Adversarial Loss for Human Trajectory Prediction

Parth Kothari
VITA Lab, EPFL

Alexandre Alahi
VITA Lab, EPFL

July 3, 2019

Abstract

Autonomous vehicles need to accurately forecast future human trajectories in order to navigate safely and gain human trust. Capturing the subtle intricacies of human motion is a challenging task as human trajectories are multimodal *i.e.* given a past trajectory, multiple socially compliant future trajectories are possible. To this effect, various human trajectory prediction models have been proposed to capture not only the multimodality but also the unspoken social rules of mobility. The current best performing approach is based on Generative Adversarial Network (GAN) taking advantage of the success of Recurrent Neural Network (RNN) models in sequence prediction tasks. In this work, we highlight an unexpected pitfall in this state-of-the-art architecture via controlled experiments. We provide an explanation for this shortcoming and propose a modification to the given architecture leveraging the architectures used in the GAN community. Furthermore, we prove the efficacy of the proposed modification on synthetic data and real world datasets, thereby indicating room for improvement on state-of-the-art.

Introduction

The unwritten rules of human motion are subtle and many, making the task of human trajectory prediction challenging. Humans have the innate ability to not only avoid collisions but to do so in a socially acceptable manner. For *e.g.*, consider the case of crowded spaces like airports or railway stations, where humans have to follow social conventions like respecting personal space and yielding right-of-way. Moreover, human trajectories are multimodal in nature, *i.e.*, given a past motion history, multiple future predictions are possible. Capturing these intricacies into a single model for generating human-like trajectories is not trivial.

Traditional trajectory prediction methods have largely modelled the social interaction aspect, the most notable work being Social Force model [1] by Helbing and Molnar. However, these methods were based on handcrafted features which fail to handle interactions in complex environments. With the success of deep learning techniques in approximating complex functions, various solutions to learn human-human interactions in a data-driven fashion have been proposed. Recently, Alahi *et. al.* [2] incorporated the social interactions among agents into trajectory forecasting model based on Recurrent Neural Networks (RNNs). On the other hand, Lee *et. al.* [3] output different route choices based on a given static scene in order to model the multimodal aspect of the human motion. However, none of these models enforce the predicted trajectories to be real looking *i.e.* human-like.

Generative Adversarial Networks (GANs) [4] have shown great success in producing realistic samples from distributions of complex data like images. Inspired by the successful application of GANs, Gupta *et al.* proposed Social GAN [5] that tackles not only social interactions and multimodality but also produces *socially acceptable* trajectories by integrating RNN models into the GAN framework. The RNNs aid in future trajectory generation conditioned on past trajectory, while the GAN framework assists in producing multiple socially acceptable trajectories. The main architecture differences among selected data-driven models for human trajectory prediction are

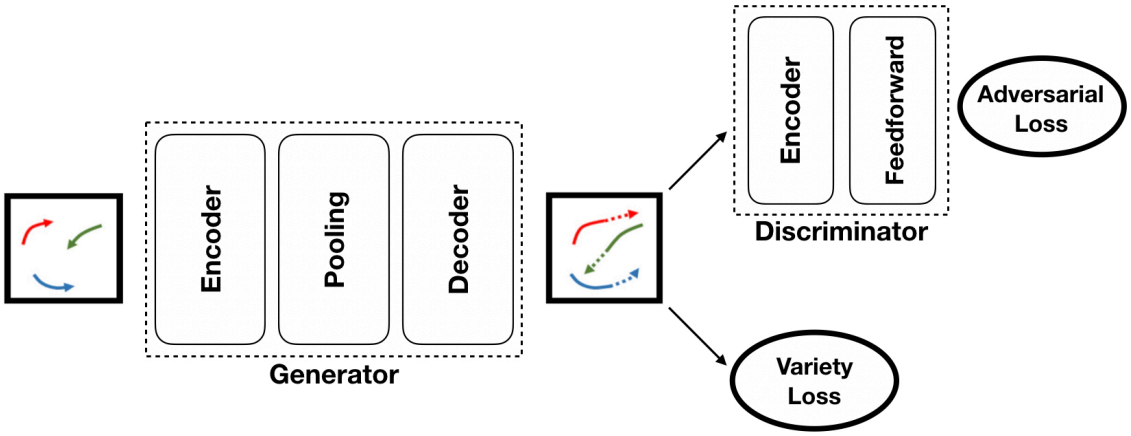


Figure 1: An illustration of Social GAN architecture. The discriminator combined with the adversarial loss is primarily responsible for differentiating the Social GAN architecture from other data-driven models [See Table 1]

