CNN, Segmentation or Semantic Embeddings: Evaluating Scene Context for Trajectory Prediction

Arsal Syed(✉) and Brendan Tran Morris

University of Nevada Las Vegas (UNLV), Las Vegas, NV 89154, USA
syeda3@unlv.nevada.edu, brendan.morris@unlv.edu

Abstract. For autonomous vehicles (AV) and social robot's navigation, it is important for them to completely understand their surroundings for natural and safe interactions. While it is often recognized that scene context is important for understanding pedestrian behavior, it has received less attention than modeling social-context – influence from interactions between pedestrians. In this paper, we evaluate the effectiveness of various scene representations for deep trajectory prediction. Our work focuses on characterizing the impact of scene representations (sematic images vs. semantic embeddings) and scene quality (competing semantic segmentation networks). We leverage a hierarchical RNN autoencoder to encode historical pedestrian motion, their social interaction and scene semantics into a low dimensional subspace and then decode to generate future motion prediction. Experimental evaluation on the ETH and UCY datasets show that using full scene semantics, specifically segmented images, can improve trajectory prediction over using just embeddings.

Keywords: Trajectory prediction · Scene context · RNN autoencoder

1 Introduction

Pedestrian trajectory prediction is an important topic in many applications including autonomous vehicles (AV), social robots, and surveillance systems. In AVs, accurate trajectory prediction of pedestrians enables the vehicle to achieve better control and operate in a safer manner. For surveillance systems, understanding pedestrian behavior can help to detect unusual activity. With the spread of COVID-19, another important application is ensuring a safe distance to meet social distancing guidelines. A recent surveillance implementation was used to predict whether people walking in a scene were maintaining safe distance and generate corresponding alerts if pedestrians are in too close proximity to each other [1]. The task of pedestrian trajectory prediction is challenging in nature due to erratic dynamics and ability to make sudden directional changes. However, when people walk, they tend to follow learned societal norms – for example they avoid collision and give right of way as necessary – but the social context and influences may not be consistent with all of those around them.

G. Bebis et al. (Eds.): ISVC 2020, LNCS 12510, pp. 706–717, 2020.
https://doi.org/10.1007/978-3-030-64559-5_56

Scene context also influences pedestrian motion as they tend to avoid static objects and follow "rules of the road." Before deep learning became popular, pedestrian motion was modeled using hand-crafted features and experimentally validated. One such famous technique was the Social Force (SF) model [2], which used vector functions to incorporate pedestrian interaction. In order to capture scene interaction, the SF model used repulsive and attractive potentials to incorporate interaction with stationary objects. Recently researchers have shifted from manual methods to data driven approaches for pedestrian motion modeling. Usually a video sequence is preprocessed to annotate trajectories and prediction is treated as a sequence-sequence learning task.

Our focus in this paper is to evaluate the importance of scene semantics on pedestrian motion and how to best incorporate those semantics into the learning framework. Recently researchers have explored different ways to incorporate scene context by using convolutional encoding layers of semantic segmentation architectures, like Deeplabv3 [3]. We believe that much of the contextual power of semantic segmentation architectures comes from the decoding layers which is lost when only using the encoding layers. Our experiments show that instead of embeddings, a fully segmented output map shows considerable improvement in trajectory prediction results.

2 Related Work

We review primary work on human trajectory prediction which are classified into two areas: (i) Human-human centric interaction where pedestrian motion is modeled through an occupancy grid map to provide social context. (ii) Pedestrian-scene interaction where influence of static scene objects (sidewalk/roads/exits etc.) is considered for scene context.

2.1 Human-Human Interaction

Social LSTM (S-LSTM) [4] was the primary research work that shifted focus from physics-based modeling approach to a deep learning model. It proposed a LSTM-based framework which models pedestrian interaction through a social pooling mechanism to share hidden state information with neighbors. Gupta et al. proposed Social GAN (S-GAN) [5], a GAN based network in conjunction with LSTM encoder-decoder to model multi-modal behavior of pedestrians. It also used an efficient global pooling technique, in contrast to the local pooling of S-LSTM [4], for social interaction. More recently graph neural networks have become popular where pedestrian motion and their interaction with surroundings is modeled by creating a spatio-temporal graph [6, 7]. However, these models fail to capture scene interactions.

