electronics

Article

D-STGCN: Dynamic Pedestrian Trajectory Prediction Using Spatio-Temporal Graph Convolutional Networks

Bogdan Ilie Sighencea, Ion Rareș Stanciu and Cătălin Daniel Căleanu *

Department of Applied Electronics, Faculty of Electronics, Telecommunications and Information Technologies, Politehnica University Timișoara, 300223 Timișoara, Romania
* Correspondence: catalin.caleanu@upt.ro

Abstract: Predicting pedestrian trajectories in urban scenarios is a challenging task that has a wide range of applications, from video surveillance to autonomous driving. The task is difficult since pedestrian behavior is affected by both their individual path's history, their interactions with others, and with the environment. For predicting pedestrian trajectories, an attention-based interaction-aware spatio-temporal graph neural network is introduced. This paper introduces an approach based on two components: a spatial graph neural network (SGNN) for interaction-modeling and a temporal graph neural network (TGNN) for motion feature extraction. The SGNN uses an attention method to periodically collect spatial interactions between all pedestrians. The TGNN employs an attention method as well, this time to collect each pedestrian's temporal motion pattern. Finally, in the graph's temporal dimension characteristics, a time-extrapolator convolutional neural network (CNN) is employed to predict the trajectories. Using a lower variable size (data and model) and a better accuracy, the proposed method is compact, efficient, and better than the one represented by the social-STGCNN. Moreover, using three video surveillance datasets (ETH, UCY, and SDD), D-STGCN achieves better experimental results considering the average displacement error (ADE) and final displacement error (FDE) metrics, in addition to predicting more social trajectories.

Keywords: pedestrian trajectory prediction; deep learning; social interactions; graph neural networks

Citation: Sighencea, B.I.; Stanciu, I.R.; Căleanu, C.D. D-STGCN: Dynamic Pedestrian Trajectory Prediction Using Spatio-Temporal Graph Convolutional Networks. *Electronics* **2023**, *12*, 611. https://doi.org/10.3390/electronics12030611

Academic Editors: Tianfei Zhou, Xiankai Lu and Wenguan Wang

Received: 21 December 2022
Revised: 20 January 2023
Accepted: 24 January 2023
Published: 26 January 2023

1. Introduction

To drive safely in most traffic scenarios, including crowded places full of vehicles and people, autonomous vehicles must accurately predict and track the future behaviors of the surrounding interacting agents [1]. For a short-term prediction, it may be acceptable to utilize pure physics-based approaches. However, because future scenarios are unknown, a long-term prediction system is essential to enable not only interaction modeling between various agents but also to figure out the traversable regions defined by road layouts and right-of-way compliance with traffic rules.

Pedestrians are responsible for 23% of the 1.35 million road deaths worldwide every year [2]. In Organization for Economic Co-operation and Development (OECD) countries, more than 20,000 pedestrians lose their lives annually. Pedestrian deaths represent between 8% and 37% of all traffic fatalities, depending on the country and year [3]. Most of these tragic accidents occur in crowded areas at pedestrian crossings with a poor visibility due to low driver attention and/or fatigue. According to [4], the number of elderly pedestrian accidents is influenced by multiple factors in the city environment; it is important to mention that pedestrians have minimal protection. As a result, reducing (or eliminating) these collisions is an important safety concern. In these situations, helping the driver includes predicting pedestrian behavior. This helps to reduce the effect of various factors that could negatively affect traffic safety (such as fatigue, poor visibility, inadvertent cognitive distraction, etc.). According to [5], the implementation of these new technologies is estimated to reduce the number of accidents by up to 93.5%.

Despite its importance, forecasting trajectories is extremely challenging due to the inherent erratic human's characteristics. First off, the pedestrian motion is multimodal, which indicates that many different socially acceptable future paths might emerge from the same trajectory history. With many different personal habits, the associated motion is sometimes difficult to define. Second, human motion is strongly influenced by the individuals around them, as can be seen in Figure 1. The interactions between pedestrians can cause someone to walk similar paths inside a group or modify direction/speed to prevent collisions. In practice, the joint modeling of complex social interactions is also challenging. Third, the scenario is highly dynamic because people constantly move in a scene and out of it, so each frame may be different. The dynamic feature requires powerful algorithms to manage a variable number of traffic actors.

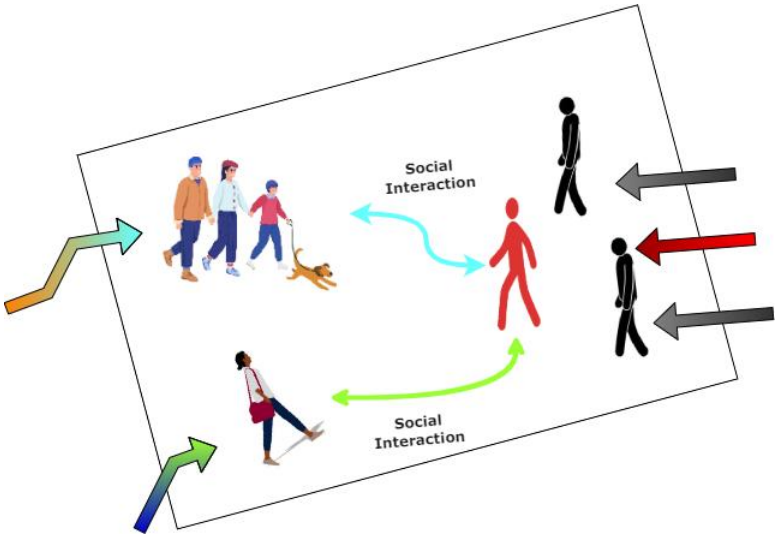


Figure 1. Pedestrians consider the future impacts of dynamic people involved in the scene when trying to make navigation decisions. The future trajectories of pedestrians need to be mentioned in a completely random way. As a result, both the spatial and temporal structure of the pedestrian-environment interaction must be taken into account.

The pedestrian trajectory prediction topic was studied in many contexts, but the previously proposed algorithms still suffer due to the constraints described below: first off, most of the current approaches encoded and decoded the recorded trajectories with a specific resolution (i.e., a defined number of iterations), which is unable to fully utilize the temporal connections of the movement behavior. Modeling the general concept (e.g., the destination of the pedestrian and their current position) and the local context (e.g., the orientation and speed of the motion at a particular time) with a single resolution is ineffective. Second, many previous approaches used aggregation mechanisms, such as pooling. This uses the pedestrian trajectory modelled by recurrent units. These are used as inputs to model human interactions by learning the scene's general representations.

On sidewalks and crosswalks, pedestrians can suddenly change direction and their movement should consider interactions with others. According to [6], 70% of pedestrians walk in groups. Most human interactions follow logic and social rules.

These interactions are not always homogeneous and are likely to be asymmetric. For example, a pedestrian in the front will have a greater impact than one in the back. To represent these dynamic interactions, this approach uses a directed graph to organize the trajectories of all pedestrians in a scenario. The nodes contain motion properties, such as the desired velocity across a defined time window. Meanwhile, the edges between neighboring nodes contain spatial interaction.

