電子情報学専攻 修士論文審査会

Future Person Localization in First-Person Videos

(一人称視点映像における人物位置予測)

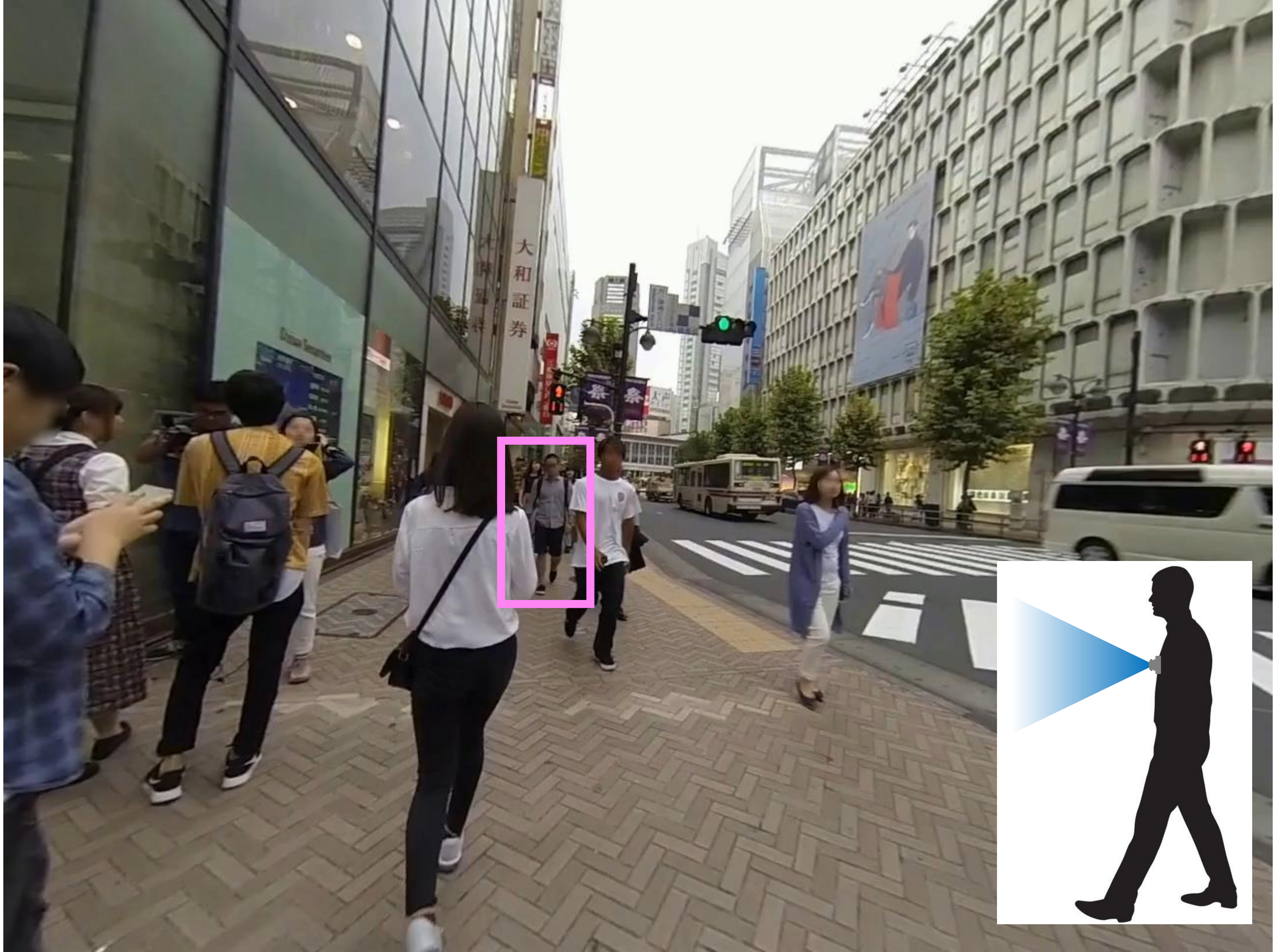
19/02/04

佐藤洋一研究室

八木 拓真 (Takuma Yagi)

First-person vision

- Use body-worn wearable cameras
- Analyze videos which reflect wearer's action and interest



大和証券


Daiwa Securities
大和証券
?


Daiwa Securities
大和証券

Future person localization in third-person videos

(1) The use of appearance feature

- Learns preference of walkable area [Kitani+, ECCV'12]
- The use of holistic visual attributes [Ma+, CVPR'17]

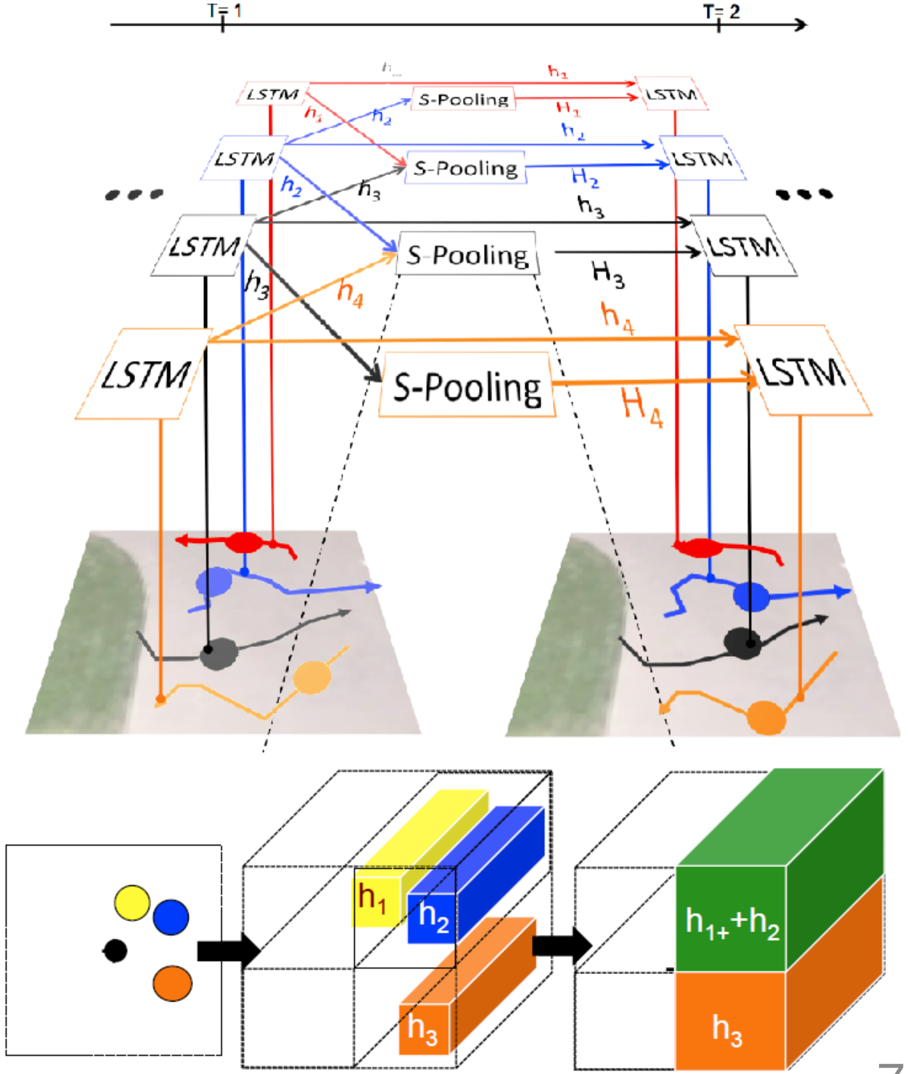
(2) The use of interaction between people