displayed in Table 1. Note that the individual components can vary in their design. For *e.g.*, the pooling module of Social LSTM differs from that of Social GAN.

Model	Encoder	Pooling	Decoder	Discriminator	Variety/L2 Loss	Adversarial Loss
LSTM	✓	×	✓	×	×	×
Social LSTM	✓	✓	✓	×	✓	×
Social GAN	✓	✓	✓	✓	✓	✓

Table 1: Main architecture differences across selected trajectory prediction models

In this work, we aim to provide an detailed analysis of incorporating adversarial loss in trajectory prediction task. To this effect, we determine the performance improvement on incorporating the GAN framework into human trajectory prediction models by performing an ablation study on the loss components. Later, we analyze the GAN architecture used in Social GAN highlighting an unexpected pitfall of using RNN models inside the GAN framework through a controlled experiment. We propose a modification to this architecture and prove its efficacy empirically by reporting the results of our proposed modification on real world datasets.

Related Work

Theoretically, the trajectory prediction problem is defined as follows [6, 7]: At any time instant t, the xy-coordinates of the i^{th} person are denoted by (x_i^t, y_i^t). The networks takes as input the positions of all the people from time 1 to T_{obs}, and outputs their predicted positions from time instants $T_{obs}+1$ to T_{pred}. This task is analogous to a sequence generation problem [8], where the input sequence is the observed positions of a person and the output sequence denotes his/her future positions at different time-instants. This problem of predicting human motion has been approached in two different ways: non-data-driven and data driven.

Traditionally, path prediction problems have been tackled extensively through different approaches such as Kalman filters [9], linear regressions [10], non-linear Gaussian Process regression models [11, 12] among others. Specific to human trajectory prediction problem, Helbing and Molnar [1] presented a pedestrian motion model with attractive and repulsive forces, referred to as the Social Force model, capturing the human-human interaction. Their seminal work has been shown to achieve competitive results even on modern pedestrian datasets. Various other methods define

useful features to aid in the trajectory prediciton problem. Robicquet *et. al.* [13] characterize human motion into different navigation styles. Alahi *et. al.* [7] define Social Affinity Maps (SAM) to link broken or unobserved trajectories. Yi *et. al.* [14] cluster humans into groups and reason that members of a group follow similar behavior. However, all these works, which rely on handcrafted human-engineered features, fail to capture the interactions in more complex environments.

With the success of deep learning, recent methods for human trajectory prediction draw inspiration from techniques in machine learning to generate socially acceptable and multimodal trajectories. Human trajectory prediction is, at its core, a sequence prediction task. Inspired by the success of RNN models and their variants including Long Short Term Memory (LSTM) [15] and Gated Recurrent Units [16] in sequence prediction tasks [17, 18, 19, 20], these methods extended the sequence generation model from Graves *et. al.* [8] to predict human trajectories. Lee *et. al.* [3] introduced a RNN Encoder-Decoder framework which uses variational autoencoder (VAE) for trajectory prediction. Alahi *et. al.* [2] incorporated a social pooling mechanism inside the LSTM framework to predict future trajectories taking into account the motion of neighbouring agents. Nonetheless, the above mentioned data-driven techniques do not enforce the output trajectories to be human-like.