A pedestrian's trajectory is affected by a variety of conditions, including the number of obstacles, the proximity to other traffic actors, their objective, and personality. On the

other hand, the movement of surrounding pedestrians is an obvious and important aspect that affects the trajectories [7]. Modern data-driven methods [8] use a pooling/aggregation process to model the influences of neighbors, which are typically chosen based on the fixed spatial size of the graph grid.

In this paper, a stable and computationally efficient solution is introduced. The primary goal of the proposed method is to predict the trajectory of all pedestrians in the scene in a single shot. This is extremely important for intelligent mobility solutions (e.g., automated guided robots, autonomous driving vehicles, etc.) because trajectory prediction determines traffic safety. A real-time operation with minimal computation resources is essentially important.

The relevance of hidden neural states is difficult to comprehend. This also makes it more challenging to extract social relationships from hidden layers. Instead of extracting these features, the social-spatio-temporal graph convolutional neural network (social-STGCNN) uses graph convolutional networks (GCNs) with a hand-crafted kernel to capture them directly from their position information, as described by Mohamed et al. [9].

Previous works in the pedestrian trajectory prediction (PTP) problem were mainly dominated by RNN and CNN approaches. They are briefly described below.

1.1. Trajectory Prediction Methods with RNNs

Due to its intrinsic capability in dealing with spatio-temporal data, many researchers prefer long short-term memory (LSTM) architecture. This is an enhanced variant of the vanilla RNN model. Alahi et al. [8] proposed a social-LSTM to extract the state of the movement characteristic of each pedestrian using three popular datasets: ETH [10], UCY [11], and the Stanford drone dataset (SDD) [12]. Additionally, a 'social pooling' layer is used to analyze the interactions of traffic participants, the result being called a social-LSTM network. Taking into account bivariate Gaussian distributions, the social-LSTM model successfully estimates the stochastic trajectories.

Recent research has improved the social-LSTM method [8] by creating an interaction aggregation module or using image-derived environment and neighbor features. To improve the prediction accuracy, Fernando et al. [13] used a dual attention (a 'soft' and a 'hardwired') model for LSTM. Xue et al. [14] proposed embedding the scene information in the LSTM model. Their initial social-scene-LSTM proposal used two extra layers to model the interaction of nearby people and the environment. This architecture proposes a circular-shaped neighborhood which outperforms the traditional rectangular one by considering the person, social, and scene scales integrated into a hierarchical LSTM structure. An LSTM refinement was proposed in [15,16]. This solution enables the employment of the current participants' group intention through a message passing mechanism.

Another approach used is structural-RNN [17]. This represents a spatio-temporal LSTM graph-based implementation. This method shows a solution that merges edgeRNNs and nodeRNNs in the form of a bipartite graph in a variety of interactions, for example, between people and static objects. This approach showed better results in generalizing the mutual influence of pedestrians in crowd modeling.

Some researchers argue that if a previous trajectory is provided and pedestrian routes are analyzed in the context of a multimodal distribution, many paths are realistic and socially acceptable. Gupta et al. [18] proposed social-GAN, which employs generative adversarial networks (GANs) with an LSTM-based generator. Many other researchers began to improve on this study once it was published. According to Amirian et al. [19], the social networks reveal an improvement in the attention pooling architecture rather than the maximum pooling structure and use an info-GAN without L2-loss. Sadeghian et al. [20] proposed SoPhie, which includes the scene characteristics to improve the attention mechanism obtained from images using a CNN-based feature extraction module and an LSTM-based GAN architecture. The approach described in [21] uses contextual information that affects the pedestrian trajectories. Using the recorded human-contextual interaction and the latent variation mode, their method deals with uncertainties in the estimated trajectories.

1.2. Trajectory Prediction Methods with CNNs

CNNs were used to estimate the trajectory and extract information from the images. Compared to RNN solutions, which autoregressively predict the future path time steps, CNN-based methods can estimate all the location time steps in an instant. Nikhil and Morris [22] used CNN to estimate the trajectories. They demonstrated a competitive performance while maintaining computational efficiency. The multi agent tensor fusion generative adversarial network (MATF-GAN) [23] created a social tensor that encodes the entire static scene's context and agent's trajectory information. Although MATF-GAN considered the global spatial structure of agents in the final observed positions, it did not take into account the temporal dynamics across the trajectories and did not overlook the spatial distribution of the agents' path.

Chandra et al. [24] presented a method that uses CNN to extract data from the images independent of the perceptions of the horizon and proximity and mix it with the ego state to integrate it into the LSTM architecture. This method performed well (ADE = 0.78, FDE = 2.44 for the TRAF dataset) in predicting future pedestrian routes. Peek into the future (PIF) [25] collects a visual input from a scene using CNN and integrates the movement data into an LSTM architecture to estimate the pedestrian trajectories. The authors initially employed a pre-trained image segmentation model to extract the pixel-level scene classes; in this work, a total of ten common scene classes including roads, sidewalks, etc., are considered. Before making decisions, agents always collect and evaluate the scene information within a certain range of locations. Segmentation maps are insufficient to fully represent their future actions.

Trajectron++ [26] developed a recurrent model (graph-structured) predicting the trajectories of a varying pedestrian number while considering various scene inputs, including the human dynamics. This approach uses then the graph's structure to model the nodes interaction (i.e., agents) in a scene. To prevent obstacle collisions, it encodes a local map that represents the agent's interaction and the environment. To outperform other complex models, Zamboni et al. [27] proposed novel data augmentation and preprocessing methods. Their model produces the deterministic projections of future states without measuring the uncertainties. In complex traffic circumstances, the predicted trajectories are not always accurate.

Considering the prediction accuracy, the data efficiency, and the component size, the introduced model outperforms previous architectures. The following are the main contributions of the presented work:

- An improved trajectory prediction by increasing the number of convolutional layers in the model, resulting in a significant improvement generated in the social-STGCNN [9] architecture.
- Performed a comparison experiment using three trajectory datasets: ETH [10], UCY [11], and the Stanford drone dataset (SDD) [12] collected in different urban scenarios, to assess the effectiveness of the introduced method versus state-of-the-art architectures.
- Performed an ablation study to reveal the contribution of different components of the presented model on the prediction problem and to validate the performance of the attention-based ST-GCNN and TXP-CNN.
- Illustrated the qualitative results visualization on different scenarios from ETH, UCY, and SDD datasets.

The rest of the article is organized into six main sections, each with extensive subsections. In Section 2, a brief overview of other relevant research works is presented. Section 3 concerns the formulation of the problems. Section 4 describes the model architecture of the proposed method. Section 5 contains information about the proposed methodology, including the implementation details, datasets, and evaluation metrics. Section 6 provides the ablation study, the comparison of the presented method with the state-of-the-art research, and the visualization of various predicted trajectories in different scenes. Finally, Section 7 presents the observations and conclusions.

2. Related Work

In recent years, in the context of smart vehicles and autonomous driving, the pedestrian trajectory prediction problem has become an important aspect. However, due to their varying interactions, predicting trajectories is challenging. Researchers are studying this problem and are generating predictions under uncertainty. Deep learning has recently demonstrated the potential of learning the human movement patterns using a data-driven approach. In the following, an overview of the relevant work in the field of PTP is provided.

