



**HAL**  
open science

# Cross-Modal Attention for Accurate Pedestrian Trajectory Prediction

Mayssa Zaier, Hazem Wannous, Hassen Drira, Jacques Boonaert

► **To cite this version:**

Mayssa Zaier, Hazem Wannous, Hassen Drira, Jacques Boonaert. Cross-Modal Attention for Accurate Pedestrian Trajectory Prediction. 34th British Machine Vision Conference 2023, Nov 2023, Aberdeen, United Kingdom. hal-04448852

**HAL Id: hal-04448852**

**<https://hal.science/hal-04448852v1>**

Submitted on 13 Feb 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Cross-Modal Attention for Accurate Pedestrian Trajectory Prediction

BMVC 2023 Submission # 467

## Abstract

Accurately predicting human behavior is essential for a variety of applications, including self-driving cars, surveillance systems, and social robots. However, predicting human movement is challenging due to the complexity of physical environments and social interactions. Most studies focus on static environmental information, while ignoring the dynamic visual information available in the scene. To address this issue, we propose a novel approach called Cross-Modal Attention Trajectory Prediction (CMATP) able to predict human paths based on observed trajectory and dynamic scene context. Our approach uses a bimodal transformer network to capture complex spatio-temporal interactions and incorporates both pedestrian trajectory data and contextual information. Our approach achieves state-of-the-art performance on five real-world pedestrian prediction datasets, making it a promising solution for improving the safety and reliability of pedestrian detection and tracking systems.

## 1 Introduction

Accurately predicting human movement has significant applications in various domains, including autonomous driving, surveillance systems, and wheelchair automation. It helps detect potential threats in security, ensures safe navigation in autonomous driving, and provides valuable insights into human-environment interactions for social and behavioral sciences. However, predicting human movement is a challenging task due to dynamic interactions between agents, complex environments, and long-term dependencies. The multimodality of human motion also presents a significant challenge.

Recent research has focused on leveraging the power of deep learning models to improve the accuracy of predicting human movement. Early models, such as Social Forces, had limitations in complex crowded environments. Researchers have since developed sequence prediction methods based on Recurrent Neural Networks (RNNs) [2], which performed well for modeling nearby trajectories but could not capture the impact of further pedestrian motion. More recent works have combined temporal encoding of kinematics data using LSTM and spatial feature extraction through convolution networks on image inputs [26], improving state-of-the-art results. However, these models have limitations in predicting unexpected scenarios, such as sudden changes in motion direction or avoidance of moving obstacles.

To overcome these limitations, we propose a novel approach that utilizes Transformer Networks, which we believe prioritize attentive focus as a crucial aspect in predicting trajectories. While most current methods treat trajectory prediction as time sequence generation

using LSTMs or Transformers, our approach fully leverages both the set of coordinates and videos through multimodal transformers. However, despite increasing research in this area, most studies still overlook the dynamic visual information available in the scene, instead focusing on static environmental data. To address this gap, we introduce the Cross-Modal Attention Trajectory Prediction (CMATP) framework, which predicts human paths based on both the observed trajectory and dynamic scene context, leveraging a ResNet and attention mechanism on video input. By doing so, CMATP captures both environmental constraints and social interactions in dynamic scenes, without requiring communication with other humans.

Our approach includes a cross-attention module that integrates trajectory data with contextual information, allowing the network to capture the general temporal consistency of pedestrian movement. By using a convolutional model for feature extraction and a bimodal transformer, CMATP captures intricate spatio-temporal interactions, improving accuracy while maintaining the same computational complexity as using a single data type. The main contribution lies in the ability to leverage the benefits of two input modalities while avoiding the computational overhead of incorporating additional data types.

## 2 Related Work

This paper discusses research trends in human trajectory forecasting, a topic that has garnered interest for over two decades. We identify three major research directions: improving sequence modeling, studying the impact of people’s actions on each other, and modeling interactions between people and their environment.

**Sequence modeling using RNNs.** RNNs are often used to generate sequences, including kinematic trajectory information [1, 18, 57]. However, they struggle to capture spatio-temporal interactions among humans in a scene [8, 23]. To address this, researchers have proposed augmenting RNNs with pooling [2, 8] or attention [3, 29] modules. Recent work [26] leverages dynamic scene features via a conditional 3D visual encoder based on attention which captures complex interactions. However, RNNs and CNNs have limitations in modeling long-term dependencies and extracting local sequence patterns [30]. Transformers are argued to be more suitable for sequence modeling and trajectory forecasting, especially with large amounts of data, due to their better capability of learning non-linear patterns.