Generative Adversarial Networks (GANs) [4] have become the de facto generative models for generating high dimensional complex data distributions like images. GANs revolve around the concept of a two player game between a generator and discriminator. The objective of the generator is to fool the discriminator into believing that the generated samples are real. On the other hand, the discriminator has to correctly classify whether the given sample is real or fake. Both the generator G and the discriminator D are modelled using neural networks. The generator G takes as input a noise vector z sampled from a given noise distribution p_z and transforms it into a real looking sample $G(z)$. G essentially maps the noise distribution p_z to a generator distribution p_g. D assigns a score to each sample $D(x)$, indicating whether a sample comes from the generator distribution p_g or the real data distribution p_r. In standard GAN training, the discriminator guides the generator by providing gradients to improve the fake generated sample through backpropagation. The game between the generator G and the discriminator D can be formulated as a minimax objective problem:

$$\min_G \max_D \mathbb{E}_{x\sim p_r}[\log(D(x))] + \mathbb{E}_{z\sim p_z}[1 - \log(D(G(z)))]. \tag{1}$$

GANs have achieved great success in various tasks such as image generation [21, 22], representation learning [23], image style transfer [24] among others. Despite these promising results, in practice training GANs is a challenge. The gradients provided by the discriminator are prone to vanish [25] or explode [26] over time, often leading to training instability.

Recently, Gupta *et. al.* [5] incorporated the LSTM encoder-decoder architecture into the GAN framework to not only produce socially acceptable but also multimodal trajectories, thereby achieving state-of-the-art results on modern trajectory datasets. They propose an adversarial loss along with novel variety loss to handle both socially acceptable and multimodal aspect of trajectory prediction. We revisit these losses in detail in the next section.

In this work, we dig deeper and gather further insights into the intricacies of the adversarial loss for trajectory prediction. We observe the performance of the two loss components in both controlled experiments as well as real world datasets. We discuss a plausible reasoning behind a particular unnatural observation in performance of Social GAN, suggesting possible modifications to improve the model. We later report the results of our proposed modification on real world datasets indicating a room for improvement for the state-of-the-art architecture.

The Loss Function

The success of GANs in modelling complex distributions along with the success of RNNs in providing remarkable outputs in sequence generation naturally leads one to combine the two

frameworks. However, keeping in mind the notorious instability of GAN training [25, 26], such an integration needs to be done carefully.

Social GAN [5] introduces two losses, namely variety loss and adversarial loss. To promote the diversity of generated trajectories, Social GAN introduces a novel variety loss defined as:

$$\mathcal{L}_{variety} = \min_{k} \|Y_i - \hat{Y}_i^{(k)}\| \tag{2}$$

where k is a hyperparameter corresponding to the number of trajectories sampled from the generator. By considering only the best trajectory, the variety loss encourages the network to spread its output trajectories.

The adversarial loss in Social GAN, relates to the GAN objective described in Eq. 1. The adversarial loss for the GAN discriminator and GAN generator defined as:

$$\begin{aligned} L_D &= -\mathbb{E}_{x\sim p_r}[\log D(x)] - \mathbb{E}_{z\sim p_z}[1 - \log D(G(z))] \\ L_G &= -\mathbb{E}_{z\sim p_z}[\log D(G(z))]. \\ L_{adv} &= L_G + L_D \end{aligned} \tag{3}$$

The adversarial loss provides additional benefits over variety loss. With the objective to decrease its adversarial loss, the generator is forced to produce human-like trajectories in order to fool the discriminator. Thus, incorporating the adversarial loss holds the potential to learn a generative model that produces trajectories conforming to unwritten social norms of human motion.

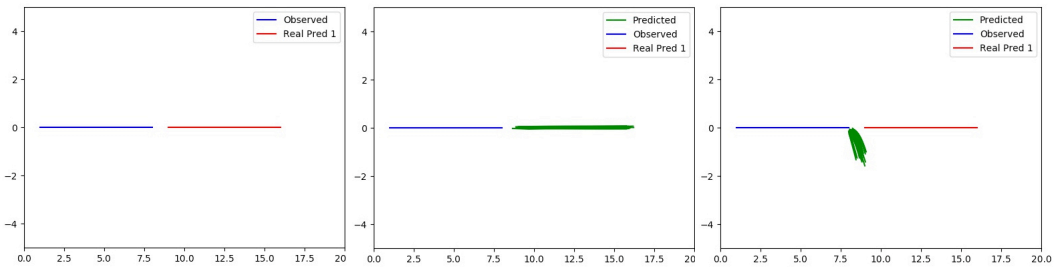
In this work, we divert from modelling the human interactions and focus on the role of the adversarial loss in trajectory prediction framework. In particular, we seek to determine the performance boost obtained on incorporating adversarial loss. To do so, we write the overall Social GAN loss as:

$$L_{SGAN} = \lambda * L_{variety} + (1 - \lambda) * L_{adv} \tag{4}$$

We vary the hyperparameter λ to determine the trade-off between accurate trajectories and socially acceptable trajectories. Analyzing this performance indirectly indicates the improvement obtained on integrating the GAN framework into standard sequence prediction models.

Controlled Experiments

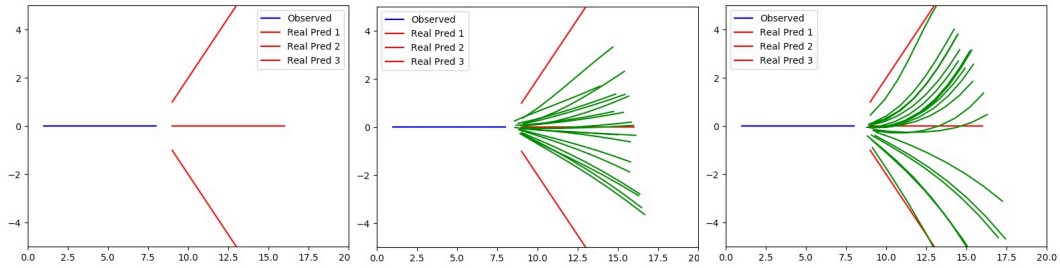
We start by considering the Social GAN performance on a simple synthetic trajectory dataset: Straight lines. This synthetic dataset consists of 500 trajectories which are produced by adding small gaussian noise $\mathcal{N}(0, 0.04)$ to straight lines. We use the same hyper-parameters, as the ones reported [1] to obtain state-of-the-art results, to train the social GAN. Fig. 2 provides a



(a) (b) (c)

Figure 2: Analyzing Performance of Adversarial Loss on Straight Trajectory dataset. (a) A sample from the dataset. (b) Performance of combined losses on the dataset (c) Performance of only adversarial loss on the dataset. The adversarial loss alone fails to recover the data distribution.

[1] Social GAN Code available at https://github.com/agrimgupta92/sgan



(a) (b) (c)

Figure 3: Analyzing performance of adversarial loss on an early terminated Social GAN model. (a) A sampled observed trajectory (blue) ends up in one of the three possible futures (red) (b) Social GAN training is early terminated. We can observe that the resulting trajectory distribution (green) does not span the entire space of ground truth outcomes. (c) Training Social GAN further using only the adversarial loss helps in spanning the 3 modes thereby proving its effectiveness.

visualization comparing the ground truth trajectory with the predicted ones [2]. The Social GAN is trained with two different loss setups: (a) Using both variety loss and adversarial loss [Fig.(2b)] (b) Using adversarial loss only [Fig.(2c)]. It is evident that training using only the adversarial loss is insufficient to recover such a simple data distribution. This result is quite unexpected compared to traditional GANs that can recover high dimensional complex image distributions from scratch.

We would like to stress here that the above observation does not render adversarial loss redundant in the current setup. In GAN training, depending on the loss function and the discriminator architecture, a good discriminator may not provide meaningful gradients to the generator [27], when the generator distribution is far away from the real distribution. Once the trajectories are close to real trajectories, the adversarial loss of Social GAN helps in spreading the probability distribution of the predicted trajectories. To corroborate this, we present another synthetic dataset comprising of 500 trajectories where the observed trajectory is a straight line which can diverge in one of three different directions. Similar to the previous setup, small gaussian noise $\mathcal{N}(0, 0.04)$ is added to the trajectories. The network is initially trained for a few epochs (20 in this case) with both losses combined [Fig.(3b)]. It is further trained using only the adversarial loss for 20 more epochs. We notice [Fig.(3c)] that the adversarial loss aids in spreading the trajectory distribution. The efficacy of adversarial loss is further highlighted in the real-world experiments section.