The PTP field has attracted a lot of research interest in the scientific community. The proposed methods are focused on the following four areas: modeling social interactions, the physical limitations of real-world scenes, the joint representation of available cues, and multimodal predictions.

The graph neural network is a deep learning model that is applied directly to graph architectures. It effectively includes relational inductive bias into the model's design. In the context of GNNs, most graphs are attributed (with nodes and edges attributes, and/or global characteristics). In graph representation learning using GNNs, there are three main operations: nodes, edges, and global updates [28].

Graph convolutional neural networks (GCNNs) can be divided into spectral [29–31] and spatial methods [32,33]. The first uses a spectral representation of the graphs to create convolutions, whereas the latter defines on-the-graph convolutions, working on a spatially close group of neighbors. The learned filters of the spectral methods are defined by a Laplacian eigen basis, which is determined by the graph's structure. As a result, a model trained on a particular structure cannot be immediately transferred to a structurally different graph. However, modeling human social interaction requires a time-variant graph. As a result, spectral methods are ineffective in forecasting pedestrian trajectories. Therefore, the proposed method belongs to the category of spatial solutions.

Velickovi et al. [34] introduced so-called graph attention networks (GATs). This allows for an integration of a 'self-attention-based' architecture into any type of data structure characterized as a graph. These networks improve the features of GCNNs [35] by allowing the model to assign each network node a dynamic value directly. Kosaraju et al. [36] proposed social-BiGAT, which generates social interactions between pedestrians as a graph and provides higher edge weights for the most important interactions. The so-called agent-former learns representations of social interaction from both the temporal and social dimensions, arguing that the representation of the time and social elements separately may lead to a suboptimal solution.

Mohamed et al. [9] combined pedestrian crossing features from their position data directly using the GCNN and a handcrafted kernel. Their approach (social-STGCNN) improves the STGAT method [37] by obtaining data from each pedestrian trajectory, modeling the interactions between pedestrians with the help of a GNN, and calculating a weighted graph using GAT. In [38], the authors defined the task of predicting the trajectory of pedestrians using a spatio-temporal graph, with nodes representing humans in the crowd.

In [39], the authors designed a conditional generative neural system (CGNS) to estimate future vehicle trajectories for a safe and intelligent navigation. However, due to an unsupervised learning method and unstable training, it is difficult to expect good results based on this solution.

The reader can also consult [40] for a more extensive overview of the deep learning-based approaches used to predict the trajectories of the pedestrian.

Similarly to [34], in our proposed method, pedestrian interactions are represented as a graph, with nodes designating people and edges indicating interactions; higher edge weights correspond to strong interactions. A weighted adjacency matrix, which is a representation of the characteristics of the graph's edge, is also utilized to determine the influence between people.

3. Problem Description

The main objective of trajectory prediction is to estimate the future positions of a group of N pedestrians in a real-world scene. This is based on their past and current positions on a map representation throughout a time interval, starting at moment T_o, as can be seen in Figure 2. The goal is to estimate the coordinates of each pedestrian after T_p time steps. The position of each pedestrian in a scene is represented by the real-world coordinate $X = (x, y)$. The ith past position for each human is identified as $X_t^i = (x_t^i, y_t^i)$ with $t \in \{1, \ldots, T_o\}$, where (x_t^i, y_t^i) are the random variables describing the probability distribution of the pedestrian n location at moment t within the scene (n is scene-dependent). The ground truth of the future trajectories is denoted as $Traj_{obs} = \{Y_t^i = (x_t^i, y_t^i)\}$, where $i \in \{1, \ldots, n\}, T_o + 1 \leq t \leq T_p$.

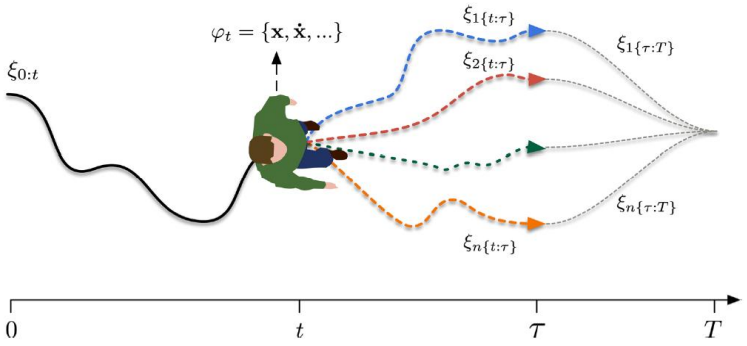


Figure 2. Time distribution of each pedestrian position starting from the moment T_o until T_p. Here, three positions can be mentioned: observed positions, real-future positions, and predicted positions.

The predicted positions denoted $Traj_{pred} = \{\hat{Y}_t^i = (\hat{x}_t^i, \hat{y}_t^i)\}$, $i \in \{1, \ldots, n\}$, $T_o + 1 \leq t \leq T_p$ are the random variables. It is assumed that the position of the i^{th} pedestrian at moment t follows bi-variate Gaussian distribution $\hat{Y}_t^i \sim \mathcal{N}(\mu_t^i, \sigma_t^i, \rho_t^i)$. At time t, the pedestrian's group's center is denoted as $\mu_t^i = (\mu_x, \mu_y)_t^i$. The standard deviation is referred to as $\sigma_t^i = (\sigma_x, \sigma_y)_t^i$, and the correlation coefficient is denoted as ρ_t^i. To obtain the trajectory prediction, the proposed architecture predicts the Gaussian distribution parameters $(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)_t^i$.

The pedestrian positions of the observation time moments $1 \leq t \leq T_o$ are utilized to estimate the position's moments $T_o + 1 \leq t \leq T_p$ for each traffic participant. To learn the parameters of the model, the negative logarithmic likelihood loss function is used, as shown in Equation (1).

$$L^i(W) = -\sum_{t=1}^{T_p} \log\Big(f(x_t^i, y_t^i \big| \mu_x, \mu_y, \sigma_x, \sigma_y, \rho\Big)) \tag{1}$$

In Equation (1), W denotes the learned network parameters. The loss value is minimized to obtain the optimal weights of the network.

4. The Proposed Method

In the following, a description of the key aspects of the proposed estimation method is provided. Subsequently, the complete design of each module will be discussed, including the overall creation of the model. The main goal is to forecast future trajectories for numerous interacting agents using position history and context information. The prediction method is also extendable to multi-target tracking frameworks.

Compared to social-STGCNN [9], in this paper, the layers dimensions of both CNN modules were adjusted to increase the accuracy of the estimation. The original method includes two components: the spatio-temporal graph convolution neural network (ST-GCNN) and the time-extrapolator convolution neural network (TXP-CNN). The first component

uses convolution to extract the features. These characteristics provide a concise description of the history of the observed trajectories. These are fed to the TXP-CNN which employs them to predict positions for all pedestrians in the group and to extrapolate the future trajectories. The overview diagram of this method is depicted in Figure 3.

