Supplementary Materials: Introvert: Human Trajectory Prediction via Conditional 3D Attention

Nasim Shafiee
Northeastern University
shafiee.n@northeastern.edu

Taskin Padir
Northeastern University
t.padir@northeastern.edu

Ehsan Elhamifar
Northeastern University
e.elhamifar@northeastern.edu

1. Ablation Studies

Effect of Conditional Visual Encoder. In order to investigate the effectiveness of different components of our proposed framework, we performed an ablation study and reported the average displacement error (ADE) and final displacement error (FDE) in first three rows of Table 1. Notice that without any visual encoder and by using only the kinematic encoder, the errors are large, obtaining 0.94 for ADE metric and 1.75 for FDE metric on average on the 5 datasets. Using our proposed 3D visual encoder, however without conditioning on the kinematics data, the performance significantly improves, obtaining 0.23 ADE and 0.37 FDE on average over the 5 datasets. This demonstrates the importance of using visual data for trajectory prediction. Finally, by conditioning our 3D visual encoder on kinematics information, we obtain the best performance, i.e., 0.21 and 0.34 as the average ADE and FDE on the 5 datasets. In particular, the conditioning reduces the error from 0.24 to 0.20 and from 0.45 to 0.42 on University and ETH datasets, respectively, showing the effectiveness and importance of our conditional 3D visual attention model.

Effect of Dual Attention Mechanism. As we discussed in Section 3.4, ρ_1, ρ_2, ρ_3 are responsible for extracting increasingly fine-grained visual (attention) features. Second three rows of Table 1 is shown ADE/FDE, when each ρ_i is excluded. Notice that removing ρ_2 results in significant performance drop as it is the main attention module and is conditioned on $\mathcal{Z}_{kin}, u_1$.

2. Results on Stanford Drone Dataset (SDD)

We compare our method against SoPhie and PECNet on the SDD (Pedestrians) [2], which has more challenging background than other datasets such as UCY (however, UCY is more crowded with higher density of people and includes more social interactions). For conducting the experiments on our method, we have splitted SDD (Pedestrian) into 80% training and 20% testing sets. We used leave-one-out method similar to our setup for UCY and ETH datasets. Table 2 shows the results of our method against Sophie and PECNet as reported in [3, 1]. Notice that our method outperforms both baselines, achieving lower ADE and FDE.

3. Visualization of Spatio-Temporal Attention

We demonstrate three samples from UCY and ETH datasets and the spatiotemporal attentions of the second and third attention modules, $\Psi 2$ and $\Psi 3$. As it is shown in Figure 1, Introvert produces attentions along both spatial and temporal dimension of visual data. For demonstration purpose, we upsampled these attentions and overlayed them on actual frames. The first attention is related to a target pedestrian in Zara2 dataset who passes by a car. Notice that the attention successfully focuses on the car. In the second row, the future path of the target pedestrian and location of the pedestrian in front of the target person are focused on by our attention model. In the third row, the future path of the target pedestrian and his surrounding pedestrians are attended to by our model. Notice that all the attention maps vary over time as well. Attentions related to $\Psi 2$ modules are fine-grained, focusing on detailed spatial information in the scene, while $\Psi 3$ attentions are coarse-level and more abstract as they are generated by a deeper layer of the CNN.

References

[1] Karttikeya Mangalam, Harshayu Girase, Shreyas Agarwal, Kuan-Hui Lee, Ehsan Adeli, Jitendra Malik, and Adrien Gaidon. It is not the journey but the destination: Endpoint conditioned trajectory prediction. *arXiv preprint arXiv:2004.02025*, 2020. 1

[2] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *European conference on computer vision*, pages 549–565. Springer, 2016. 1

[3] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, and Silvio Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1349–1358, 2019. 1