Since pedestrian behavior is stochastic in nature, Variational Autoencoders (VAE) [8] and inverse optimal control [9] techniques have also been used. Similarly, these data driven techniques were extended for vehicle trajectory forecasting where Kim et al. [10] used LSTM models to forecast vehicle trajectory by understanding neighboring vehicle behavior through an occupancy grid map. Deo and Trivedi [11] addressed the social interaction of vehicles through a convolution social pooling layer and then predicted different vehicle maneuvers. Becker et al. [12] proposed the RED predictor to compare the

prediction results with linear interpolation as base line. It consists of an RNN-Encoder with a stacked multi-layer perceptron (MLP) and used smooth trajectories to forecast future time steps, which helped in preventing accumulation of error. Compared to other baseline methods like RNN-MLP, RNN-encoder-MLP and Temporal Convolutional Networks (TCN), RED has strong results on the TrajNet [13] challenge despite its simple architecture.

2.2 Human-Scene Interaction

To understand the importance of scene structure, Sadeghian et al. [14] proposed the SoPhie network which was comprised of an attention-based GAN network with VGG-19 as a scene feature extraction module. The attention mechanism focuses on the agents and static objects which are important for trajectory prediction. In [15], Bartoli et al. introduced context-aware trajectory prediction where it used a context based pooling strategy to include static scene features. More recent works like Peek into the Future (PIF) [16] and State Refinement LSTM (SR-LSTM) [17] extended S-LSTM by incorporating scene features and new pooling strategies to improve prediction results. Multi-agent tensor fusion is another recent work which leverages GANs and CNNs to capture dynamic vehicle behavior by fusing the motion of multiple agents and their respective scene context into a tensor. Social Scene-LSTM (SS-LSTM) [18] used an explicit scene context branch paired with social context and self context (dynamics) for prediction. They used a CNN encoder for the scene that was trained from scratch specifically for trajectory prediction. SSeg-LSTM [19] extended the idea of SS-LSTM to explicitly incorporate scene semantics through the use of a semantic segmentation network (SegNet [20]) to encode scene features and showed improvement compared to just CNN-encoding.

Most of recent scene structure techniques have utilized variants of CNN/VGG-16 for embedding scene information. In this work, we focus on evaluating scene encoding techniques and their ability to extract relevant scene features for pedestrian motion. We then further evaluate these methods with semantic segmentation architectures which have not been explored in depth for trajectory prediction problems.

3 Scene Augmented RAE Trajectory Prediction

In order to characterize the importance of scene-context for trajectory prediction, we utilize a network model with three branches to handle dynamics, social interactions, and scene contribution as shown in Fig. 1.

3.1 Encoding Context

For every pedestrian i in a scene, the observed motion is denoted as $\boldsymbol{X}_t^i = \left(\boldsymbol{x}_t^i, \boldsymbol{y}_t^i\right)$ where $t = 1, \ldots, t_{obs}$. The goal is to predict trajectories for future time stamps which is defined as $\hat{\boldsymbol{Y}}_t^i = \left(\hat{\boldsymbol{x}}_t^i, \hat{\boldsymbol{y}}_t^i\right)$ for $t = t_{obs+1}, \ldots, t_{pred}$. The scene branch utilizes a semantic segmentation network to provide scene encoding. Since the scene is static, a single frame is used once for each trajectory. Pedestrian dynamics are encoded as typical using an LSTM on historical positions [4, 5] for each time step. Social interactions between