- Computer simulation (Social force) [Helbing+, '95]
- Data-driven approach (Social LSTM) [Alahi+, CVPR'16]

Future person localization in third-person videos

Social LSTM [Alahi+, CVPR'16]

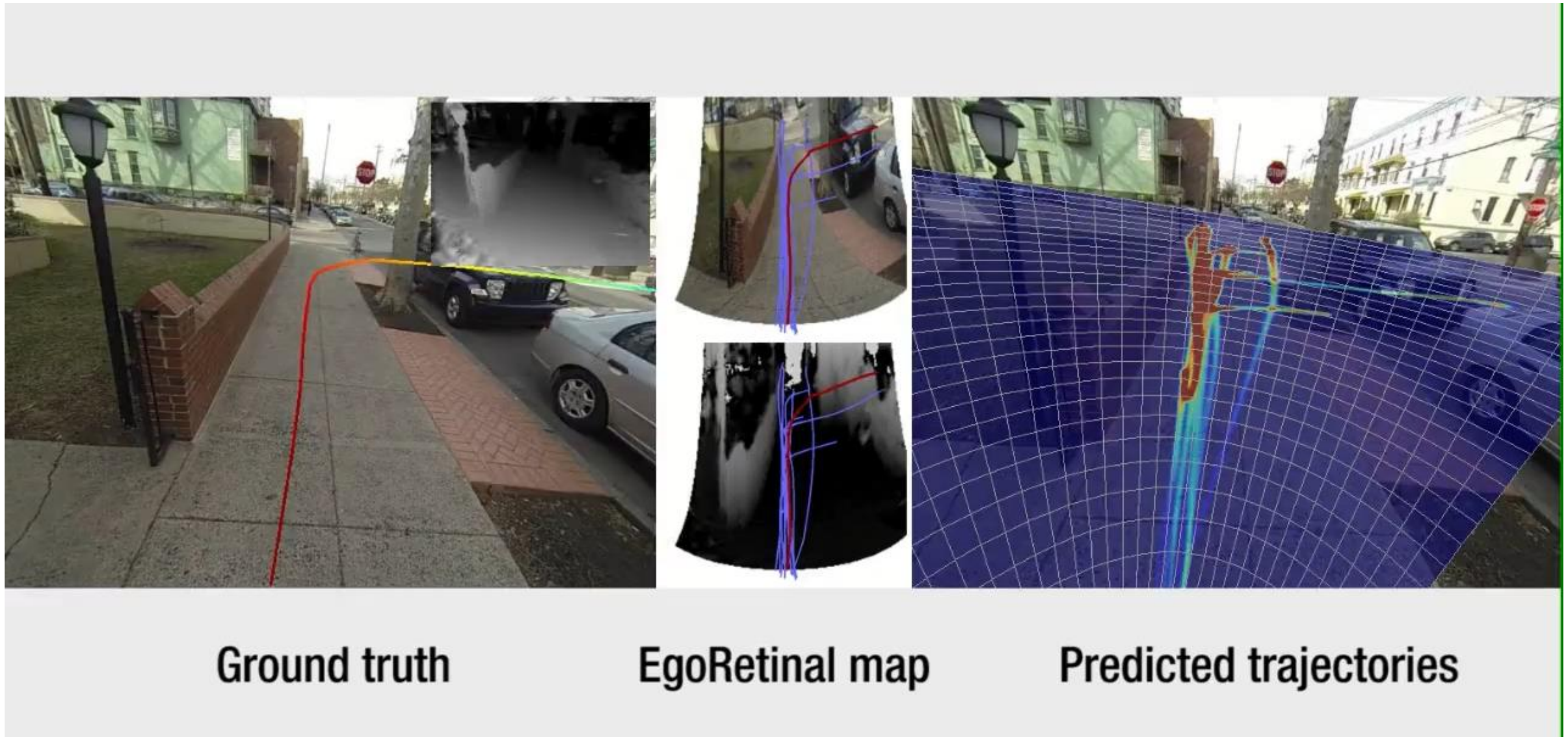
- Model each pedestrian by a LSTM
- Social pooling layer squashes the features of neighboring people into a fixed-size vector



Cannot directly apply to first-person videos

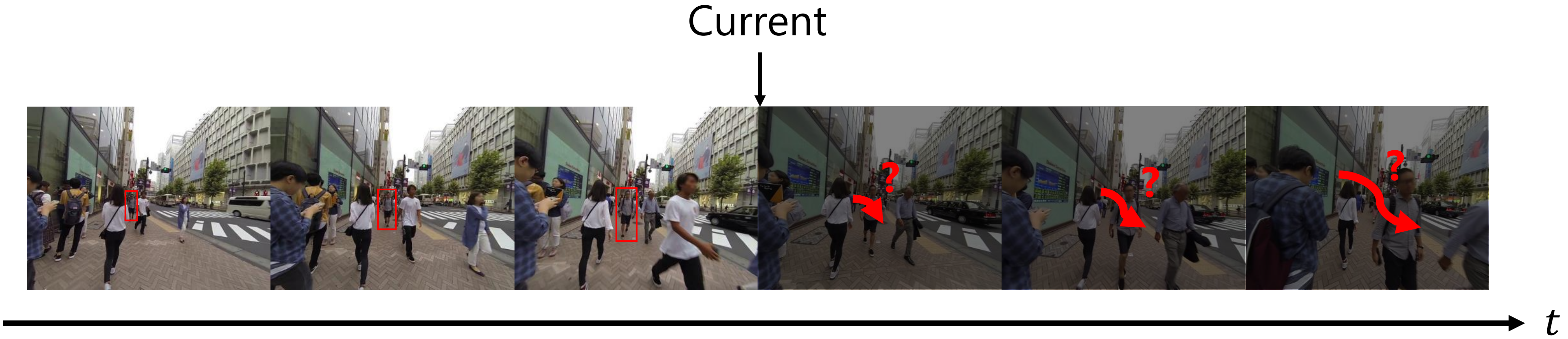
First-Person future person localization

Egocentric Future Localization [Park+, CVPR'16]