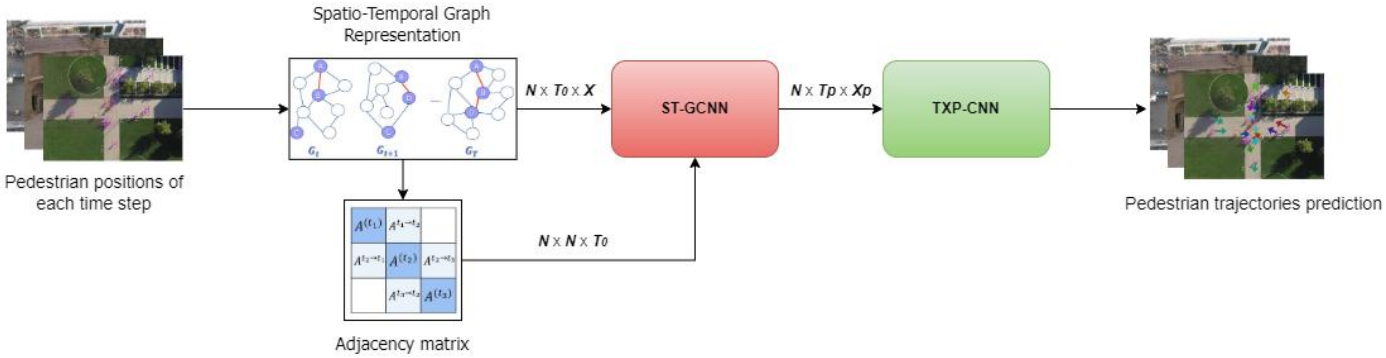


Figure 3. Overall architecture of the proposed method. Through optimization of the layer size, increased accuracy in trajectory prediction is also possible.

4.1. Graph Representation of Pedestrian Trajectory

The graph is formed by the relative positions of the pedestrian trajectories and captures the spatial interactions of pedestrians in the real-world scene. Considering N pedestrians at moment t, their trajectories are represented as the graph $G_t = (V_t, E_t)$. Here, $V_t = \{\{v_t^n | \forall_n \in \{1, \ldots, N\}\}$ is a node-set that corresponds to the pedestrians' positions. Each observed location (x_t^n, y_t^n) is a point of v_t^n. $E_t = \left\{ e_t^{nj} \middle| \forall n,\ j \in \{1, \ldots, N\} \right\}$ represents the set of edges within graph G_t. The vector e_t^{nj} emphasizes the edge between v_t^n and v_t^j (n and j denote pedestrians) and is reflected by an $N \times N$ adjacency matrix A_t^m for each moment. A_t^m is defined in Equation (2).

$$A_t^m(n,\ j) = \begin{cases} 1/\|v_t^n - v_t^j\|_2 & if\ \|v_t^n - v_t^j\|_2 \neq 0, \\ 0 & Otherwise \end{cases} \tag{2}$$

4.2. Spatio-Temporal Graph Convolutional Neural Network

The first component is the ST-GCNN used to incorporate the object's representation [9]. The ST-GCNN function is to extract the spatio-temporal embedding from the input graph. The next step is the normalization of matrix A_s, representing the network topology at time step t, for generating the Laplacian matrix, as can be seen in Equation (3):

$$\hat{A}_t^s = \Lambda_t^{-\frac{1}{2}} (A_t^s + I) \Lambda_t^{-\frac{1}{2}}\ , \tag{3}$$

where node links are considered by adding the identity matrix I to the adjacency A_t^s and Λ_t is the diagonal degree matrix with components reflecting the row summations of $A_t^s + I$. The ST-GCNN is further specified by convoluting the velocity map $V(\imath)$ with the kernel $W(\imath)$ at layer $\imath$ under the graph Laplacian $\hat{A}_t^s$ as can be seen in Equation (4):

$$f(V(\imath), A) = \sigma\left(\Lambda_t^{-\frac{1}{2}} \hat{A}_t^s \Lambda_t^{-\frac{1}{2}} V(\imath) W(\imath) \right). \tag{4}$$

4.3. Time-Extrapolator Convolutional Neural Network

The second component employed is the TXP-CNN. This works on the temporal dimension of the graph embedding $\hat{V}$ and it increases to meet the demand of the prediction accuracy. Since it relies on feature space convolutions, it has fewer parameters than the recurrent units. It is important to note that the TXP-CNN layer is permutation-variant,

since variations in the graph embedding lead to different outcomes. Aside from that, the predictions are invariant if the order of the pedestrians is permuted.

The main objective of the network time extrapolator is to perform temporal convolutions in the trajectory's history to decode future locations. This is because the temporal convolution network (TCN) [41] is considered a more powerful and efficient system for learning time dependencies than the recurrent architectures. Each temporal layer is residually interconnected with the previous one.

In this paper, the entire process of the prediction of the pedestrian trajectory is illustrated by Algorithm 1.

Algorithm 1. Dynamic Graph Convolutional Networks

Input: real-world pedestrians coordinate $X = (x, y)$;
Output: the evaluation metrics, i.e., the average displacement error (ADE) and final displacement error (FDE);

For each $t \in [1,\ T_o]$ **do**
  Represent trajectories as a graph : $G_t = (V_t, E_t)$;
  Compute ground truth of future trajectories : $Traj_{obs} = \left\{ Y_t^i = \left(x_t^i, y_t^i \right) \right\}$;
End for
Create 'distribution distance' $N \times N$ adjacency matrix A_t^m by using Equation (2);
Generate Laplacian matrix by using Equation (3);
For each $i \in \{1, \ldots,\ n\}$ **do**
  For all $t \in [1,\ T_o]$ **do**
    Learn the parameters $\left(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho\right)$ of the model by using Equation (1);
    probability distribution of the predicted trajectory : $\hat{Y}_t^i = \left(\hat{x}_t^i, \hat{y}_t^i \right) \ \sim \mathcal{N}\left(\mu_t^i, \sigma_t^i, \rho_t^i \right)$;
  End for
End for
Collect all of the predicted location and the real location for each pedestrian;
Compute the ADE and FDE with the formulas from Equations (5) and (6);
Return ADE and FDE.

5. Experiments

5.1. Implementation Details

The PyTorch deep learning framework was used for the implementation of the proposed solution. For training and evaluation, a Nvidia GeForce GTX 1080 GPU with 8 GB memory RAM on a desktop running Ubuntu Version 20.04 was used. The model was trained for 250 epochs with a batch size set to 128. The stochastic gradient descent (SGD) algorithm was utilized as a neural network optimizer. The learning rate was initially set at 0.01 while the decay was set at 0.002 after 150 epochs, while PRELU [42] was used as an activation function. All the hyperparameters were determined empirically following a trial-and-error approach.

5.2. Datasets

For comparison purposes, the proposed method was first evaluated on the following traditional datasets: ETH [10] and UCY [11]. Furthermore, the experimental results for the more challenging dataset, SDD [12], are also provided. These datasets contain videos and annotated trajectories that include social interactions in real-world scenarios. All the paths are represented in world coordinates. Unprocessed pedestrian positions for these datasets were obtained from the S-GAN-P repository [18], which used them to calculate the relative location.