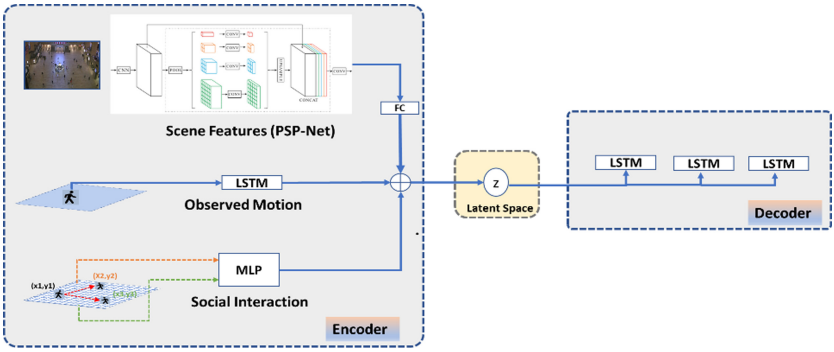

Fig. 1. Our proposed model based on RNN-AE which uses PSP-Net to incorporates Scene Semantics

neighboring pedestrians using a neighborhood pooling strategy similar to [5]. The social interaction tensor is generated based only on the time of prediction t_{obs}. Since the scene is static, only a single fixed frame is used as input for each trajectory.

This work uses PSP-Net [21] for semantic segmentation which has shown improvement in the pixel-level prediction through use of a different global pooling mechanism [22] where it explores the capability of global context information through region-based context aggregation. Previous methods, especially the early Fully Convolution Network (FCN) [23] had numerous problems when it came to parsing a scene. There were several mismatched relationships based on the appearance of objects, for example in the ADE20K [24] dataset it predicted car over water instead of boat. FCN also failed to recognize small scene objects like signboards and streetlights. Given FCN shortcomings, PSP-Net introduced a pyramid pooling module which is effective in capturing the global context prior. It used four modules in a pyramid fashion to fuse the feature context. The coarsest level is the global pooling module, which generates a single output. The subsequent levels divide the feature map into different sub regions and generated a pooled representation of its locations. Google DeepLab variants (notably DeepLab_v3) have become popular for semantic segmentation but only have marginal improvement over PSP-Net in benchmarks [3].

3.2 Decoding Trajectories

Trajectory predictions are generated through a sequence-sequence mapping through an autoencoder (AE). An AE is a self-supervised learning technique which has been extensively used in representation learning tasks. Usually the input to an AE is a set of features which are compressed into a bottleneck dimension and at the output, a decoder is used to reconstruct the original input. The key attribute to AE is the bottleneck representation or latent space without which the network will only memorize the input states. The latent space provides the necessary information to traverse the full network, forcing a learned compression of input data.

RNN-AE (RAE) on the other hand is a formulation of an AE for sequence data which uses an LSTM encoder-decoder architecture. In our pedestrian trajectory problem, the

encoder section consists of three branches. The top layer is used to capture scene semantics using the PSP-Net architecture. The middle branch encodes the observed trajectories and the last branch captures social interaction of pedestrians. Pedestrian information, scene and human interactions are encoded into a low dimensional latent space which is then provided as an input to the LSTM-decoder to generate future trajectories.

4 Experiments

Following the existing literature, we utilize UCY [25] and ETH [26] pedestrian datasets for evaluation. The total number of trajectories approximately 1500. The sampling period of pedestrian trajectories 0.4 s. We use the typical training method of k-fold cross validation to train on four and test on the remaining sequence. To evaluate the model, we observe the past trajectories for 3.2 s (8-time stamps) and predict for 4.8 s (12 time stamps).

The evaluation metrics used are Average displacement error (ADE) and final displacement errors (FDE). ADE measures the average prediction performance along the trajectory for all the pedestrians in the scene. FDE on other hand measures the performance for the last or end point of trajectory for the pedestrians in the scene. To keep evaluation consistent for fair comparison, like SGAN [5] we also draw 20 sample trajectories closest to ground truth and then compute ADE and FDE.

$$ADE = \frac{\Sigma_{i=1}^{N} \Sigma_{t=t_{obs+1}}^{T} \left\| \hat{Y}_t^i - Y_t^i \right\|_2}{N * T} \quad FDE = \frac{\Sigma_{i=1}^{N} \left\| \hat{Y}_{t_{pred}}^i - Y_{t_{pred}}^i \right\|_2}{N} \tag{1}$$

Where N refers to number of pedestrians in a scene and $T = t_{pred} - t_{obs} + 1$ is the prediction horizon.

