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

Junwei Liang[1]* Lu Jiang[2] Juan Carlos Niebles[3,2] Alexander Hauptmann[1] Li Fei-Fei[3,2]
[1]Carnegie Mellon University [2]Google AI [3]Stanford University
{junweil,alex}@cs.cmu.edu, lujiang@google.com, {lifeifei,jniebles}@cs.stanford.edu

We present more details and analysis for our experiments on the ActEV/VIRAT and ETH & UCY Benchmarks.

1. ActEV/VIRAT Details

1.1. Object & Activity Class

We show the object classes we used for our person interaction module and the activity classes for our activity prediction module in Table 1. Detailed class definition can be found on https://actev.nist.gov/.

1.2. Trajectory Type

In ActEV/VIRAT dataset, there are two distinctive types of trajectory: relatively static and the moving ones. We label the person trajectory as moving if at time T_{obs} there is an activity label of one of the following: "Walk", "Run", "Ride_Bike", otherwise we label it as static trajectory. Table 1.4 shows the mean displacement in pixels between the last observed point and the prediction trajectory points. As we see, there is a large difference between the two types of trajectory.

1.3. Nearest Neighbor Experiment

Since the ActEV/VIRAT experiment is not camera-independent, we conduct a nearest neighbor experiment. Specifically, for each observed sequence in the test set, we use the nearest sequence in the training set as future predictions. As shown in Table 3, it is non-trivial to predict human trajectory as people navigate differently even in the same scene. Please refer to the paper for evaluation metrics.

1.4. Single Model Experiment

We train 20 identical *Precog* models with different initialization for the single output experiment. We show the mean and standard deviation numbers in Table 3.

1.5. Single Feature Ablation Experiments

We experiment with ablating person-object, person-scene, person keypoint and person appearance feature, as shown in Table 4.

	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

Table 1. Object & Activity Classes.

	move_traj	static_traj
Average Displacement (train)	69.18	7.57
Final Displacement (train)	124.79	14.63
num% (train)	48.8%	51.2%
Average Displacement (test)	75.78	12.01
Final Displacement (test)	137.21	23.11
num% (test)	61.9%	38.1%

Table 2. Trajectory statistics for different trajectory class in ActEV dataset (on the training set).

1.6. Activity Detection Experiment

Since we are predicting activities in the not so distant future, a system may perform well enough if it just outputs the current activity labels as the future prediction. We train an identical model to detect the activity labels at time T_{obs} as the future prediction outputs, which results in a performance of 0.155 mAP for activity prediction and 18.27 ADE for trajectory prediction as shown in Table 4. Such a significant performance drop (0.192 vs. 0.155) suggests that activity prediction even for 4.8 seconds into the future is not a trivial task.

*Work partially done during a part-time research program at Google.

Metric	Nearest Neighbor	Our-Single-Model
ADE	40.04	17.99±0.043
FDE	73.69	37.24±0.102
move_ADE	39.52	20.34±0.059
move_FDE	72.67	42.54±0.146

Table 3. Our single model experiment on the ActEV/VIRAT benchmark.

1.7. More Qualitative Analysis

We show more qualitative analysis in Fig. 1. In each graph the yellow trajectories are the observable sequences of each person and the green trajectories are the ground truth future trajectories. The predicted trajectories are shown in the blue heatmap. To better visualize the predicted future activities of our method, we plot the person keypoint template for each predicted activity at the end of the predicted trajectory.

Successful cases: In Fig 1(a), Fig 1(b), Fig 1(c) and Fig 1(d), both the trajectory prediction and future activity prediction are correct. In Fig 1(d), our model successfully predicts the two persons at the bottom is going to walk past the car and also one of them is going to gesture at the other people by the trunk of the car.

Imperfect cases: In Fig 1(e) and Fig 1(f), although the activity predictions are correct, our model predicts the wrong trajectories. In Fig 1(e), our model fails to predict that the person is going to the other direction. In Fig 1(e), our model fails to predict that the person near the car is going to open the front door instead of the back door.

Failed cases: In Fig 1(g) and Fig 1(h), our model fails to predict both trajectories and activities. In Fig 1(h), the person on the bike is going to turn to avoid the incoming car while our model predicts a straight direction.

1.8. Comparing ActEV/VIRAT to ETH & UCY Benchmark

We compare the ActEV/VIRAT dataset and the ETH & UCY trajectory benchmark in Table 1.8. As we see, the ActEV/VIRAT dataset is much larger compared to the other benchmark. Also, the ActEV/VIRAT includes bounding box and activity annotations that could be used for multi-task learning. The ActEV/VIRAT is inherently different from the crow dataset since it includes diverse annotation of human activities rather than just passers-by, which makes trajectory prediction more purpose-oriented. We show the trajectory numbers after processing based on the setting of eight-second-length sequences. Note that in the public benchmark it is unbalanced since there is one crowded scene called ”University” that contains over half of the trajectories in 4 scenes.

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

Table 4. More ablation experiments on the ActEV/VIRAT benchmark.

	ActEV	ETH, UCY
#Scene	5	4
Dataset Length	4 hours 22 minutes	38 minutes
Resolutions	1920x1080, 1280x720	640x480, 720x576
FPS	30	25
Annotation FPS	30	2.5
#Traj	84600	19359, (10039 in Univ)
Annotations	Person+object bounding boxes, activities	Person coordinates

Table 5. Comparison to commonly used person trajectory benchmark datasets.