**Social aware models.** Pedestrian trajectory prediction can be approached either by modeling pedestrians as a crowd or as individuals. Traditional crowd models [11, 9, 22, 53] rely on handcrafted kinetic forces and energy potentials to help pedestrians reach their goals while avoiding collisions. But, these methods cannot capture complex interactions in crowded environments. Recent works focus on RNN-based architectures to encode interactions between humans [2, 13, 14, 57]. However, RNNs struggle to capture spatio-temporal interactions among pedestrians. Graph representations have been used to capture social interactions [10, 17, 19, 54], but some suffer from limited understanding of the environmental context. Other approaches incorporate models of human interaction with the environment [10, 23, 24], such as visual features [6, 32] and dynamic 3D scene information [26]. There is criticism of RNNs’ ability to model human-human interaction [4, 25], with suggestions that it limits the model’s generalization capability [25]. Transformer-based methods [27, 52, 54] are seen as potentially more suitable for trajectory forecasting, but these methods make predictions based solely on past trajectories, which may fail to detect unpredictable sharp turns, suggesting that additional information, such as environmental configuration, should be incor-

porated. Our work focuses on predicting individual pedestrian motion, sidestepping social and environmental interactions. Fascinatingly, our approach achieves the best performance on the toughest benchmark.

**Context aware models.** Context-aware trajectory prediction models aim to incorporate physical scene information, such as crosswalks and roads. Previous methods have been proposed to extract and integrate static scene information [13, 23, 24]. Recent models used dynamic spatial and temporal context [6, 26]. However, all these models suffer from limitations related to memory and computational complexity. For example, [26] employs 3D-CNNs, but these models can be computationally expensive and require large amounts of memory because they process volumetric data, which can be larger than the 2D images used in traditional CNNs. Incorporating additional visual modalities can significantly improve performance compared to those only trajectory-based methods [9]. However, existing networks often merge features from different modalities through a simple concatenation in the fusion mechanism. However, this approach lacks the ability to capture the interaction between various granular motion features and does not effectively mine the characteristics and relations of distinct modalities.

After reviewing existing research, we found that pedestrian behavior prediction can greatly benefit from the use of Transformer models and attention mechanisms, as well as the inclusion of contextual information and observed trajectory. To address these challenges, we propose a novel model that incorporates all of these features and utilizes a co-attentional mechanism for capturing dynamic motion information. Our model provides a solution to the limitations of existing methods and has the potential to significantly improve pedestrian behavior prediction.

## 3 Approach

### 3.1 Problem Formulation

The aim of this work is to predict the future positions of individuals in a scene using a transformer-based framework. During training, the method requires preprocessed videos with human detection and tracking algorithms, but during testing, it relies on the target human’s trajectory.

At any time-instant  $t$ , the  $i$ th person in the scene is represented by his/her xy-coordinates  $(x_t^{(i)}, y_t^{(i)})$ . We observe the positions of all individuals from time 1 to  $T_{obs}$ , and predict their positions for time instants  $t_{obs} + 1$  to  $t_{pred}$ . In formal terms, we denote the 2D position of human  $i$  at frame  $t$  by:  $u_{obs}^{(i)} = (x_t^{(i)}, y_t^{(i)}) \in R^2$ . Assume we observe trajectories and the scene from frame 1 to  $t_{obs}$ . We represent the observed sequence for a person, denoted as  $i$ , using  $T_{obs}^{(i)} = (u_1^{(i)}, \dots, u_{t_{obs}}^{(i)})$ , and future positions by  $T_{pred}^{(i)} = (u_{t_{obs}+1}^{(i)}, \dots, u_{t_{pred}}^{(i)})$ .

### 3.2 Overview

In order to enhance the precision of pedestrian trajectory forecasting, the proposed model (*CMATP*) employs a bimodal encoder-decoder architecture with a cross-modal attention mechanism, which handles two modalities: kinematic and visual information. The (*CMATP*) model has two parallel encoder branches (Figure 1). The first branch utilizes self-encoding to transform the pedestrian trajectory  $\tau$  into a latent vector  $X_{kin}$ , while the second branch

extracts visual information through a feature extraction process using a pre-trained convolutional neural network, specifically a ResNet50. The resulting feature vector  $v$  is then passed through a fully connected layer and self-attention block to generate a latent vector  $X_{vis}$  that encodes both visual and temporal information. A cross-attention block is introduced to capture the relationship between the kinematic  $X_{kin}$  and visual latent  $X_{vis}$  vectors outputted by the top and bottom self-attention modules, respectively. This cross-attention mechanism effectively improves the accuracy of future trajectory prediction.