Predicts the wearer's future position

Future person localization in first-person videos



Our Challenge:

To develop a future person localization method tailored to first-person videos

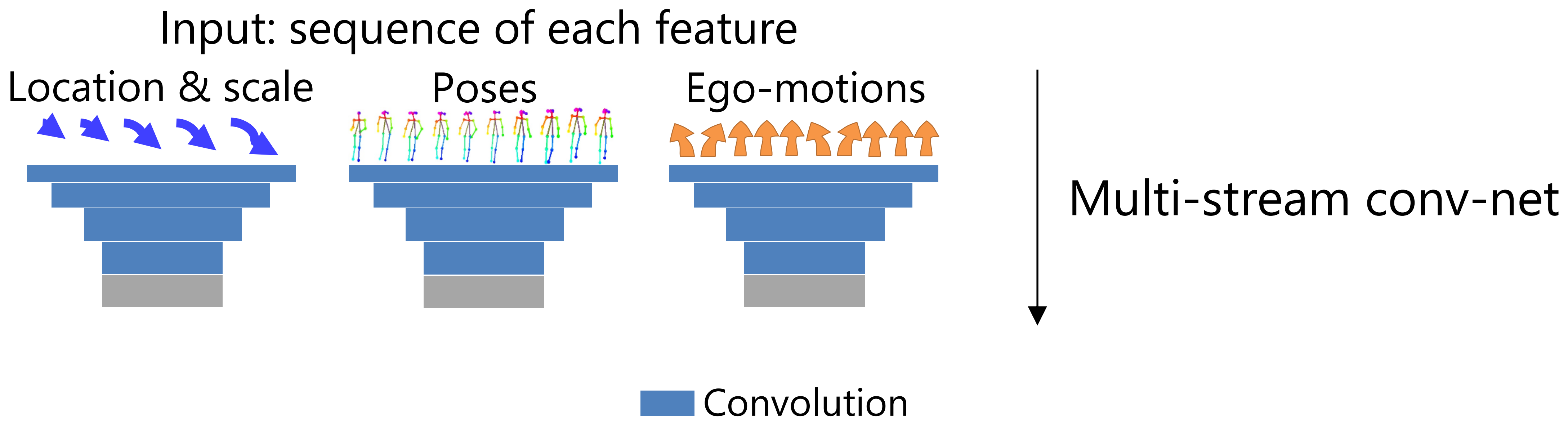
Our approach

- Incorporating both pose and ego-motion as a salient cue in first-person videos
- Multi-stream CNN to predict the future locations of a person