Discussion

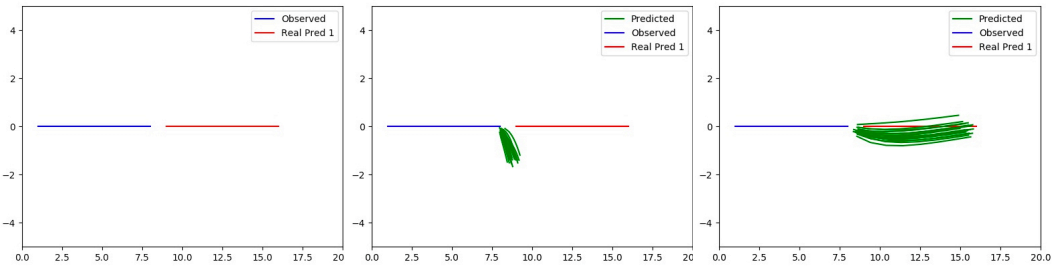
GANs are capable of generating highly complex distributions from randomly initialized weights. In the previous section we observed the failure of the adversarial loss to guide the GAN training on a simple straight line dataset. This observation adds credence to the fact that the GAN architecture of the Social GAN framework can be improved.

A particular reason for the above inefficacy could be the choice of discriminator architecture. The discriminator of Social GAN has a recurrent architecture. Training RNNs is difficult in comparison to feedforward networks. This fact combined with the instability in training GANs can lead to bad results. According to the proposed Social GAN architecture, the feedforward network inside the discriminator [Fig. 1] provides the direction of improvement to the hidden state of the fake trajectory rather than the trajectory itself. In such cases, gradients might vanish on backpropagating from the hidden state to the trajectory via the discriminator encoder. Such ill-gradients are ineffective in guiding the generator.

A plausible solution to this problem could be to change the discriminator architecture into a purely feedforward one. Fig. 4 explores this proposed solution on the synthetic straight line dataset. The performance of the feedforward discriminator on this dataset [Fig. 4c] indicates that

[2]We sample 20 output trajectories as mentioned in the Social GAN paper

the modified architecture holds promise in making the state-of-the-art model better.



(a) (b) (c)

Figure 4: Performance of a purely feedforward discriminator on the Straight Line dataset. (a) A sample from the dataset (b) Output trajectory distribution before training (c) Output trajectory distribution on training using a purely feedforward discriminator.

Real World Experiments

In this section, we first empirically corroborate that incorporating adversarial loss is advantageous in the task of human trajectory prediction. Later, we report preliminary results of the performance of Social GAN with a feedforward discriminator (denoted by FSGAN) in comparison to the default Social GAN (denoted by SGAN) on real world datasets. We evaluate our method on two publicly available datasets: ETH [28] and UCY [29]. These datasets consist of real world human trajectories with rich human-human interaction scenarios. All the data to real world coordinates and interpolate to obtain values at every 0.4 seconds. In total there are 5 sets of data (ETH - 2, UCY- 3) with 4 different scenes which consists of 1536 pedestrians in crowded settings with challenging scenarios like group behavior, people crossing each other, collision avoidance and groups forming and dispersing.

Evaluation Metrics. Similar to prior work [2, 5] we use two error metrics:

1. *Average Displacement Error (ADE)*: Average $L2$ distance between ground truth and our prediction over all predicted time steps.
2. *Final Displacement Error (FDE)*: The distance between the predicted final destination and the true final destination at end of the prediction period T_{pred}.

Evaluation Methodology. We follow similar evaluation methodology as [2, 5] to maintain uniformity in comparisons. We use leave-one-out approach, train on 4 sets and test on the remaining set. We observe the trajectory for 8 times steps (3.2 seconds) and show prediction results for 8 (3.2 seconds) and 12 (4.8 seconds) time steps.