4.1 Implementation Details

Our model uses an LSTM framework to encode observed trajectories and PSP-Net to capture scene features. We utilize the same pooling method training parameters as mentioned in generator module of [5]. The network is trained for 200 epochs with batch size of 32 and learning rate as 0.005. The latent space (z) has dimension of 64. To implement segmentation networks, we annotated different frames across pedestrian datasets and generated image masks. The networks were trained for 20 epochs with batch size of 2 using Adam optimizer.

4.2 Pedestrian Scene Segmentation

We first examined the quality of semantic scene segmentation with different segmentation networks such as fully trained (FT) SegNet and PSP-Net. A visual comparison of segmentation results is provided in Fig. 2 which highlights cleaner results from PSP-Net over SegNet. PSP-Net better identifies light post, trees, and the bench in Hotel. The grass area on the top of Univ is more cleanly segmented against the building. Finally, in ZARA we see that SegNet was not able to parse the lower vehicle correctly.

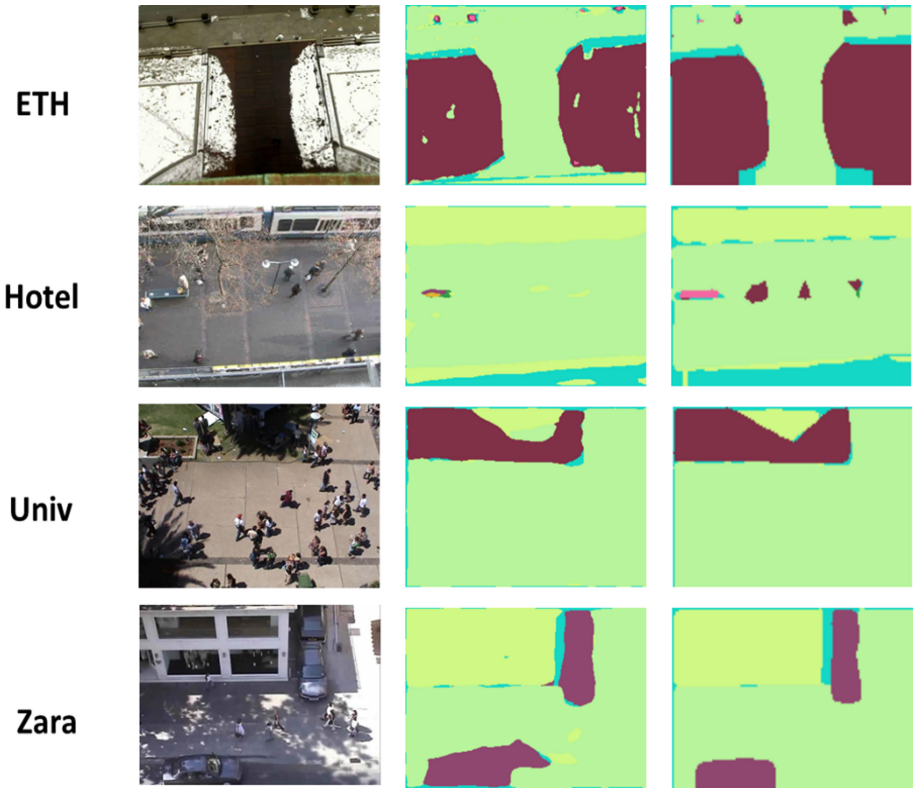

Fig. 2. Segmentation results: scene image (left), fully trained SegNet (middle), and fully trained PSP-Net (right)

In general, PSP-Net had consistent scene segmentation making it more understandable. Hence it is expected that the PSP-Net architecture will produce better trajectory prediction results. We also used SegNet which was pre-trained (PT) using CamVid [26] dataset. Unsurprisingly, it failed to accurately capture scene semantics on UCY-Univ and ETH-Univ dataset which are from a top-down view rather than street level as in CamVid.

4.3 Quantitative Analysis

We compare our prediction results with following baseline models:

- **S-LSTM** [4]: Pedestrian behavior modeled using LSTM and social interaction through hidden layer pooling. It has no scene context.
- **S-GAN** [5]: An adversarial network architecture with generator that uses a global pooling module to capture pedestrian's social interactions. It has no scene context.
- **SoPhie** [14]: An attention-based GAN network which uses VGG-19 tuned for FCN segmentation as backbone to extract scene context.