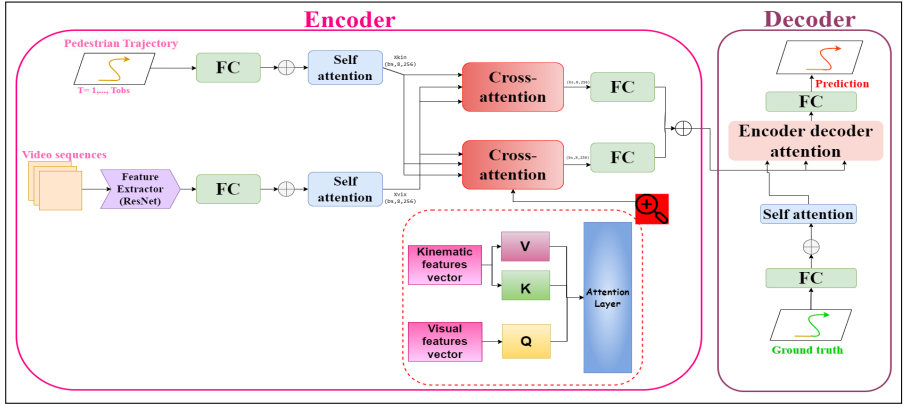


Figure 1: Overview of *CMATP* approach.

The proposed method introduces a novel approach to spatio-temporal attention modeling by decomposing it into two parts: kinematic modeling and contextual modeling. Kinematic modeling employs a temporal Transformer network, which is more effective than RNNs in capturing temporal dependencies from individual trajectory data. Contextual modeling, on the other hand, introduces a Transformer-based encoder module that encodes contextual information from video data to enhance the attention mechanism. To predict human trajectory, the method employs two encoder modules joined by a cross-modal attention mechanism, which is then used with a decoder transformer. The method argues that attention is a crucial component for effective and efficient trajectory prediction.

**Positional encoding.** The transformer architecture uses multi-head self-attention instead of the recurrence mechanism, resulting in faster training and the ability to capture longer dependencies. Positional encoding is a vital component of the transformer, which incorporates sequence order information by adding positional embeddings to input embeddings.

**Attention module.** The use of self-attention mechanisms improves the model's ability to capture long-term dependencies, model complex interactions. In particular, attention modules capture the context and the effect of past data on the current one by dividing the sequence entries into *Query* ( $Q$ ), *Keys* ( $K$ ), and *Values* ( $V$ ). The *Query* is compared to all other *Keys* using a scaled dot product, and the output determines the weights assigned to the *Values*. Attention is therefore given by the equation:  $Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_k}})$ , where  $d_k$  is the embedding dimension which is equal for both the  $Q$  and  $K$ .

**Cross-attention mechanism.** Cross-attention mechanism allows the model to be more contextually aware, as it takes into account the effect of visual context on pedestrian trajec-

184 tories. This is particularly important in crowded scenes where the visual environment plays  
 185 a significant role in predicting pedestrian movements. Figure 1 illustrates the operation of  
 186 cross-attention.

187  
 188 **Training method/ Loss function.** As prior work [26], our loss function consists of two  
 189 components - the mean-squared loss and a regularization term called  $\mathcal{L}_{reg}$ , which regulates  
 190 the smoothness of future trajectories. In training our network, we use the following loss  
 191 function:  $\mathcal{L}_{model} = \mathcal{L}_{mse} + \lambda \mathcal{L}_{reg}$ , where  $\lambda$  is a regularization parameter. We kept the  
 192 value of  $\lambda$  fixed at 0.5 in our experiments to avoid restricting the model’s ability to capture  
 193 sudden changes in the target pedestrians’ trajectory.  $\mathcal{L}_{mse}$  is calculated as the average of the  
 194 squared differences between predicted and observed values, while  $\mathcal{L}_{reg}$  is calculated as the  
 195 sum of Euclidean distances between each step of the predicted trajectory and a line fitted to  
 196 the observed trajectory. In our experiments, we sample 20 future trajectories and select the  
 197 top 5 trajectories that are closest to the ground-truth to calculate  $\mathcal{L}_{mse}$ . Through empirical  
 198 observation, we have found that this approach enables our network to converge more quickly  
 199 while producing more accurate predictions.