ADE / FDE	University	Zara 1	Zara 2	Hotel	ETH
Baseline(No 3D Visual Encoder)	0.84 / 1.57	0.62 / 1.29	0.43 / 0.81	0.40 / 0.66	2.4 / 4.42
Unconditional 3D Visual Encoder	0.24 / 0.39	0.18 / 0.30	0.16 / 0.24	0.11 / 0.17	0.45 / 0.73
Conditional 3D Visual Encoder	0.20 / 0.32	0.16 / 0.27	0.16 / 0.25	0.11 / 0.17	0.42 / 0.70
ρ_1 excluded	0.23 / 0.36	0.19 / 0.33	0.18 / 0.28	0.13 / 0.19	0.45 / 0.72
ρ_2 excluded	0.57 / 0.87	0.59 / 1.04	0.31 / 0.50	0.81 / 1.22	0.89 / 1.24
ρ_3 excluded	0.24 / 0.39	0.17 / 0.29	0.18 / 0.28	0.13 / 0.19	0.45 / 0.65

Table 1. The ablation study on the effectiveness of the conditional visual encoder and dual attention mechanism.

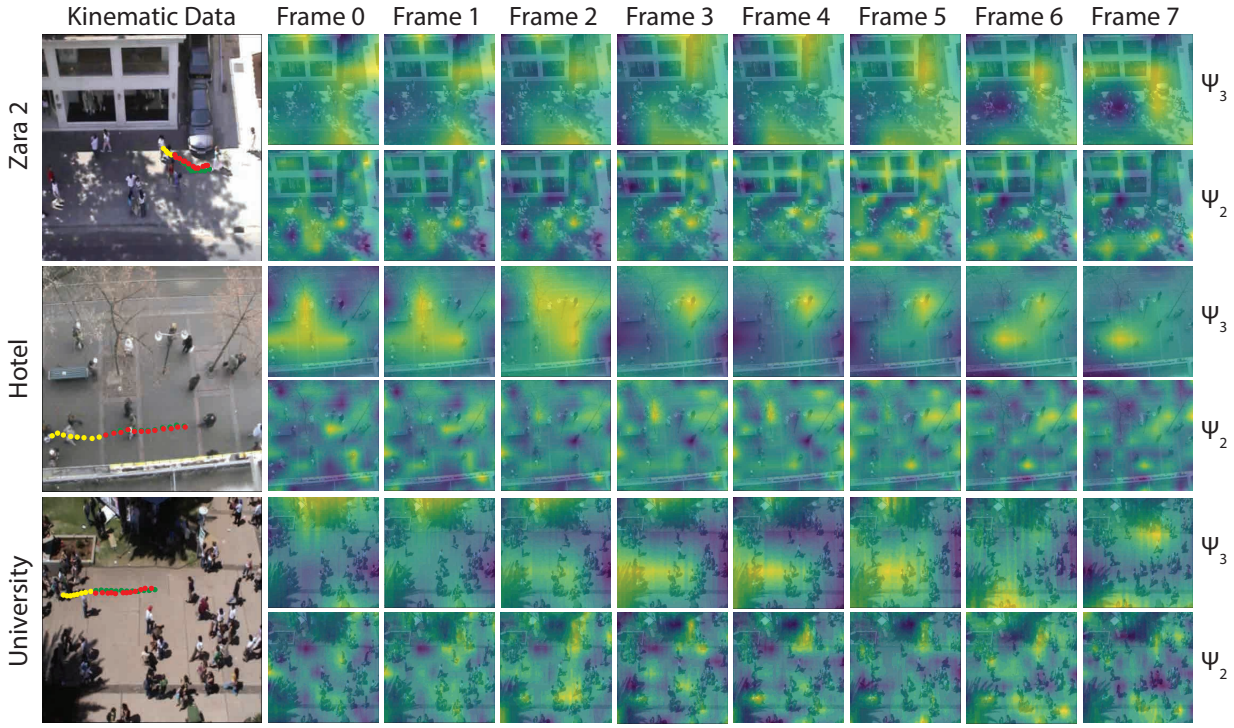

Figure 1. Demonstration of the Spatio-Temporal Attention obtained by our method. Left figures contain kinematic data of the target pedestrian. Yellow, Red and Green dots are observations, predictions and ground truth data, respectively. Right figures show the spatio-temporal attentions produced by $\Psi 2$ and $\Psi 3$ modules.

Method	SoPhie	PECNet	**Introvert (ours)**
ADE/FDE	16.27/29.38	9.96/15.88	**7.43/12.44**

Table 2. Performance (ADE/FDE) comparison on the Stanford Drone Dataset (SDD).