ETH includes two scenes, known as ETH and HOTEL. UCY includes three scenes, identified as UNIV, ZARA1, and ZARA2. In total, there are around 1500 pedestrians, including thousands of trajectories in different scenarios, such as creating groups, crossing together, and walking alongside each other.

The SDD contains a bird's-eye view from the drone recordings of eight different locations and thirty-one videos recorded on the Stanford University campus. It includes six annotated classes, such as vehicles, pedestrians, cyclists, buses, carts, and skateboarders. SDD has eight scenes, including *Bookstore*, *Coupa*, *DeathCircle*, *Gates*, *Hyang*, *Little*, *Nexus*, and *Quad*. The annotated data produce 2D bounding box coordinates (measured in pixels) in video sequences available at 30 frames per s. While pedestrians and bicycles are visible in every scene, fast-moving categories have a significant impact on pedestrians. The decision to consider this dataset is related to the existence of difficult-to-model multiple dynamic patterns.

5.3. Evaluation Metrics

In computer vision research, trajectories are commonly described by motion statistics such as the number of collisions, the average acceleration, the average speed, and the total distance covered [19]. There are eight observed time steps (3.2 s) for each traffic participant. For all datasets, a total of twelve steps (4.8 s) are used to represent the real future positions. Euclidian L2 norms are employed to evaluate the displacements between the ground truth and the estimated trajectories. The evaluations are based on two common metrics: the average displacement error (ADE) and final displacement error (FDE). These are briefly described in the following.

- The ADE represents the average distance between the ground truth and the predicted trajectories over all future time steps, as shown in Equation (5).

$$ADE = \frac{\sum_{n=0}^{N} \sum_{t=0}^{T_p} \| \hat{p}_t^n - p_t^n \|_2}{N \times T_p}, \tag{5}$$

- The FDE represents the distance between the final positions of the ground truth and the predicted trajectories, as shown in Equation (6).

$$FDE = \frac{\sum_{n=0}^{N} \| \hat{p}_t^n - p_t^n \|_2}{N}, \; t = T_p, \tag{6}$$

  where N denotes the number of pedestrians, T_p designates the number of the predicted time step, and p_t^n and $\hat{p}_t^n$ are the ground truth and predicted result at time step t, respectively.

Many of the proposed solutions [43–45] incorporate multimodality in their estimation task, causing their models to produce more than one prediction given a single past trajectory. This proposed method reports the results using the minimum ADE/FDE among the randomly sampled k number of predicted positions, where $k = 20$ in all the tested scenes from the above-mentioned datasets.

6. Results

6.1. Ablation Study

D-STGCN is created by stacking several layers of ST-GCNN and TXP-CNN. In general, the selection of the network dimension can have a major impact on the ideal results. To identify the most efficient architecture, multiple combinations of the ST-GCNN and TXP-CNN number of layers were tested based on the design-of-experiment (DOE) method comparable to [46].

Each experiment was carried out with the same hyperparameter settings in both the testing and training phases. Taking into account the complexity of the model, the maximum number of ST-GCNN and TXP-CNN is established at four. The metrics described in Section 5.3 were used for evaluation purposes. The empirical results are detailed in Table 1.

Table 1. Results of different parameter combinations applied on the ETH, UCY, and SDD datasets. The results are shown in terms of the average ADE/FDE metrics. The AWG column represents the average ADE/FDE results for all scenes of the ETH-UCY datasets. The best results are indicated in bold. Lower numerical results are better.

ST-GCNN	TXP-CNN	SDD	ETH	HOTEL	UNIV	ZARA1	ZARA2	AWG
1	1	17.03/28.32	0.68/1.32	0.52/0.86	0.47/0.83	0.41/0.61	0.34/0.54	0.48/0.83
1	2	20.47/36.21	0.72/1.23	0.54/1.01	0.47/0.83	0.38/0.64	0.32/0.50	0.48/0.84
1	3	18.98/29.46	**0.63**/**1.03**	0.40/0.65	0.50/0.89	0.37/0.60	0.32/0.50	**0.44**/**0.73**
1	4	19.69/34.28	0.74/1.26	**0.37**/**0.58**	0.47/0.85	**0.35**/0.57	**0.29**/**0.48**	0.44/0.74
2	1	15.82/**25.50**	0.69/1.33	0.58/0.91	**0.46**/**0.78**	0.38/**0.56**	0.34/0.51	0.49/0.81
2	2	18.65/31.05	0.99/1.80	0.52/0.93	0.52/0.96	0.37/0.58	0.37/0.57	0.55/0.96
2	3	24.32/44.00	0.77/1.50	0.43/0.72	0.50/0.92	0.36/0.59	0.36/0.56	0.48/0.85
2	4	17.56/31.53	0.81/1.64	0.67/1.23	0.50/0.90	0.38/0.62	0.34/0.55	0.54/0.98
3	1	25.09/29.93	0.69/1.34	0.53/0.91	0.62/1.10	0.42/0.70	0.39/0.58	0.53/0.92
3	2	**15.18**/25.93	0.70/1.26	0.58/0.97	0.57/1.03	0.44/0.67	0.35/0.53	0.52/0.89
3	3	43.46/62.67	0.77/1.34	0.66/1.22	0.49/0.88	0.40/0.60	0.39/0.56	0.54/0.92
3	4	23.86/42.04	0.73/1.42	0.48/0.78	0.51/0.96	0.44/0.70	0.32/0.53	0.49/0.82
4	1	18.65/31.84	0.99/1.74	0.53/0.74	0.57/0.95	0.50/0.86	0.35/0.53	0.58/0.96
4	2	21.54/31.21	0.94/1.94	0.64/1.03	0.53/0.84	0.51/0.89	0.35/0.51	0.59/1.04
4	3	19.18/33.29	0.75/1.29	1.13/2.04	0.53/0.98	0.39/0.63	0.39/0.55	0.63/1.09
4	4	26.88/43.56	0.71/1.25	0.89/1.56	0.56/0.99	0.45/0.69	0.35/0.54	0.59/1.00

This study concluded that the best architecture model consists of:

- One ST-GCNN layer and three TXP-CNN layers for the ETH-UCY datasets.
- Three ST-GCNN layers and two TXP-CNN layers for the SDD dataset.

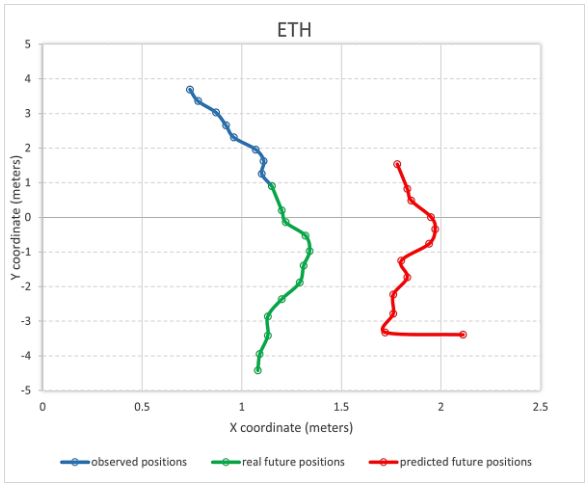
As the number of D-STGCN layers increases, the performance of the model decreases (i.e., a combination of three and four layers). This behavior is caused by a lack of scene vision data for the model inputs.