## 201 4 Experiments

### 202 4.1 Experimental Setup

203  
 204 **Datasets.** Our approach was evaluated on well-established public human-trajectory datasets,  
 205 namely ETH [20] and UCY [14] datasets, which are widely-used benchmarks for pedestrian  
 206 motion prediction. These datasets were acquired from surveillance videos of pedestrians  
 207 walking on sidewalks and annotated with location coordinates. The datasets contain real-  
 208 world pedestrian trajectories in top-view coordinates expressed in meters, with rich human-  
 209 human and human-object interaction scenarios. The acquisition was done using a fixed cam-  
 210 era on 5 different scenarios captured at 2.5 Hz, and an image is annotated with pedestrian  
 211 positions every 0.4 seconds. ETH/UCY consists of five different scenes. The ETH dataset  
 212 consists of two scenes (*ETH* and *Hotel*) taken from a bird’s eye view, with hundreds of  
 213 pedestrian trajectories engaged in walking activities. The UCY dataset provides three scenes  
 214 (*Zara1*, *Zara2*, and *Univ*) taken from a bird’s eye view with standing/walking activities.

215  
 216 **Evaluation Metrics.** Similar to existing works [2, 3, 8, 10, 11, 13, 15, 23, 24, 32, 33],  
 217 our method is evaluated using two widely used metrics in the field, namely the *Average Dis-*  
 218 *placement Error (ADE)* and the *Final Displacement Error (FDE)*. ADE is defined as the  
 219 average L2 distance (in meters) between the actual trajectory and the predicted trajectory at  
 220 each time step of the trajectory from  $T_{obs+1}$  to  $T_{pred}$  on average over all pedestrians. FDE is  
 221 defined as the Euclidean distance (in meters) between the ground truth (actual position) and  
 222 the prediction (predicted position) at the last time step of the prediction  $T_{pred}$ , averaged over  
 223

224 all pedestrians. Formally:  $ADE = \frac{\sum_{i=1}^n \sum_{t=T_{obs+1}}^{T_{pred}} \|\hat{Y}_t^i - Y_t^i\|}{n * T}$ ;  $FDE = \frac{\sum_{i=1}^n \|\hat{Y}_{pred}^i - Y_{pred}^i\|}{n}$

225 Where  $n$  represents the number of pedestrians,  $\hat{Y}_t^i$  are the predicted coordinates for pedestrian  
 226  $i$  at time  $t$ ,  $Y_t^i$  are the real future positions, and  $\|\cdot\|$  is the Euclidean distance.  $T_{pred}$  is the final  
 227 predicted timestep.  $T$  is the prediction horizon.

228  
 229 **Evaluation method.** For benchmarking purposes, we follow a similar evaluation method

as in prior works (See Table 1) When evaluating trajectory forecasting models, the *time horizon* of the forecast is crucial, as different objects move at different speeds. The choice of time horizon should depend on the class of objects being considered. To have a fair comparison with all the existing works, we observe each training trajectory for 8 times-steps (3.2 seconds) and evaluate the model’s performance by measuring prediction errors for the next 12 time-steps (4.8 seconds). To fully utilize the datasets during model training, we adopt a *leave-one-out* approach for evaluation that has been commonly used in previous studies. We train our model on four sets of data and evaluate it on the remaining set. We repeat this process for all the 5 sets.

**Implementation details.** Our model is based on the original Transformer Networks architecture [28] with a dimension of model of 512 and 6 layers with 8 heads. We trained the entire network end-to-end with a batch size of 40 for 400 epochs, using stochastic gradient descent (SGD) optimizer with a learning rate scheduler and two mean squared error (MSE) loss functions. The learning rate is adjusted every 40 steps with an initial learning rate of 0.01 and the maximum gradient value is clipped to 1 to prevent gradient explosion. We adopted the teacher force strategy and used our proposed loss function with a value of  $\lambda = 0.5$  to achieve faster convergence as prior work [26]. The model was implemented using PyTorch on an Ubuntu server equipped with an NVIDIA TITAN RTX GPU and 24 GB RAM.