- **RAE-VGG-16:** We replace the PSP-Net in the scene encoding branch in Fig. 1 with an off-the-shelf VGG-16 encoder pretrained on ImageNet. The output of the ReLU following the last convolution layer is used for scene encoding/embedding.
- **RAE-SegNet:** We replace the PSP-Net with SetNet, both pre-trained (PT) and fine-tuned (FT), to characterize the effect of segmentation quality.
- **RAE-PSPNet-emb:** Instead of utilizing the decoder and segmented image, we use only the semantic embedding representation after encoding into a feature map after concatenating different levels of pyramid pooling module in PSP-Net.

The performance of the various trajectory prediction algorithms is shown in Table 1. The architectures which use scene information tend to perform better as they take into account of where pedestrians walk and their point of interests. Note S-GAN-VP20 reports the best of 20 predicted trajectories.

Table 1. ADE/FDE (meters) comparison

Models	ETH	Hotel	Univ	Zara1	Zara2	Avg
Social Embedding						
S-LSTM [4]	1.09/2.35	0.79/1.76	0.67/1.40	0.47/1.00	0.56/1.17	0.72/1.54
S-GAN-VP20 [5]	0.87/1.62	0.67/1.37	0.76/1.52	0.35/**0.68**	0.42/0.84	0.61/1.21
Scene Embedding						
Sophie (T_o+I_o) [14]	0.86/1.65	0.84/1.80	0.58/1.27	0.34/**0.68**	0.40/0.82	0.64/1.24
RAE-VGG-16	0.86 / 1.65	0.89/1.75	0.56/**1.14**	0.42/0.80	0.40/0.81	0.62/1.23
RAE-PSPNet-emb	0.88/1.70	0.79/1.53	0.56/1.39	0.41/0.97	0.42/0.96	0.61/1.31
Scene Segmentation						
RAE-SegNet-PT	1.11/1.87	0.70/1.38	0.86/1.77	0.57/1.18	0.59/1.21	0.90/1.75
RAE-SegNet-FT	0.84/1.51	0.68/1.36	0.54/1.41	0.33/1.11	**0.36**/0.83	0.55/1.24
RAE-PSPNet (ours)	**0.79/1.48**	**0.64/1.34**	**0.52**/1.40	**0.32**/0.99	**0.36/0.80**	**0.52/1.20**

Social Embedding vs Scene Embedding: From Table 1, we see that Sophie outperforms on ETH, Univ, Zara1, Zara2 and RAE-VGG-16 shows better performance on ETH, Univ and Zara2 when compared with models (S-LSTM and S-GAN-VP20) that do not incorporate scene behavior. This shows that for trajectory prediction, capturing scene features are essential. We want to further investigate how much influence trajectory motion has if we can come up with better methods to encode scene information.

Scene Embedding vs Scene Segmentation: While, semantic embedding only provided marginal improvement of social embedding, there is considerable improvement with scene segmentation. We speculate that much of the representation power from semantic segmentation networks comes from the decoder which needs to produce semantic interpretation and labels from embeddings. The VGG-16 or ResNet (PSPNet-emb)

encoders do not provide strong semantic insight, therefore, taking advantage of decoder is necessary to accurately encode scene features.

However, the use of semantic segmentation images alone is not sufficient. As seen in Table 1, RAE-SegNet-PT performs poorly compared to scene embedding models. Since SegNet-PT is trained on driver's view perspective images and not from surveillance point-of-view, it was not able to semantically segment and identify points of interest for the pedestrians. In order to overcome this shortcoming and to keep the evaluation consistent, we trained SegNet from scratch which reduced the trajectory prediction errors. To further test our hypothesis, we experimented with fully training the more advanced segmentation architecture of PSP-Net and found further improved segmentation performance, especially on Hotel and Zara datasets. From Table 1 we see that RAE-PSPNet outperforms all other models with 0.52/1.20 ADE/FDE.