6.2. Visualization

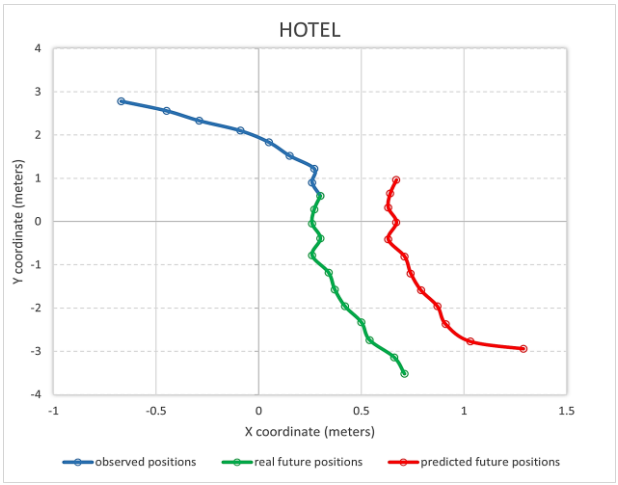
The prediction of the proposed method is qualitatively analyzed for five different real scenes from the ETH and UCY datasets. The visualization results are illustrated in Figure 4. The plots depicted here show that pedestrians pay attention to their surroundings. People pay more attention to the nearby humans and obstacles in front of them than to those behind them. Changes in the context behind a pedestrian or far in front of him have little impact on future movement decisions. The observed pedestrian positions (the known eight-point trajectory) are represented by the blue segments between two consecutive locations. The real trajectory (observed) is represented in green (twelve positions). The prediction, starting from position nine (for the next twelve points), is represented in red. All locations are depicted in a 2D plot (x, y coordinates) in meters.

Figure 4a represents a road crossing scene. It includes pedestrian interactions and an obstacle avoidance scenario in the ETH scene. This figure shows the behavior of the algorithm path prediction algorithm in the absence of information from social interactions.

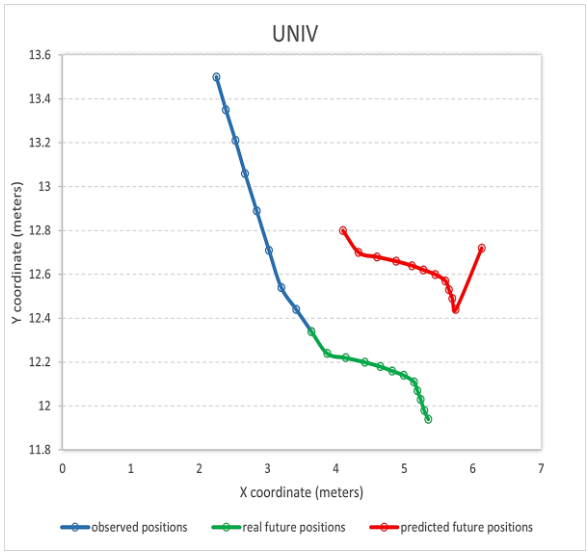
Figure 4b shows the qualitative model prediction evaluation for a pedestrian interaction in the HOTEL scene scenario. For example, solutions such as S-GAN-P [18] and SoPhie [20] use simple social information to pool data; therefore, the prediction results deviate significantly from the real trajectory. The proposed model captures long-term social information using the spatial-temporal graph; thus, the results are more likely to correspond to the real future paths.



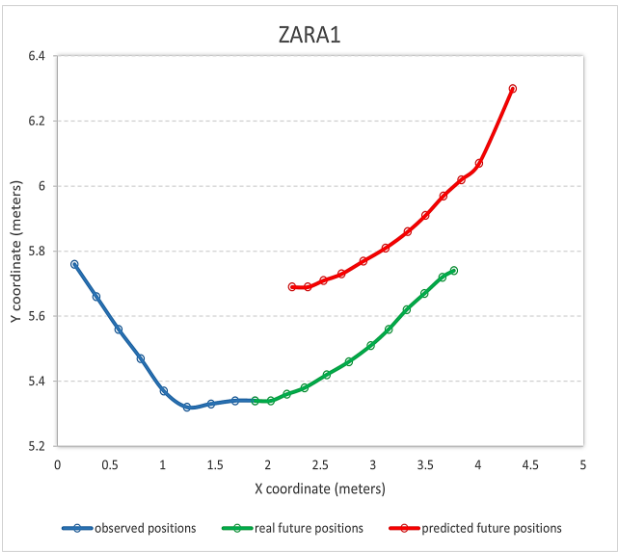
(**a**)



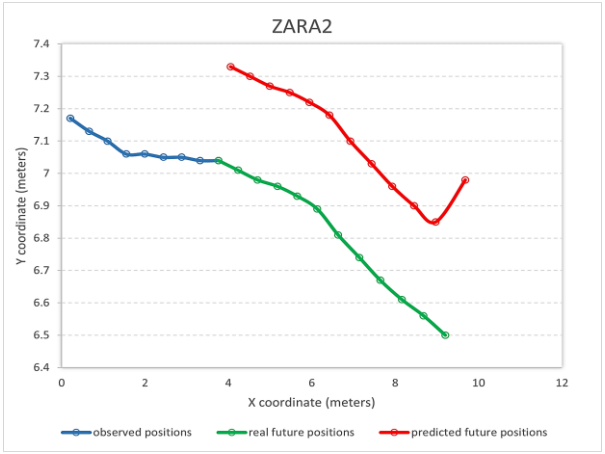
(**b**)



(**c**)



(**d**)



(**e**)

Figure 4. Examples of trajectory prediction results. (**a**) Pedestrian crossing the road in the ETH scene. (**b**) Interaction of pedestrians and objects in the HOTEL scene. (**c**) No social interaction in the UNIV scene. (**d**) Sudden change in direction in the ZARA1 scene. (**e**) Pedestrians walking together throughout the ZARA2 scene.

Figure 4c corresponds to the UNIV scene, where the subject walks on campus alone without interference. The proposed model collects information on the subject–environment interaction using the spatial-temporal graph and provides various impact weights based on that information. As a result, the predicted trajectory is pretty close to the real future positions.

Figure 4d represents the ZARA1 scene in which the trajectory prediction fails. The analyzed pedestrian normally walks in the straight direction but is temporally diverted in the direction by other pedestrians passing nearby. The expected outcome of the model for this situation is not satisfactory.

Finally, in Figure 4e, one can see the best result obtained with our model in the ZARA2 scene. Here is a crossing scenario where the pedestrians are considered to be walking together with two others throughout the entire scene. Figure 4e shows that, taking into account the spatial interaction, more socially appropriate trajectories are obtained.

6.3. Comparison with the State-of-the-Art Methods

Table 2 presents the comparison between the state-of-the-art methods and the one proposed in the ETH, UCY, and SDD datasets in terms of the ADE/FDE evaluation metrics. In most situations, the presented method outperforms the other current state-of-the-art approaches.