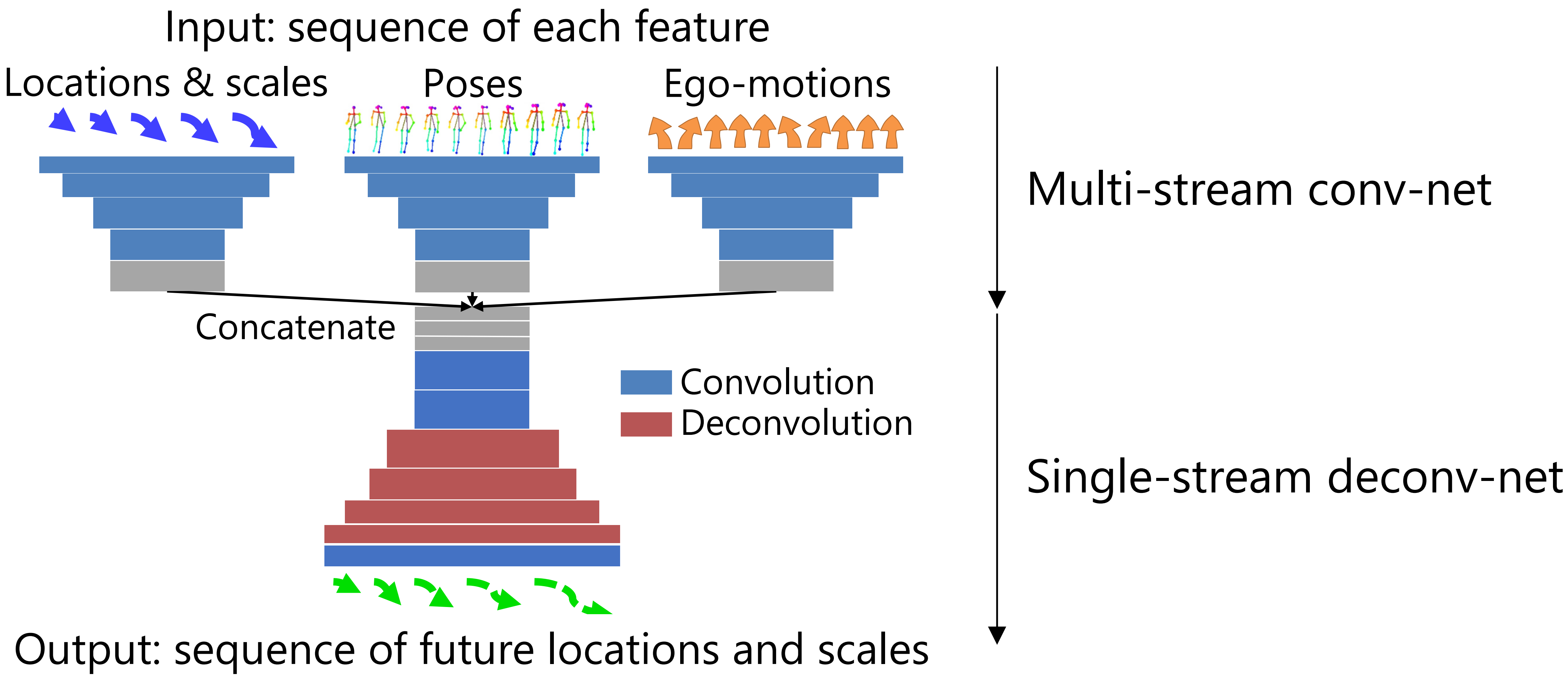
Pose indicates future direction

Ego-motion captures interactive locomotion

Proposed method: tri-stream 1D-CNN

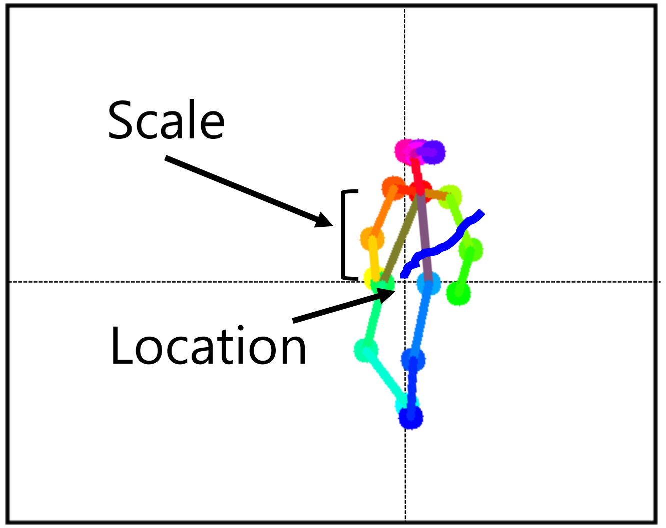


Proposed method: tri-stream 1D-CNN



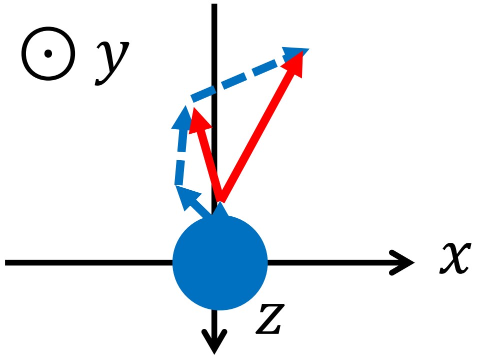
Feature representation

Target feature



- Location-scale cue (3 dims)
  - Location (2 dims) + scale (1 dim)
  - Captures perspective effect by the apparent size
- Pose cue (2D×18 keypoints=36 dims)
  - Used pretrained OpenPose [Cao+, CVPR'17]
  - Normalized position and scale
  - Imputed missing detections

Ego-motion feature

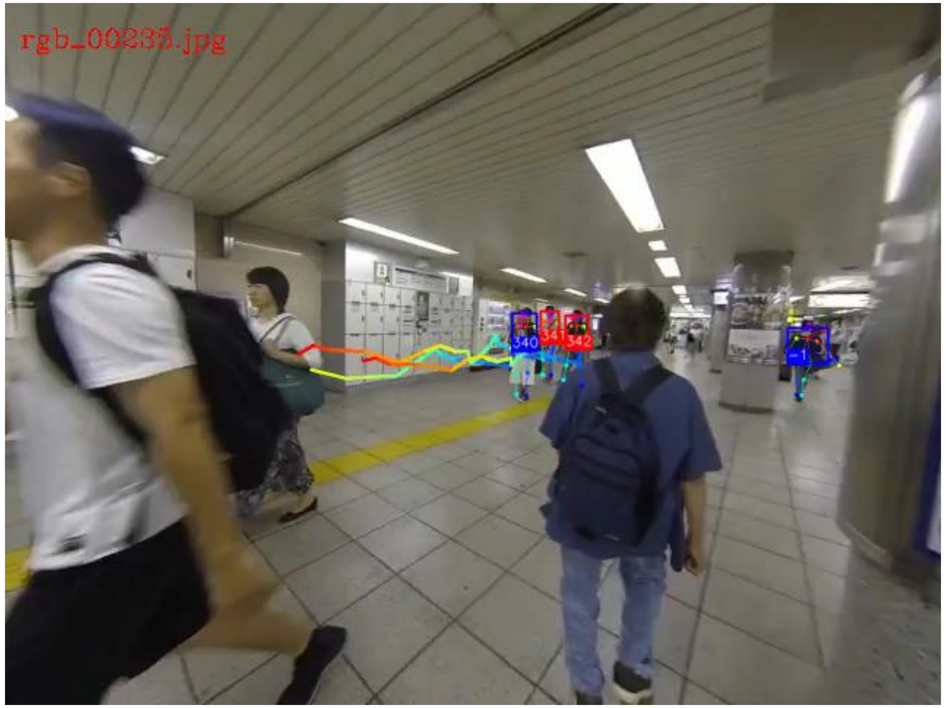


- Ego-motion cue (6 dims)
  - Camera pose estimation from multiple frames [Zhou+, CVPR'17]
  - Translation (3 dims) + rotation (3 dims)
  - Accumulate local movement between frames

Data collection

- Recorded walking video sequences in diverse cluttered scenes
  - One subject, total 4.5 hours, captured over 5,000 people
  - Annotations by tracking people