## 4.2 Results

**Baseline.** For evaluation purposes, we generate 20 predictions for each observed trajectory and select the prediction closest to the ground truth. This evaluation technique enables us to examine the multi-modality and diversity of the predictions. We evaluate our approach against six *deterministic baselines*, which are linear regression, LSTM, Social-LSTM [2], Social ATTN [29], TrafficPredict [17], and SR-LSTM [37]. We also compare our approach against various *generative baselines* [3, 7, 8, 10, 11, 12, 13, 16, 18, 19, 23, 24, 26, 31, 34, 35, 38] using various approaches such as LSTM, GAN, spatio-temporal GCNs, and transformers to predict human trajectories.

**Quantitative Analysis.** In Table 1, we report obtained results against state-of-the-art approaches as mentioned above, using the *best-of-20 protocol*, which involves sampling 20 possible future trajectories and selecting the one with the best test performance. Our proposed method achieves outstanding performance, ranking either first or second among state-of-the-art methods. In particular, on the FDE metric, our method significantly outperforms existing algorithms on 4 out of 5 datasets, achieving the best average error of 0.33. On the ADE metric, the proposed method outperforms existing algorithms on 3 out of 5 datasets and achieves an average ADE error of 0.22 across all 5 datasets. The University dataset has higher displacement errors compared to other datasets, making it challenging to predict future trajectories accurately. Our method remains comparable to other existing approaches but outperforms all the dense interaction-based methods like *S-GAN*, *Sophie*, *S-BiGAT*, *S-STGCNN*, and *Social Ways*. The *Hotel* dataset has many pedestrians waiting for trains, resulting in limited motion. Therefore, most methods, including ours, achieve relatively small displacement errors by predicting small motions accurately. Our proposed method achieves the lowest FDE (0.16) and ADE (0.11) errors on this dataset. The *ETH* dataset often produces larger displacement errors, which is a common occurrence among many models, due to lower frequency of video frames and kinematic data. However, our



Method	Performance ADE/FDE ↓ (m)					
	Univ	Zara1	Zara2	Hotel	ETH	Avg
Linear*	0.82/1.59	0.62/1.21	0.77/1.48	0.39/0.72	1.33/2.94	0.79/1.59
LSTM*	0.61/1.31	0.41/0.88	0.52/1.11	0.86/1.91	1.09/2.41	0.70/1.52
Social-LSTM* [10]	0.67/1.40	0.47/1.00	0.56/1.17	0.79/1.76	1.09/2.35	0.72/1.54
Social-ATTN* [29]	0.33/3.92	0.20/0.52	0.30/2.13	0.29/2.64	0.39/3.74	0.30/2.59
TrafficPredict* [14]	3.31/6.37	4.32/8.00	3.76/7.20	2.55/3.57	5.46/9.73	3.88/6.97
SR-LSTM* [52]	0.51/1.10	0.41/0.90	0.32/0.70	0.37/0.74	0.63/1.25	0.45/0.94
DESIRE [13]	0.59/1.27	0.41/0.86	0.33/0.72	0.52/1.03	0.93/1.94	0.53/1.11
Social-GAN [8]	0.60/1.26	0.34/0.69	0.42/0.84	0.72/1.61	0.81/1.52	0.58/1.18
FSGAN [12]	0.54/1.14	0.35/0.71	0.32/0.67	0.43/0.89	0.68/1.16	0.46/0.91
SoPhie [14]	0.54/1.24	0.30/0.63	0.38/0.78	0.76/1.67	0.70/1.43	0.54/1.15
Trajectron [10]	0.54/1.13	0.43/0.83	0.43/0.85	0.35/0.66	0.59/1.14	0.47/0.92
MATF [53]	0.44/0.91	0.26/0.45	0.26/0.57	0.43/0.80	1.01/1.75	0.48/0.90
Next [14]	0.60/1.27	0.38/0.81	0.31/0.60	0.30/0.59	0.73/1.65	0.46/1.00
Social-BiGAT [10]	0.55/1.32	0.30/0.62	0.36/0.75	0.49/1.01	0.69/1.29	0.48/1.00
Social-STGCNN [14]	0.44/0.79	0.34/0.53	0.30/0.48	0.49/0.85	0.64/1.11	0.44 / 0.75
Social Ways [8]	0.55/1.31	0.44/0.64	0.51/0.92	0.39/0.66	0.39/0.64	0.46/0.83
PECNet [18]	0.35/0.60	0.22/0.39	0.17/0.30	0.18/0.24	0.54/0.87	0.29/0.48
M2P3 [20]	0.64/1.34	0.45/0.95	0.37/0.79	0.54/1.13	1.04/2.16	0.60/1.27
Transformer-TF [10]	0.35/0.65	0.22/0.38	0.17/0.32	0.18/0.30	0.61/1.12	0.31/0.55
STAR [54]	0.31/0.62	0.26/0.55	0.22/0.46	0.17/0.36	0.36/0.65	0.26/0.53
AgentFormer [53]	0.25/0.45	<b>0.18/0.30</b>	<b>0.14/0.24</b>	0.14/0.22	0.45/0.75	0.23/0.39
Trajectron++ [20]	0.30/0.54	0.25/0.41	0.18/0.32	0.18/0.28	0.67/1.18	0.32/0.55
SGN LSTM [16]	0.48/1.08	0.30/0.65	0.26/0.57	0.63/1.01	0.75/1.63	0.48/0.99
Introvert [26]	<b>0.20/0.32</b>	<b>0.16/0.27</b>	0.16/0.25	<b>0.11/0.17</b>	0.42/0.70	<b>0.21/0.34</b>
GroupNet [51]	0.26/0.49	0.21/0.39	0.17/0.33	0.15/0.25	0.46/0.73	0.25/0.44
<b>Our model (CMATP)</b>	0.37/0.52	<b>0.19/0.27</b>	<b>0.14/0.21</b>	<b>0.11/0.16</b>	<b>0.32/0.51</b>	0.22/ <b>0.33</b>