Table 2. Quantitative results of state-of-the-art methods for ETH, UCY, and SDD datasets in terms of ADE/FDE metrics. All models use eight frames as input and predict the next twelve frames. The column AWG represents the average results between the ETH-UCY datasets' scenes. n/a means that the corresponding articles have not provided detailed results in the SDD dataset.

Methods	SDD	ETH	HOTEL	UNIV	ZARA1	ZARA2	AWG
S-LSTM [8]	31.19/56.97	1.09/2.35	0.79/1.76	0.67/1.40	0.47/1.00	0.56/1.17	0.71/1.53
Social-STGCNN [9]	n/a	0.64/1.11	0.49/0.85	**0.44**/0.79	0.34/**0.53**	0.30/0.48	0.44/0.75
SR-LSTM [15]	n/a	0.63/1.25	0.37/0.74	0.51/1.10	0.41/0.90	0.32/0.70	0.44/0.93
SR-LSTM-2 [16]	n/a	0.58/1.13	0.31/0.62	0.50/1.10	0.41/0.90	0.33/0.73	0.43/0.89
S-GAN-P [18]	27.23/41.44	0.87/1.62	0.67/1.37	0.76/1.52	0.35/0.68	0.42/0.84	0.61/1.20
S-Ways [19]	n/a	**0.39**/**0.64**	0.39/0.66	0.55/1.31	0.44/0.64	0.51/0.92	0.45/0.83
SoPhie [20]	16.27/29.38	0.70/1.43	0.76/1.67	0.54/1.24	0.30/0.63	0.38/0.78	0.53/1.15
SSALVM (20) [21]	n/a	0.61/1.09	0.28/0.51	0.59/1.24	**0.30**/0.64	0.37/0.78	0.43/0.85
MATF-GAN [23]	27.82/59.31	1.33/2.49	0.51/0.95	0.56/1.19	0.44/0.93	0.34/0.73	0.64/1.26
PIF [25]	n/a	0.73/1.65	**0.30**/0.59	0.60/1.27	0.38/0.81	0.31/0.68	0.46/1.00
Social BiGAT [36]	n/a	0.69/1.29	0.49/1.01	0.55/1.32	**0.30**/0.62	0.36/0.75	0.47/0.99
STGAT [37]	n/a	0.65/1.12	0.35/0.66	0.52/1.10	0.34/0.69	0.29/0.60	0.43/0.83
CGNS [39]	15.84/**25.17**	0.62/1.40	0.70/0.93	0.48/1.22	0.32/0.59	0.35/0.71	0.49/0.97
Our method	**15.18**/25.50	0.63/1.03	0.37/**0.58**	0.46/**0.78**	0.35/0.56	**0.29**/**0.48**	**0.42**/**0.68**

Regarding the ADE metric, the presented approach decreased the error by 3% for the ETH-UCY datasets compared to the state-of-the-art SR-LSTM-2 solution [16] and decreased the error by 5% for the SDD dataset compared to the state-of-the-art method CGNS [39]. Regarding the FDE metric, the presented approach reduced the error by 10% compared to the baseline state-of-the-art method social-STGCN [9]. Surprisingly, D-STGCN, which does not utilize scene image data, outperforms the methods that do, such as SoPhie [20], PIF [25], or STGAT [37].

7. Discussion and Conclusions

In this paper, a suitable graph-based spatio-temporal approach for the prediction of pedestrians' trajectories is presented. The experimental results improve the previous findings of [9,47] by showing the efficiency of attention-based spatial and temporal graph neural networks along with the importance of an optimization procedure performed with respect to the number of layers for both modules of the neural network. Comparative experiments on ETH, UCY, and SDD datasets indicate that the proposed method outperforms the baseline approach and other state-of-the-art solutions in terms of the accuracy

and performance on the ADE/FDE metrics. The visualization results demonstrate that the presented approach can provide reasonable interaction weights and generate coherent future unimodal trajectories.

In future work, previous [48] and current results will be considered for future developments aimed at addressing the more complicated real-world situations commonly encountered in autonomous driving. Furthermore, it is interesting to determine the performance obtained using a larger dataset such as the nuScenes dataset [49]. A richer dataset would offer images taken in more traffic scenarios and eventually surprise human social interaction. Some rich information can be extracted from interaction, such as grouping (for example, at road-crossing points) and ungrouping (once the road has been crossed).

Author Contributions: Conceptualization, B.I.S. and C.D.C.; methodology, I.R.S.; software, B.I.S. and I.R.S.; validation, I.R.S. and C.D.C.; formal analysis, C.D.C.; investigation, B.I.S. and I.R.S.; resources, B.I.S.; writing—original draft preparation, B.I.S., I.R.S. and C.D.C.; writing—review and editing, I.R.S. and C.D.C.; visualization, B.I.S.; supervision, C.D.C. All authors have read and agreed to the published version of the manuscript.

Funding: This paper was financially supported by the Project "Network of excellence in applied research and innovation for doctoral and postdoctoral programs/InoHubDoc", project co-funded by the European Social Fund financing agreement no. POCU/993/6/13/153437.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lefèvre, S.; Vasquez, D.; Laugier, C. A survey on motion prediction and risk assessment for intelligent vehicles. *Robomech J.* **2014**, *1*, 1. [CrossRef]
2. WHO. *Global Status Report on Road Safety 2018*; WHO: Geneva, Switzerland, 2018; p. 11.
3. ITF. *Pedestrian Safety, Urban Space and Health*; OECD Publishing: Paris, France, 2012.
4. Gálvez-Pérez, D.; Guirao, B.; Ortuño, A.; Picado-Santos, L. The Influence of Built Environment Factors on Elderly Pedestrian Road Safety in Cities: The Experience of Madrid. *Int. J. Environ. Res. Public Health* **2022**, *19*, 2280. [CrossRef] [PubMed]
5. Winkle, T. Safety benefits of automated vehicles: Extended findings from accident research for development, validation, and testing. In *Autonomous Driving*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 335–364.
6. Moussaïd, M.; Perozo, N.; Garnier, S.; Helbing, D.; Theraulaz, G. The walking behaviour of pedestrian social groups and its impact on crowd dynamics. *PloS ONE* **2010**, *5*, e10047. [CrossRef] [PubMed]
7. Sharma, N.; Dhiman, C.; Indu, S. Pedestrian Intention Prediction for Autonomous Vehicles: A Comprehensive Survey. *Neurocomputing* **2022**, *508*, 120–152. [CrossRef]
8. 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 (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 961–971.
9. 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 Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 14412–14420.
10. Pellegrini, S.; Ess, A.; Schindler, K.; Van Gool, L. You'll never walk alone: Modeling social behavior for multi-target tracking. In Proceedings of the IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 261–268.
11. Lerner, A.; Chrysanthou, Y.; Lischinski, D. Crowds by example. *Comput. Graph. Forum* **2007**, *26*, 655–664. [CrossRef]
12. Robicquet, A.; Sadeghian, A.; Alahi, A.; Savarese, S. Learning Social Etiquette: Human Trajectory Understanding in Crowded Scenes. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Volume 9912.
13. Fernando, T.; Denman, S.; Sridharan, S.; Fookes, C. Soft+ hardwired attention: An lstm framework for human trajectory prediction and abnormal event detection. *Neural Netw.* **2018**, *108*, 466–478. [CrossRef] [PubMed]
14. Xue, X.; Huynh, D.; Reynolds, M. SS-LSTM: A Hierarchical LSTM Model for Pedestrian Trajectory Prediction. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 1186–1194.