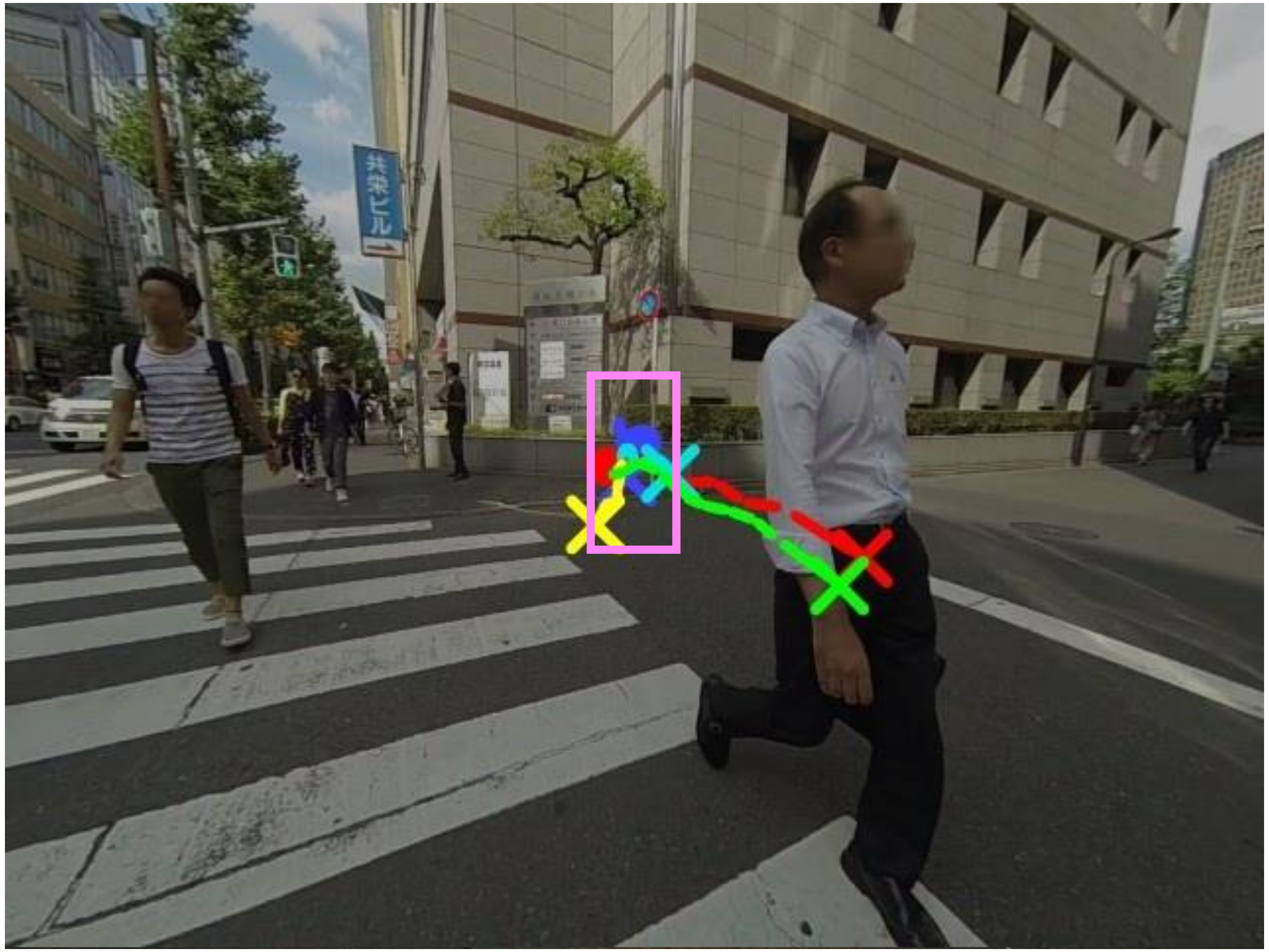
□: tracked ≧2s, □: tracked <2s



Baseline methods

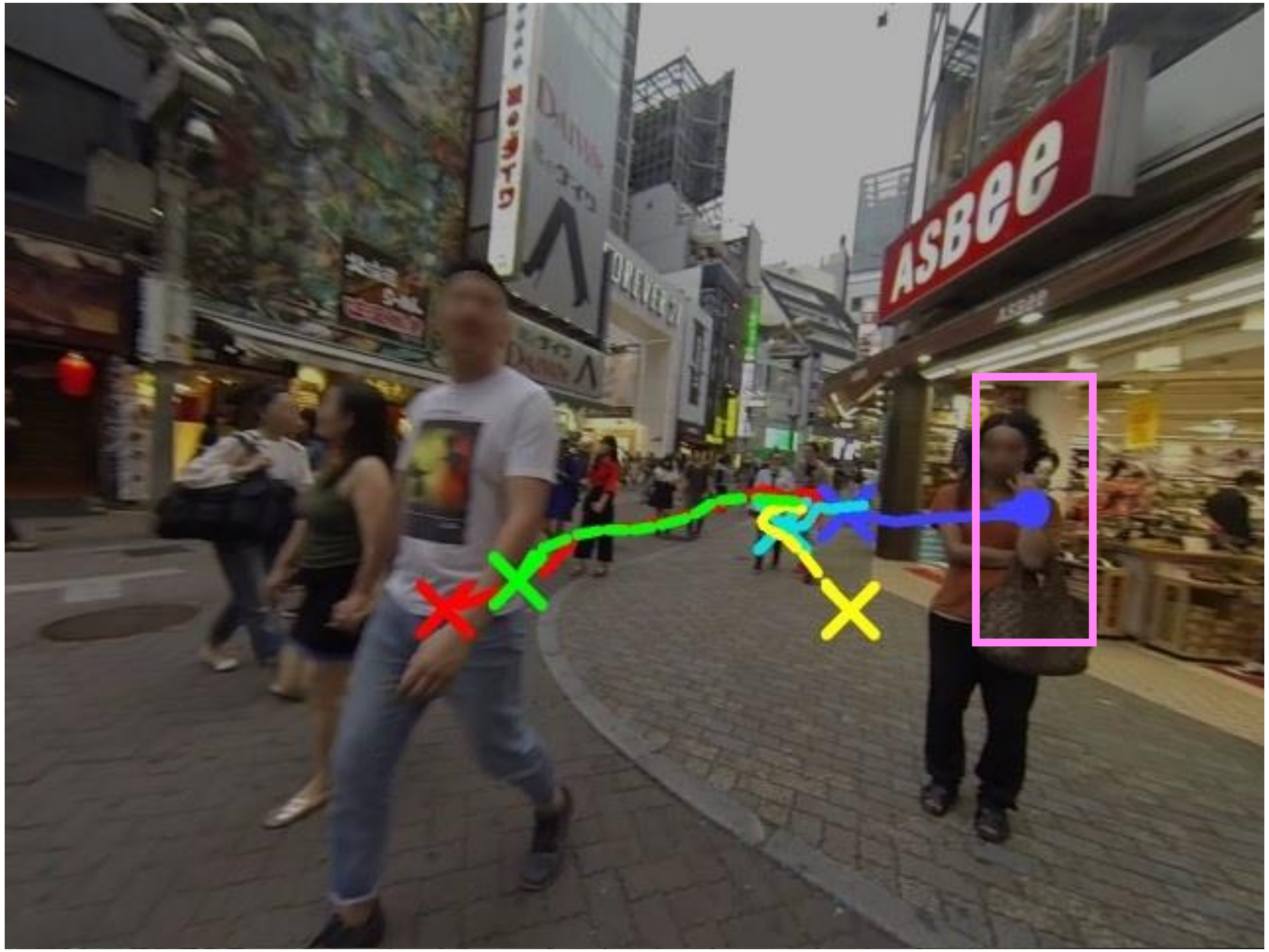
- **Constant**: Use location at the final input frame as prediction
- **ConstVel**: Assume a constant velocity model using the mean speed of inputs
- **NNeighbor**: Extracts k (=16) nearest neighbor input sequences, then produce output as the mean of the corresponding locations.
- **Social LSTM** [Alahi+, CVPR'16]: The state-of-the-art method on fixed cameras

Prediction example (input: 1sec, output: 1sec)



Input GT NNeighbor Social LSTM **Proposed**
[Alahi+, CVPR'16]