Metrics	$\lambda \rightarrow$	1.0	0.5	0.1	0.05	0
ADE/FDE	ETH	0.96 / 1.97	0.93 / 1.91	**0.90 / 1.83**	1.22 / 2.44	4.20 / 8.35
	Hotel	0.59 / 1.22	**0.54 / 1.13**	0.57 / 1.18	0.61 / 1.27	2.94 / 5.65

Figure 5: Quantitative results of SGAN on varying λ in Eqn 4 (k=1) on ETH and Hotel dataset. We report (ADE / FDE) for $t_{pred} = 12$ in meters. We notice that using adversarial loss along with L2 loss leads to optimal performance proving the efficacy of adversarial loss (lower is better)

To experimentally verify that adversarial loss in conjunction with the variety loss provides an improved performance in task of human trajectory prediction, we train Social GAN on real world datasets using Eqn 4 and vary hyperparameter λ. We particularly choose $k = 1$ in this experiment

to demonstrate a performance boost over the commonly used L2 Loss.

We now compare our proposed feedforward discriminator GAN with Social GAN. Being consistent with the Social GAN paper, to capture multimodality, we output $k = 20$ trajectories. Table 2 provides a quantitative comparison between SGAN and FSGAN on ETH and UCY datasets. We notice that Social GAN with a feedforward discriminator performs as good as, if not better than, the default Social GAN model on almost all of the real world datasets.

Metrics	Model	ETH	Hotel	Univ	Zara1	Zara2
ADE	SGAN [5]	0.61 / 0.81	0.48 / 0.72	0.36 / 0.60	**0.21 / 0.34**	0.27 / 0.42
	FSGAN (ours)	**0.55 / 0.68**	**0.33 / 0.43**	**0.34 / 0.54**	0.22 / 0.35	**0.22 / 0.32**
FDE	SGAN [5]	1.22 / 1.52	0.95 / 1.61	0.75 / 1.26	**0.42 / 0.69**	0.54 / 0.84
	FSGAN (ours)	**1.05 / 1.16**	**0.65 / 0.89**	**0.69 / 1.14**	0.43 / 0.71	**0.45 / 0.67**

Table 2: Quantitative results of SGAN v/s FSGAN across real world datasets. We report two error metrics Average Displacement Error (ADE) and Final Displacement Error (FDE) for $t_{pred} = 8$ and $t_{pred} = 12$ (8 / 12) in meters. $k = 20$ trajectories sampled at test time (lower is better)

However, when k is large, variety loss can lead the model to an undesirable minima. Since the variety loss is calculated with respect to the predicted trajectory closest to the groundtruth, in complex non-linear human motion scenarios, the network can learn to generate k uniformly spread-out trajectories. Such a uniform multimodal output satisfies the objective of minimizing the defined variety loss. Thus, the variety loss (with a large k) can potentially prevent the network from learning the intricacies of human trajectories thereby failing to realize the goal of understanding human motion. We also compare the performances of both the models with a much smaller k ($k = 3$) to justify that FS-GAN does not learn to *better spread the trajectories* (see Table 3).

Metrics	Model	ETH	Hotel	Univ	Zara1	Zara2
ADE	SGAN [5]	0.69 / 0.87	0.40 / 0.53	0.34 / 0.55	0.25 / 0.42	0.22 / 0.35
	FSGAN (ours)	**0.66 / 0.84**	**0.39 / 0.52**	**0.34 / 0.54**	**0.25 / 0.40**	**0.22 / 0.33**
FDE	SGAN [5]	1.37 / 1.73	0.80 / 1.08	0.70 / **1.15**	**0.51** / 0.89	0.46 / 0.74
	FSGAN (ours)	**1.33 / 1.68**	**0.76 / 1.08**	**0.69** / 1.16	0.52 / **0.85**	**0.45 / 0.71**

Table 3: Quantitative results of SGAN v/s FSGAN across real world datasets. We report two error metrics Average Displacement Error (ADE) and Final Displacement Error (FDE) for $t_{pred} = 8$ and $t_{pred} = 12$ (8 / 12) in meters. $k = 3$ trajectories sampled at test time (lower is better)

Conclusion and Future Work

In this work, we presented a detailed study of the effect of adversarial loss on the task of human trajectory prediction. We came across an unusual shortcoming in the state-of-the-art human trajectory prediction network. We provided a plausible explanation regarding the unexpected observation followed by a proposition to overcome this limitation. We showed the efficacy of the proposed modification on our synthetic dataset and preliminary results on real world datasets indicate room for improvement of state-of-the-art. Finding the optimal feedforward discriminator architecture to handle real world trajectory datasets is a work in progress.