Table 1: The average/final displacement error (ADE/FDE) metrics for several methods compared to our model are shown. Lower is better. The models with \* have deterministic outputs. All the stochastic models sample 20 possible trajectories and report the best result using a *best-of-20 protocol*. All models observe 8 frames and forecast the subsequent 12 frames.

method achieves the lowest ADE/FDE errors on the ETH dataset, showing the effectiveness of our approach, especially the cross-attention module, in capturing and incorporating information about the movements and behaviors of neighboring pedestrians. The inferior performance of our model without cross-attention in table 2 confirms this.

When comparing individual approaches, the transformer predictor outperforms any individual LSTM-based approach. Specifically, *Transformer-TF* performs better than *Social-LSTM* and has a significant advantage over *Social-ATTN* in terms of FDE. However, on the *Zara1* dataset, which is the least structured dataset in the benchmark and mostly consists of straight lines, LSTM-based methods like *Introvert* perform better than transformer based methods, achieving the lowest ADE (0.16) compared to 0.19 achieved by our proposed TF-based method. Our approach shares similarities with *Transformer-TF*, which utilizes an encoder-decoder transformer architecture. However, we have enhanced our model by incorporating contextual information in addition to the pedestrian positions. As seen, our approach outperforms previous Transformer-based methods such as *Transformer-TF*, *STAR*, and *AgentFormer* on the ETH and UCY datasets. Our cross-attention + Transformer encoder/decoder structure explores better dynamic context between agents than Transformer encoder/decoder in terms of trajectory prediction. Overall, our model offers a competitive



alternative to graph-based methods [14, 19] and has the potential to improve trajectory prediction accuracy.

**Qualitative Analysis.** We conducted a qualitative analysis of our approach’s predictions to gain a more comprehensive understanding of its performance. Figure 2 showcases the qualitative outcomes of our trajectory prediction on multiple videos from the ETH and UCY datasets, providing visual evidence of its effectiveness in accurately predicting pedestrian trajectories. Each column contains two plots showcasing two different pedestrians from the same dataset. In most cases, our method is able to accurately predict the future positions of pedestrians in the scene. The examples in Figure 2 show different scenarios, such as *human-human interaction*, *human-space interaction*, and *avoiding obstacles*. For example, the bottom example in *Zara1* demonstrates our model’s success in predicting that the target pedestrian will go through the door of the store on the left side of the scene. In the top example in *Zara2*, our method correctly predicts that the target human entering the scene will avoid a car and turn left. Also, for the bottom example in *Hotel*, our method correctly predicts that the target person entering the scene will avoid a pole and will continue straight towards the train. In the two cases from *ETH*, our method correctly predicts that the target human entering the scene will avoid an obstacle and turn right/left. Finally, in the top example from the *Univ*, we see an instance of human-human interaction, where the target pedestrian slows down before reaching a group of standing people, bypasses them from the left side, and then speeds up. In such crowded scenes, our method is able to capture interactions and predict future positions effectively.