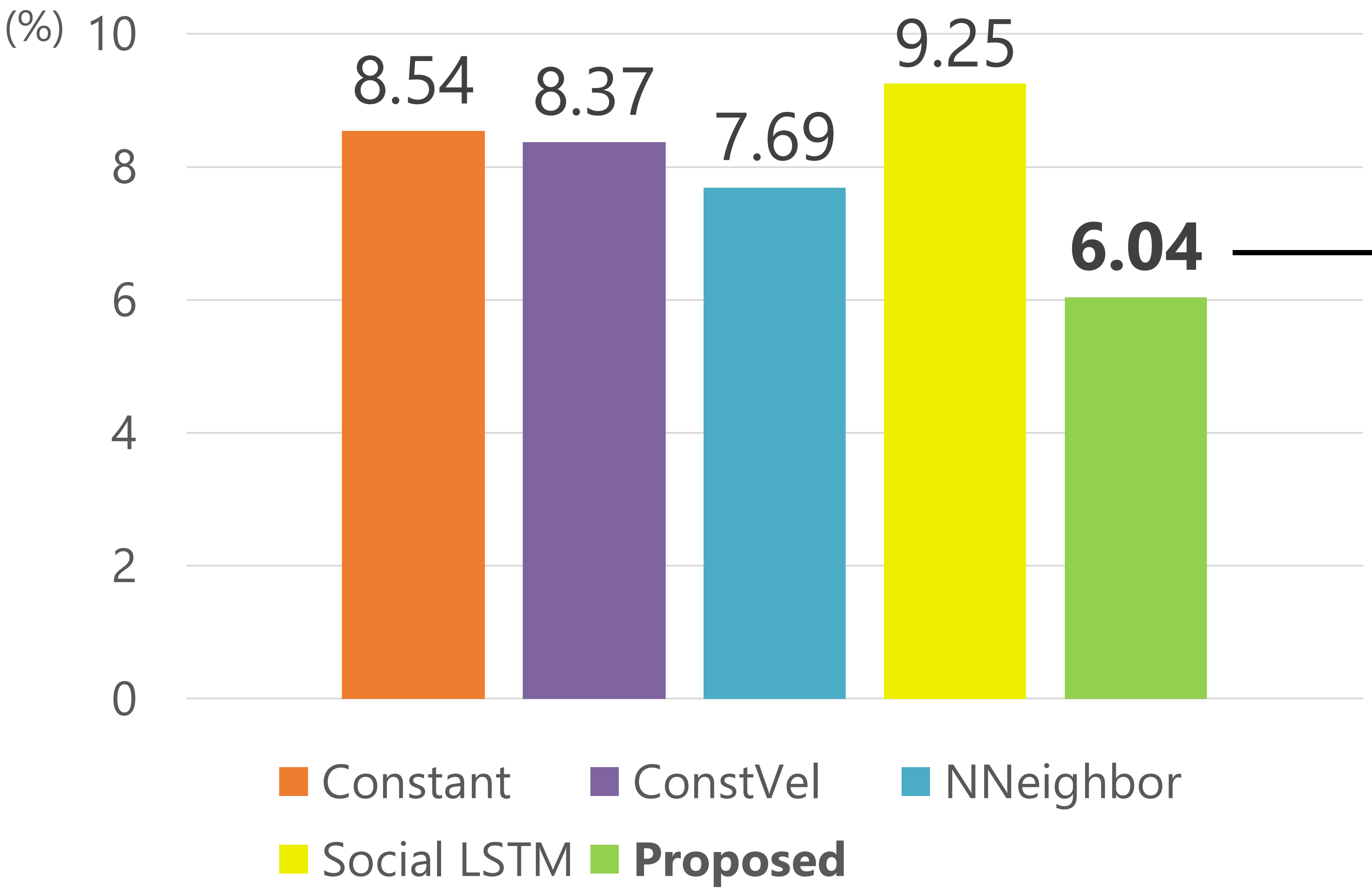
Prediction example (input: 1sec, output: 1sec)



Input GT NNeighbor Social LSTM **Proposed**

Quantitative evaluation

One-second prediction error (unit: % against frame width)



Equivalent to 60cm physical error

Ablation study

One-second prediction error (unit: % against frame width)

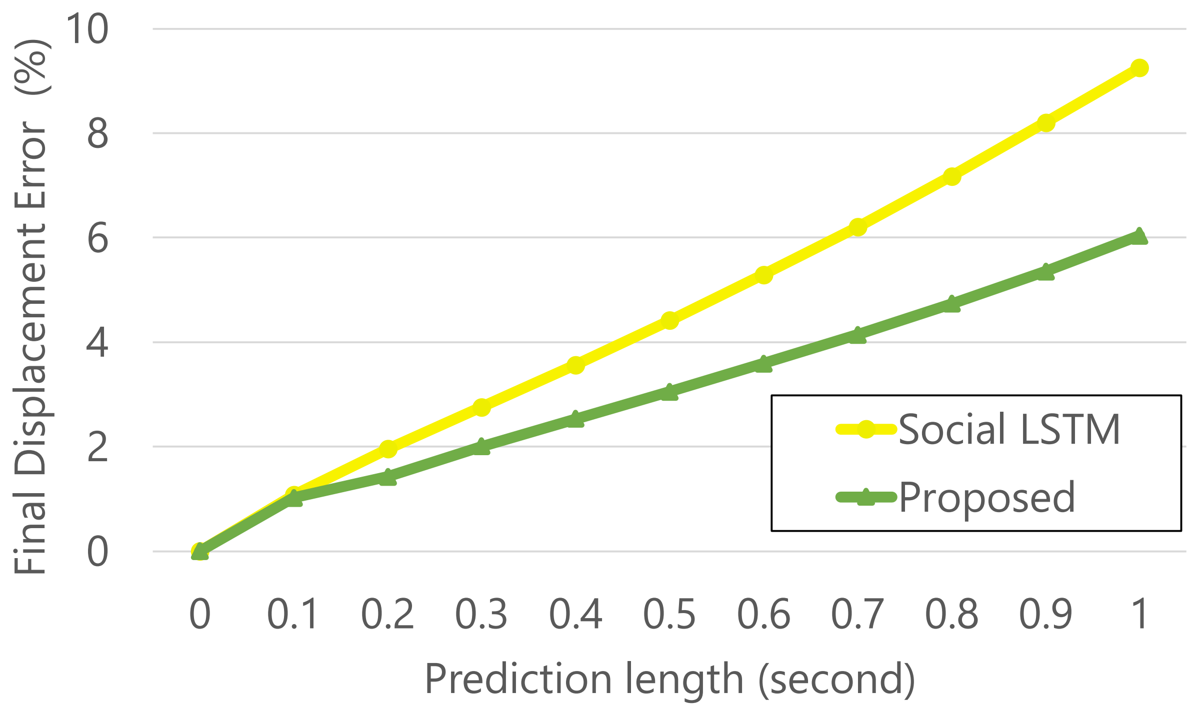
Features	Walking direction		
	Toward	Away	Average
Location + scale	9.26	6.02	6.40
+ Ego-motion	8.80	5.80	6.18
+ Pose	8.38	6.00	6.29
Proposed	**8.06**	**5.76**	**6.04**

(%)

- Pose ( ) contributes to predicting who comes **Towards** the wearer
- Ego-motion ( ) contributes to predicting who walks **Away** from the wearer

Effect of prediction length

Prediction error (unit: % against frame width)



- Prediction error linearly increases with prediction length
- Error increase rate is lower than the Social LSTM baseline