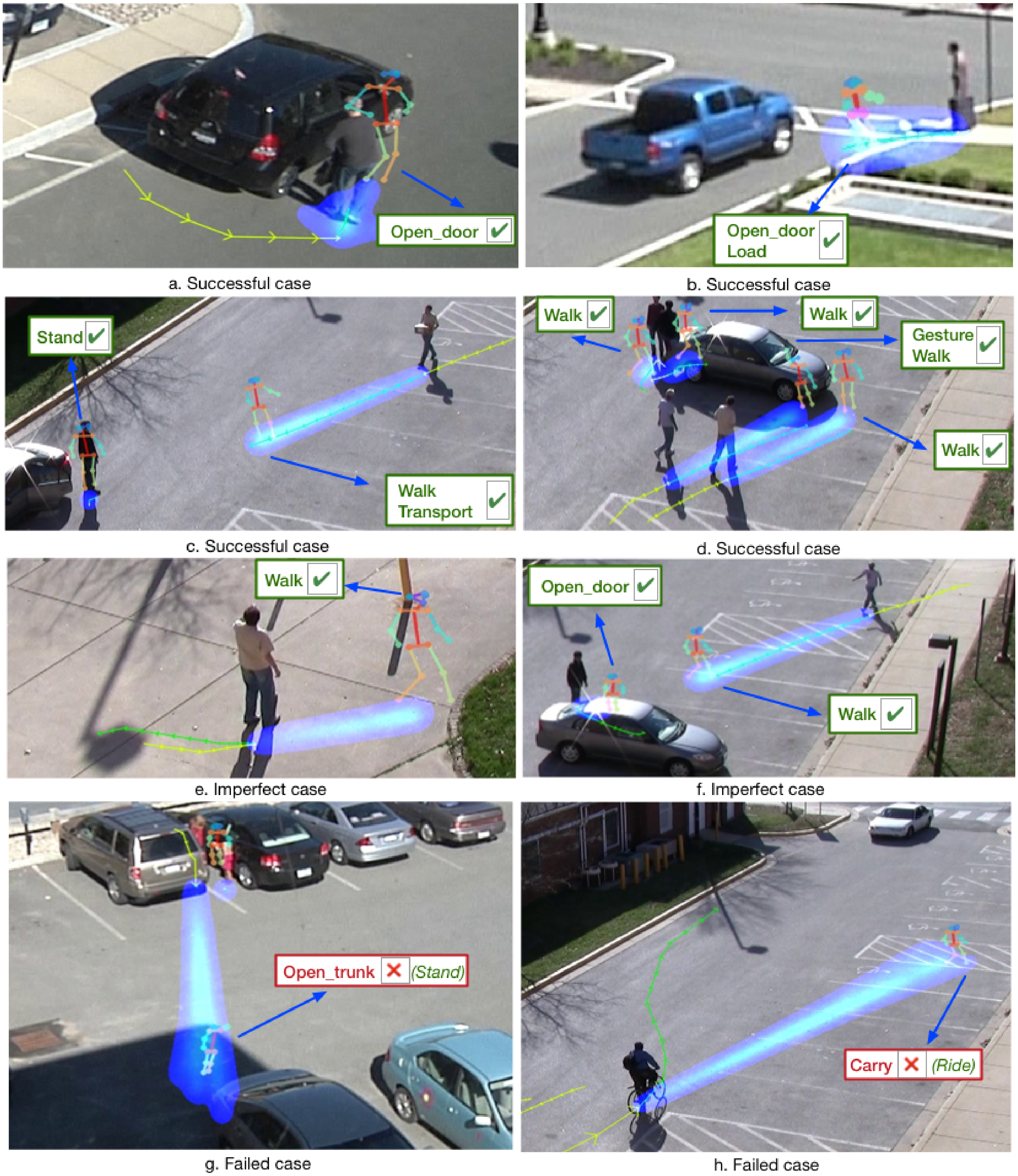
2. ETH & UCY Details

2.1. Dataset Difference Compared to SGAN

The dataset we use is slightly different from the one in [1], as some original videos are unavailable even though their trajectory annotations are provided. Specifically, two videos from UNIV scene, ”students001”, ”uni_examples”, and one video from ZARA3, ”crowds_zara03”, which is used in training for all corresponding splits in [1], cannot be downloaded from the dataset website. Therefore, the test set for UNIV we use is smaller than previous methods [1, 2] while the training set we use is about 34% smaller. Test sets for other 4 splits are the same therefore the numbers are comparable.

2.2. Pre-Processing Details

Since the annotation is only a point for each person and the human scale in each video doesn’t change much, we apply a fixed size expansion from the annotated points for each video to get the person bounding box annotation for appearance and person-scene feature pooling. Specifically, we use a bounding box size of 50 pixels by 80 pixels with the original annotation point putting at the center of the bottom line. All videos are resized to 720x576. The spatial dimension of the scene semantic segmentation feature is (64, 51) and two grid scales are used: (32, 26), (16, 13).



a. Successful case

b. Successful case

c. Successful case

d. Successful case

e. Imperfect case

f. Imperfect case

g. Failed case

h. Failed case

Figure 1. (Better viewed in color.) Qualitative analysis of our model.

References

[1] A. Gupta, J. Johnson, S. Savarese, Li Fei-Fei, and A. Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *CVPR*, 2018. 2

[2] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, and S. Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. *arXiv preprint arXiv:1806.01482*, 2018. 2