15. 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 and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 12077–12086.
16. Zhang, P.; Xue, J.; Zhang, P.; Zheng, N.; Ouyang, W. Social-Aware Pedestrian Trajectory Prediction via States Refinement LSTM. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 2742–2759. [CrossRef] [PubMed]
17. Jain, A.; Zamir, A.R.; Savarese, S.; Saxena, A. Structural-RNN: Deep Learning on Spatio-Temporal Graphs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 5308–5317.
18. Gupta, A.; Johnson, J.; Fei-Fei, L.; Savarese, S.; Alahi, A. Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 2255–2264.
19. Amirian, J.; Hayet, J.; Pettré, J. Social Ways: Learning Multi-Modal Distributions of Pedestrian Trajectories with GANs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 16–17 June 2019; pp. 2964–2972.
20. Sadeghian, A.; Kosaraju, V.; Hirose, N.; Rezatofighi, H.; Savarese, S. 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 (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 1349–1358.
21. Díaz Berenguer, A.; Alioscha-Perez, M.; Oveneke, M.C.; Sahli, H. Context-Aware Human Trajectories Prediction via Latent Variational Model. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 1876–1889. [CrossRef]
22. Nikhil, N.; Tran Morris, B. Convolutional neural network for trajectory prediction. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018.
23. Zhao, T.; Xu, Y.; Monfort, M.; Choi, W.; Baker, C.; Zhao, Y.; Wang, Y.; Wu, Y.N. Multi-Agent Tensor Fusion for Contextual Trajectory Prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 12118–12126.
24. Chandra, R.; Bhattacharya, U.; Bera, A.; Manocha, D. TraPHic: Trajectory Prediction in Dense and Heterogeneous Traffic Using Weighted Interactions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 8475–8484.
25. Liang, J.; Jiang, L.; Niebles, J.; Hauptmann, A.; 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 Workshops (CVPRW), Long Beach, CA, USA, 16–17 June 2019; pp. 2960–2963.
26. Salzmann, T.; Ivanovic, B.; Chakravarty, P.; Pavone, M. Trajectron++: Multi-agent generative trajectory forecasting with heterogeneous data for control. In Proceedings of the 16th European Conference on Computer Vision (ECCV), Seattle, WA, USA, 16–18 June 2020; pp. 683–700.
27. Zamboni, S.; Kefato, Z.; Girdzijauskas, S.; Noren, C.; Dal Col, L. Pedestrian trajectory prediction with convolutional neural networks. *Pattern Recognit.* **2022**, *121*, 108252. [CrossRef]
28. Battaglia, P.W.; Hamrick, J.B.; Bapst, V.; Sanchez-Gonzalez, A.; Zambaldi, V.; Malinowski, M.; Tacchetti, A.; Raposo, D.; Santoro, A.; Faulkner, R.; et al. Relational inductive biases, deep learning, and graph networks. *arXiv* **2018**, arXiv:1806.01261.
29. Bruna, J.; Zaremba, W.; Szlam, A.; LeCun, Y. Spectral networks and locally connected networks on graphs. *arXiv* **2013**, arXiv:1312.6203.
30. Defferrard, M.; Bresson, X.; Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 3844–3852.
31. Kipf, T.; Welling, M. Semi-supervised classification with graph convolutional networks. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
32. Hamilton, W.; Ying, R.; Leskovec, J. Inductive representation learning on large graphs. In Proceedings of the Annual Conference on Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 1024–1034.
33. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; pp. 7444–7452.
34. Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lió, P.; Bengio, Y. Graph attention networks. In Proceedings of the 6th International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
35. Isola, P.; Zhu, J.; Zhou, T.; Efros, A. Image-to-Image Translation with Conditional Adversarial Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5967–5976.
36. Kosaraju, V.; Sadeghian, A.; Martin-Martin, R.; Reid, I.; Rezatofighi, S.; Savarese, S. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 137–146.
37. Huang, Y.; Bi, H.; Li, Z.; Mao, T.; Wang, Z. STGAT: Modeling Spatial-Temporal Interactions for Human Trajectory Prediction. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6271–6280.
38. Vemula, A.; Muelling, K.; Oh, J. Social attention: Modeling attention in human crowds. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–26 May 2018; pp. 4601–4607.

39. Li, J.; Ma, H.; Tomizuka, M. Conditional Generative Neural System for Probabilistic Trajectory Prediction. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 4–8 November 2019; pp. 6150–6156.
40. Sighencea, B.I.; Stanciu, R.I.; Căleanu, C.D. A Review of Deep Learning-Based Methods for Pedestrian Trajectory Prediction. *Sensors* **2021**, *21*, 7543. [CrossRef] [PubMed]
41. Bai, S.; Kolter, J.Z.; Koltun, V. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv* **2018**, arXiv:1803.01271.
42. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1026–1034.
43. Zou, X.; Sun, B.; Zhao, D.; Zhu, Z.; Zhao, J.; He, Y. Multi-Modal Pedestrian Trajectory Prediction for Edge Agents Based on Spatial-Temporal Graph. *IEEE Access* **2020**, *8*, 83321–83332. [CrossRef]
44. Huang, L.; Zhuang, J.; Cheng, X.; Xu, R.; Ma, H. STI-GAN: Multimodal Pedestrian Trajectory Prediction Using Spatiotemporal Interactions and a Generative Adversarial Network. *IEEE Access* **2021**, *9*, 50846–50856. [CrossRef]
45. Dendorfer, P.; Ošep, A.; Leal-Taixé, L. Goal-GAN: Multimodal Trajectory Prediction Based on Goal Position Estimation. In Proceedings of the Asian Conference on Computer Vision, Kyoto, Japan, 30 November–4 January 2020; Volume 12623, pp. 405–420.
46. Chai, R.; Tsourdos, A.; Savvaris, A.; Chai, S.; Xia, Y.; Chen, C.L.P. Multiobjective Overtaking Maneuver Planning for Autonomous Ground Vehicles. *IEEE Trans. Cybern.* **2021**, *51*, 4035–4049. [CrossRef] [PubMed]
47. Sighencea, B.I.; Stanciu, R.I.; Căleanu, C.D. Pedestrian Trajectory Prediction in Graph Representation Using Convolutional Neural Networks. In Proceedings of the IEEE 16th International Symposium on Applied Computational Intelligence and Informatics (SACI), Timisoara, Romania, 25–28 May 2022; pp. 000243–000248.
48. Sighencea, B.I.; Stanciu, R.I.; Sorândaru, C.; Căleanu, C.D. The Alpha-Beta Family of Filters to Solve the Threshold Problem: A Comparison. *Mathematics* **2022**, *10*, 880. [CrossRef]
49. Caesar, H.; Bankiti, V.; Lang, A.; Vora, S.; Liong, V.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. Nuscenes: A multimodal dataset for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11618–11628.