The RAE-PSPNet results are placed in context in Table 2 by comparison with state-of-the-art (SOTA). While there is a clear gap in performance with SOTA, our model uses a RAE which is a simpler model in comparison with more advanced graph convolutional networks or Transformer networks [31]. Further, the SOTA approaches should be able add a scene branch or replace scene embeddings with scene images easily. In order to provide a bound on segmentation-based performance, we replaced the semantic segmentation network with ground truth labeled images. As expected, ground truth segmentation resulted in improved performance across the board versus RAE-PSPNet and approaches performance of some more recent (and more complicated) networks.

Table 2. ADE/FDE (meters) comparison with other state of the art architectures

Models	ETH	Hotel	Univ	Zara1	Zara2	Avg
RAE-PSPNet (ours)	0.79/1.48	0.64/1.34	0.52/1.40	0.32/0.99	0.36/0.80	0.52/1.20
Ground truth	0.75/1.45	0.62/1.33	0.50/1.40	0.29/0.96	0.31/0.75	0.49/1.17
Trajectron++ [28]	**0.35/0.77**	**0.18**/0.38	**0.22/0.48**	**0.14/0.28**	**0.14**/0.75	**0.21/0.45**
Soc-BIGAT [29]	0.69/1.29	0.49/1.01	0.55/1.32	0.30/0.62	0.36/0.75	0.48/1.00
MATF-GAN [30]	1.01/1.75	0.43/0.80	0.44/0.91	0.26/0.45	0.26/0.57	0.48/0.90
TF-based [31]	0.61/1.12	**0.18/0.30**	0.35/0.65	0.22/0.38	0.17/**0.32**	0.31/0.55

4.4 Qualitative Analysis

Here we will discuss different scenarios which shows the importance of semantics and its influence on trajectory prediction. When people walk, they tend to traverse a walkable path, for example pavements, entrances/exits, etc. It is essential for our network to semantically identify the points of interest for pedestrian motion.

In Fig. 3(a) RAE-PSPNet and RAE-SegNet-FT are able to avoid the stopped pedestrians and look to avoid the bench. Since S-GAN does not model scene features, we see a diverging path. Figure 3(b) shows that RAE-PSPNet is able to predict a trajectory that follows the edge of the path next to the snow covering. In Fig. 3(c) we observe

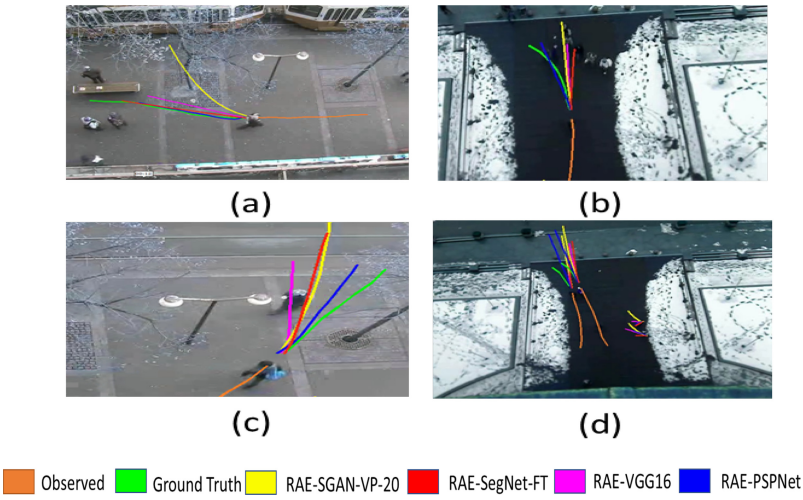

Fig. 3. Trajectory Prediction Comparison. (a) Similar results for all (but S-GAN). (b–c) RAE-PSPNet with improvement due to strong scene influence. (d) Poor performance from all models.