References

[1] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical Review E*, 51, 05 1998.

[2] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–971, 2016.

[3] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher Bongsoo Choy, Philip H. S. Torr, and Manmohan Krishna Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2165–2174, 2017.

[4] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

[5] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2255–2264, 2018.

[6] Matthias Luber, Johannes Andreas Stork, Gian Diego Tipaldi, and Kai Oliver Arras. People tracking with human motion predictions from social forces. *2010 IEEE International Conference on Robotics and Automation*, pages 464–469, 2010.

[7] Alexandre Alahi, Vignesh Ramanathan, and Li Fei-Fei. Socially-aware large-scale crowd forecasting. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2211–2218, 2014.

[8] Alex Graves. Generating sequences with recurrent neural networks. *CoRR*, abs/1308.0850, 2013.

[9] Rudolf E. Kálmán. A new approach to linear filtering and prediction problems" transaction of the asme journal of basic. 1960.

[10] P. McCullagh and J.A. Nelder. *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall, 1989.

[11] Christopher K. I. Williams. Prediction with gaussian processes: from linear regression to linear prediction and beyond. 1997.

[12] Carl E. Rasmussen. Gaussian processes for machine learning. In *Advanced Lectures on Machine Learning*, 2009.

[13] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *ECCV*, 2016.

[14] Shuai Yi, Hongsheng Li, and Xiaogang Wang. Understanding pedestrian behaviors from stationary crowd groups. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3488–3496, 2015.

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9:1735–1780, 1997.

[16] Junyoung Chung, Çaglar Gülçehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555, 2014.

[17] Alex Graves, Abdel rahman Mohamed, and Geoffrey E. Hinton. Speech recognition with deep recurrent neural networks. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649, 2013.

[18] Chunshui Cao, Xianming Liu, Yi Yang, Yinan Yu, Jiang Wang, Zilei Wang, Yongzhen Huang, Liang Wang, Chang Huang, Wei Xu, Deva Ramanan, and Thomas S. Huang. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2956–2964, 2015.

[19] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, 2015.

[20] Fandong Meng. Neural machine translation by jointly learning to align and translate. 2014.

[21] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large Scale GAN Training for High Fidelity Natural Image Synthesis. *arXiv:1809.11096 [cs, stat]*, September 2018. arXiv: 1809.11096.

[22] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv:1511.06434 [cs]*, November 2015. arXiv: 1511.06434.

[23] Yutian Chen, Matthew W. Hoffman, Sergio Gomez Colmenarejo, Misha Denil, Timothy P. Lillicrap, Matt Botvinick, and Nando de Freitas. Learning to Learn without Gradient Descent by Gradient Descent. *arXiv:1611.03824 [cs, stat]*, November 2016. arXiv: 1611.03824.

[24] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-Resolution Image Synthesis and Semantic Manipulation With Conditional GANs. pages 8798–8807, 2018.

[25] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved Techniques for Training GANs. *arXiv:1606.03498 [cs]*, June 2016. arXiv: 1606.03498.

[26] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved Training of Wasserstein GANs. *arXiv:1704.00028 [cs, stat]*, March 2017. arXiv: 1704.00028.

[27] Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *CoRR*, abs/1701.07875, 2017.

[28] Stefano Pellegrini, Andreas Ess, and Luc Van Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. In *ECCV*, 2010.

[29] Laura Leal-Taixé, Michele Fenzi, Alina Kuznetsova, Bodo Rosenhahn, and Silvio Savarese. Learning an image-based motion context for multiple people tracking. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3542–3549, 2014.