Predicting longer-term future

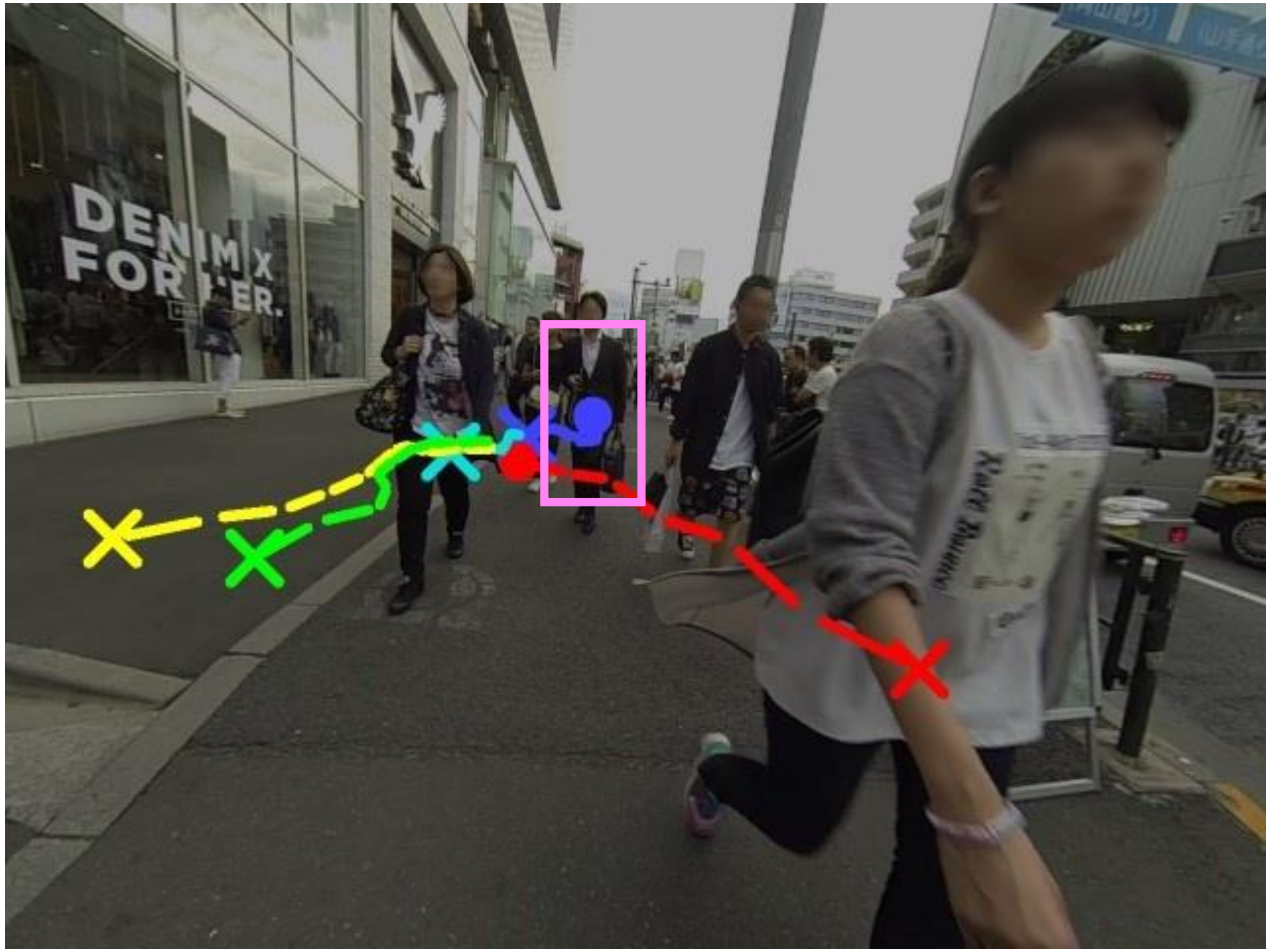
two-second prediction error (unit: % against frame width)

Method	Walking direction			
	Toward	Away	Average	Average (1.0s)
Social LSTM	22.12	17.56	17.75	9.23
Proposed	13.68	9.54	9.75	6.04

(%)

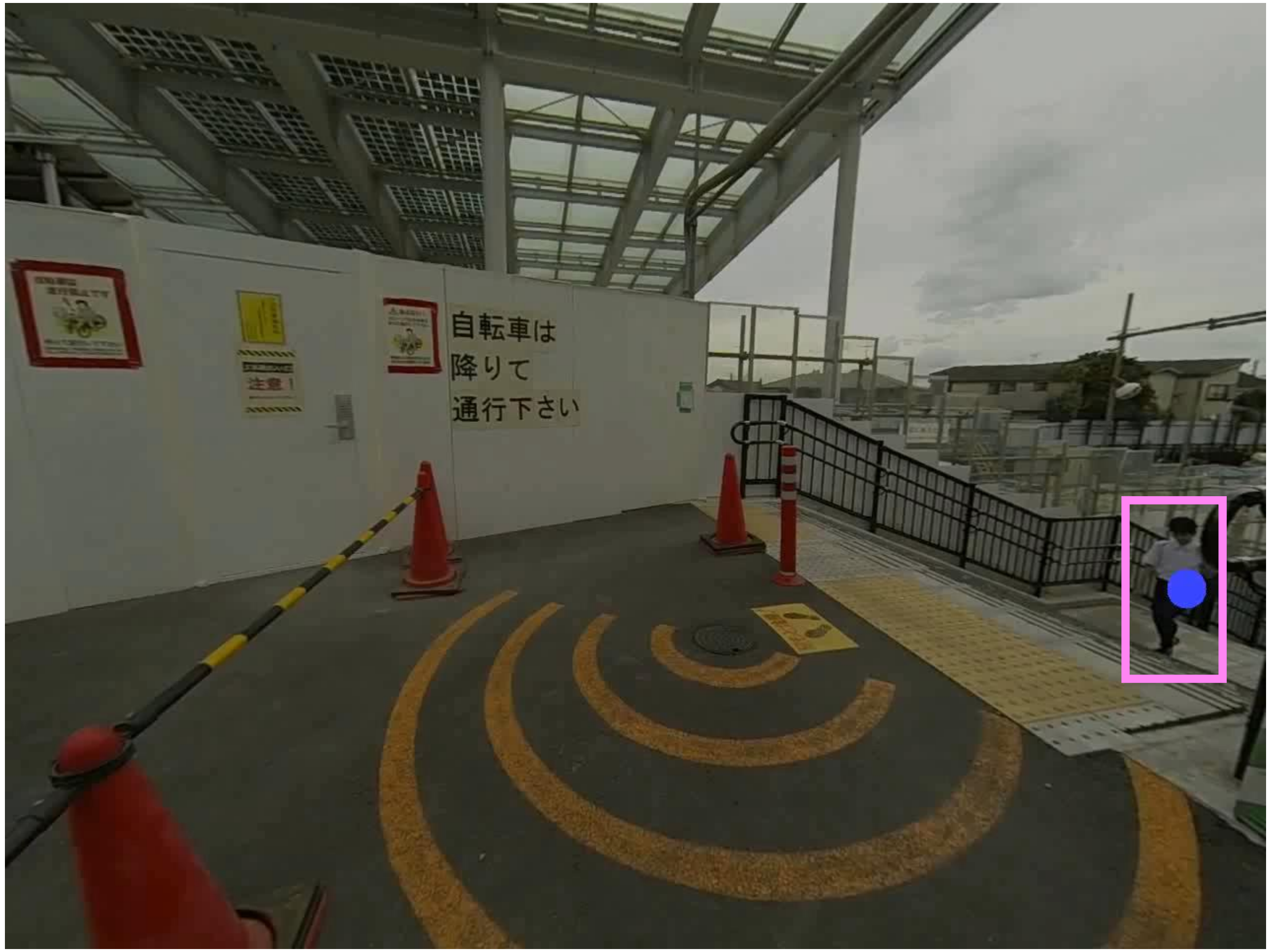
- Input: 0.6sec, output: 2.0sec
- Able to predict longer-term future with modest error increase