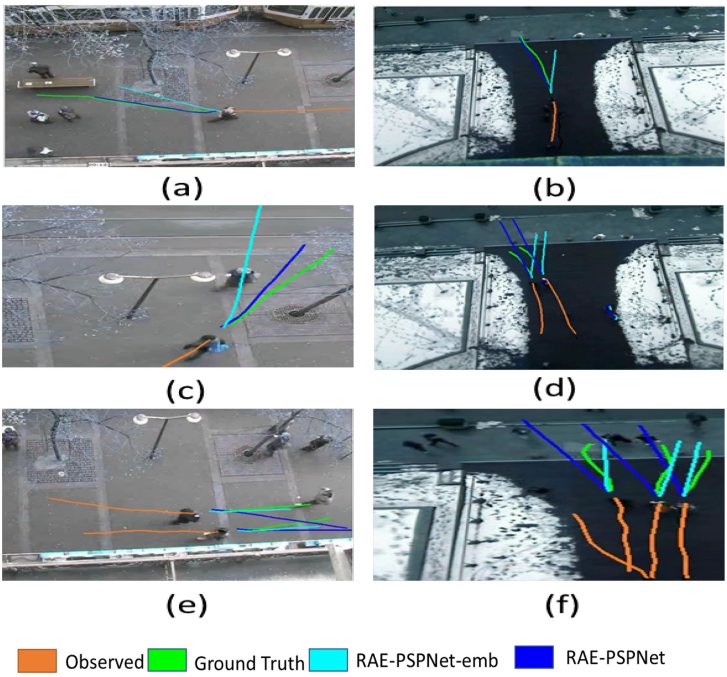

Fig. 4. (a) RAE-PSPNet-emb shows poor results. (b–c) RAE-PSPNet-emb was not able to capture scene information and has bad predictions. (d) Both perform poor but RAE-PSPNet-emb is even worse. (e–f) RAE-PSPNet-emb does better where pedestrians walk in groups.

a rich scene interaction and therefore RAE-PSPNet has successful predictions compared to RAE-SegNet-FT. This is because in Fig. 2 PSP-Net was better able to identify static objects (pole and lamppost) where as SegNet failed to capture those features. Figure 4(d) shows and example where all techniques work poorly and are not able to predict the slowing while the couple turn to the left.

Specific comparison between semantic images and embeddings are shown in Fig. 4. In (a), (c) the semantic obstacle information does not seem to effectively utilized by RAE-PSPNet-emb. In (b), (d) the embedded prediction continues straight up while RAE-PSPNet is able to have predictions that follow the snow bank. When comparing RAE-PSPNet-emb with RAE-PSPNet, it shows that using semantic embedding alone is not enough and that there is value in the decoding steps which generate the fully segmented output image. Figures 4(e–f) shows that RAE-PSPNet has poorer performance than RAE-PSPNet-emb for groups of pedestrians. It may be that the social interaction module is not weighted enough (or scene too much) or needs more complex modeling though something like graph convolution networks.

5 Conclusion

In this work, we evaluated different strategies for encoding scene information for better understanding of pedestrian motion. We leveraged hierarchical RNN based autoencoders to encode semantic context with observed raw motion and social interaction. We then replaced the scene encoding branch with different off-the-shelf feature extraction modules such VGG-16, SegNet, and PSPNet. Our experiments and analysis of pedestrian interaction with various scenarios showed that availability of full and accurate segmented output map can be helpful for forecasting pedestrian trajectories compared to just semantic and scene embeddings. While this work used a RAE, the scene encoding branch can be replaced and scene segmentation applied to other architectures, such as GANs and graph neural networks, without any difficulty and can bring added value for pedestrian motion prediction.

References

1. Punn, N.S., Sonbhadra, S.K., Agarwal, S.: Monitoring COVID-19 social distancing with person detection and tracking via fine-tuned YOLO v3 and Deepsort techniques. arXiv preprint arXiv:2005.01385 (2020)
2. Helbing, D., Molnar, P.: Social force model for pedestrian dynamics. Phys. Rev. E **51**(5), 4282 (1995)
3. Chen, L.-C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. CoRR, abs/1706.05587 (2017)
4. Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S.: Social LSTM: human trajectory prediction in crowded spaces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 961–971 (2016)
5. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A.: Social GAN: socially acceptable trajectories with generative adversarial networks. arXiv preprint arXiv:1803.10892 (2018)

