwei, et al. "Webly-supervised learning of multimodal video detectors." AAAI 2017.