Failure case (existence of obstacles)



Input GT NNeighbor Social LSTM **Proposed**

Failure case (sudden direction change of the wearer)



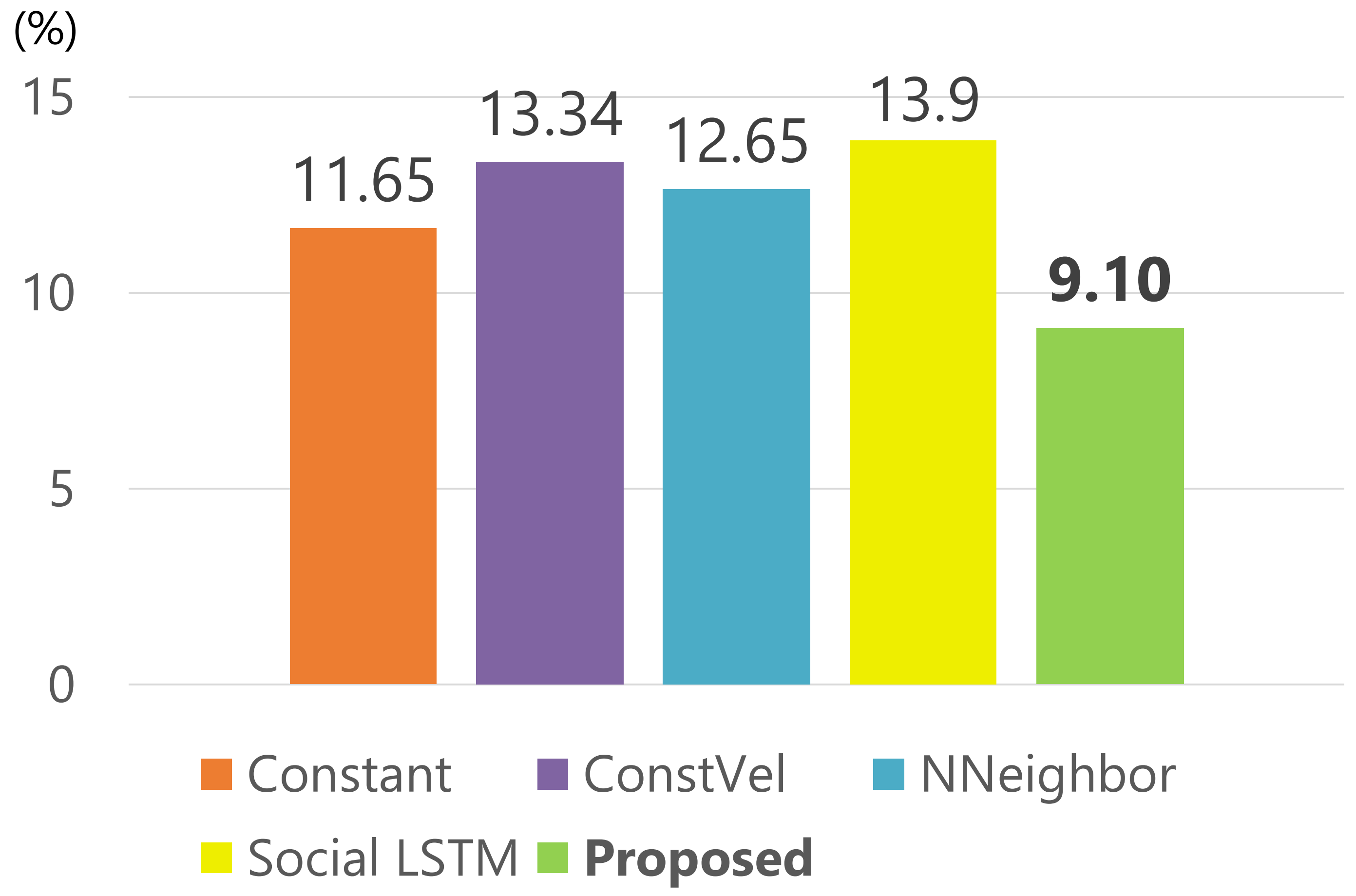
Input GT **Proposed**

Study in social interactions dataset [Fathi+, CVPR'12]

- Head-mounted videos in a theme park (more challenging setting)

Quantitative evaluation

1 second prediction error (unit: % against frame width)



Modest performance even in head-mounted videos

Prediction examples (input: 1sec, output: 1sec)

-0.9s | Current | +1.0s

Input GT **Proposed**

Summary

New Problem

- Future person localization in first-person videos

Finding

- Both target's pose and wearer's ego-motion were shown to be effective cues

Limitations

- Cross-subject evaluation (assume a single wearer in this work)
- Offline inference (currently not real-time)

Future Directions

- Forecasting under uncertainty
- Separating prediction of the wearer and the target

Publications

- International conference (refereed)
  - Takuma Yagi, Karttikeya Mangalam, Ryo Yonetani and Yoichi Sato, Future Person Localization in First-Person Videos, In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7593-7602, 2018. **(Spotlight oral, acceptance rate: 8.9%)**
- Domestic research workshop (non-refereed)
  - 八木拓真, マンガラムカーティケヤ, 米谷竜, 佐藤洋一, 一人称視点映像における人物位置予測, 第21回画像の認識・理解シンポジウム (MIRU), 2018.
  - 八木拓真, マンガラムカーティケヤ, 米谷竜, 佐藤洋一, 一人称視点映像における人物位置予測, 第211回CVIM研究会, 2018.