6. Mohamed, A., Qian, K., Elhoseiny, M., Claudel, C.: Social-STGCNN: a social spatio-temporal graph convolutional neural network for human trajectory prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14 424–14 432 (2020)
7. Kosaraju, V., Sadeghian, A., Martín-Martín, R., Reid, I., Rezatofighi, S.H., Savarese, S.: Social-BIGAT: multimodal trajectory forecasting using bicycle-GAN and graph attention networks. arXiv preprint arXiv:1907.03395 (2019)
8. Bhattacharyya, A., Hanselmann, M., Fritz, M., Schiele, B., Straehle, C.-N.: Conditional flow variational autoencoder for structured sequence prediction. In: BDL@NeurIPS (2019)
9. Kitani, K.M., Ziebart, B.D., Bagnell, J.A., Hebert, M.: Activity forecasting. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7575, pp. 201–214. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33765-9_15
10. Kim, B., et al.: Probabilistic vehicle trajectory prediction over occupancy grid map via recurrent neural network. arXiv preprint arXiv:1704.07049 (2017)
11. Deo, N., Trivedi, M.M.: Convolutional social pooling for vehicle trajectory prediction. CoRR abs/1805.06771 (2018)
12. Becker, S., Hug, R., Hübner, W., Arens, M.: An evaluation of trajectory prediction approaches and notes on the TrajNet benchmark. arXiv:1805.07663 (2018)
13. Traj Net Challenge. http://trajnet.stanford.edu/
14. Sadeghian, A., Kosaraju, V., Sadeghian, A., Hirose, N., Savarese, S.: SoPhie: an attentive GAN for predicting paths compliant to social and physical constraints. arXiv preprint arXiv:1806.01482 (2018)
15. Bartoli, F., Lisanti, G., Ballan, L., Del Bimbo, A.: Context aware trajectory prediction. arXiv:1705.02503 (2017)
16. Liang, J., Jiang, L., Niebles, J.C., Hauptmann, A.G., Fei-Fei, L.: Peeking into the future: predicting future person activities and locations in videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5725–5734 (2019)
17. Zhang, P., Ouyang, W., Zhang, P., Xue, J., Zheng, N.: SR-LSTM: state refinement for LSTM towards pedestrian trajectory prediction. In: Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition, pp. 12085–12094 (2019)
18. Xue, H., Huynh, D.Q., Reynolds, M.: SS-LSTM: a hierarchical LSM model for pedestrian trajectory prediction. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1186–1194 (2018)
19. Syed, A., Morris, B.T.: SSeg-LSTM: semantic scene segmentation for trajectory prediction. In: 2019 IEEE Intelligent Vehicles Symposium (IV), pp. 2504–2509. IEEE (2019)
20. Badrinarayanan, V., Handa, A., Cipolla, R.: SegNet: a deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. arXiv preprint arXiv:1505.07293 (2015)
21. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR (2017)
22. Liu, W., Rabinovich, A., Berg, A.C.: ParseNet: looking wider to see better. arXiv:1506.04579 (2015)
23. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
24. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ADE20K dataset. arXiv:1608.05442 (2016)
25. Lerner, A., Chrysanthou, Y., Lischinski, D.: Crowds by example. In: Computer Graphics Forum, vol. 26, pp. 655–664. Wiley Online Library (2007)
26. Pellegrini, S., Ess, A., Schindler, K., Van Gool, L.: You will never walk alone: modeling social behavior for multi-target tracking. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 261–268 (2009)

27. Brostow, G., Fauqueur, J., Cipolla, R.: Semantic object classes: a high-definition ground truth database. PRL **30**(2), 88–97 (2009)
28. Salzmann, T., Ivanovic, B., Chakravarty, P., Pavone, M.: Trajectron++: multi-agent generative trajectory forecasting with heterogeneous data for control. arXiv preprint arXiv:2001.03093 (2020)
29. Kosaraju, V., Sadeghian, A., Martín-Martín, R., Reid, I., Rezatofighi, H., Savarese, S.: Social-BIGAT: multimodal trajectory forecasting using bicycle-GAN and graph attention networks. In: Advances in Neural Information Processing Systems, pp. 137–146 (2019)
30. Zhao, T., et al.: Multi-agent tensor fusion for contextual trajectory prediction. arXiv preprint arXiv:1904.04776 (2019)
31. Giuliari, F., Hasan, I., Cristani, M., Galasso, F.: Transformer networks for trajectory forecasting. arXiv preprint arXiv:2003.08111 (2020)