While our model’s predictions closely matched the ground-truth data in most cases, there were scenarios where our predictions were not as precise as we had hoped, such as in the bottom example from the University. However, our approach still captured some of the essential features of the pedestrian’s behavior, demonstrating its effectiveness in capturing the underlying dynamics of the scene.

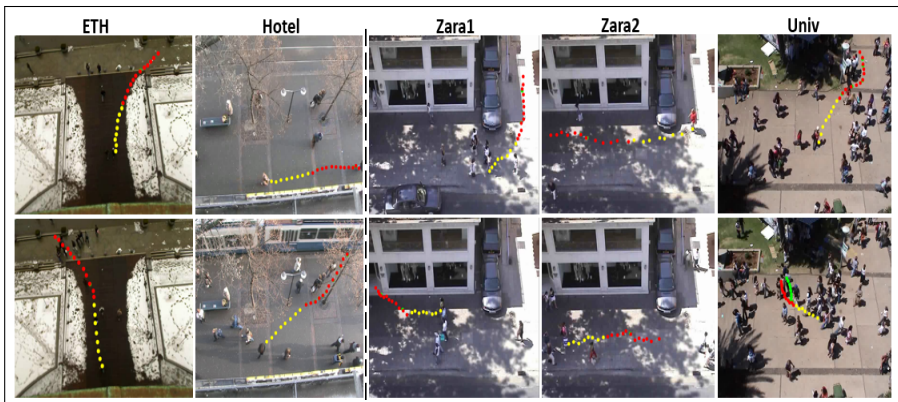


Figure 2: Illustration of the prediction trajectories. yellow dots represents the past observed while red & green dots represent our prediction and the ground truth.

**Ablation Study.** Here, we investigate the effect of the Cross Attention module in the design of trajectory prediction models. We performed *w/o cross attention*, a variant test where we removed the cross attention and concatenated the encoder stream outputs. Results in Table 2 provide insight into the model design for trajectory prediction tasks.

Method	Performance ADE/FDE ↓ (m)					
	Univ	Zara1	Zara2	Hotel	ETH	Avg
<b>Ours w/o CA (BTT)</b>	0.36/0.52	0.19/0.29	0.15/0.23	0.12/0.17	0.48/0.81	0.26/0.40
<b>Ours (CMATP)</b>	0.37/0.52	0.19/0.27	<b>0.14/0.21</b>	<b>0.11/0.16</b>	<b>0.33/0.53</b>	<b>0.22/0.33</b>

Table 2: Ablation study on the ETH/UCY datasets. CA denotes Cross Attention.

Based on analysis of 5 datasets, Cross Attention improved our approach’s performance in predicting accurate trajectories in real-world traffic scenes, outperforming alternatives like concatenation. Results showed our approach with Cross Attention significantly reduced errors to 0.22/0.33 compared to 0.26/0.40 without Cross Attention across 4 out of 5 datasets. However, the *Univ* dataset presented a unique challenge due to higher crowd density and increased uncertainty of future predictions, resulting in comparable error rates between the two models. Further investigation is required to identify reasons behind this discrepancy. Overall, our transformer architecture with Cross Attention enabled smoother temporal predictions and learning of complex sequential patterns, outperforming the baseline model.

**Discussion.** According to our comparison, *CMATP* demonstrates the following key points. First, it predicts accurate trajectories in real-world traffic scenes, surpassing the state-of-the-art methods on 4 out of 5 datasets while achieving comparable performance on the remaining dataset. Second, it incorporates a transformer architecture with *cross attention* to learn interaction, which enables a smoother temporal prediction and outperforms other attention mechanisms, such as additive or multiplicative attention, allowing the model to selectively focus on the most relevant parts of the input sequence. Third, the transformer architecture allows for capturing long-term dependencies and modeling complex interactions between agents in the scene. Fourth, it takes advantage of the transformer’s architecture and considers context, which is crucial for accurate trajectory prediction in real-world traffic scenes. Finally, *CMATP* demonstrates the effectiveness of incorporating a transformer architecture with *cross attention* in learning interaction and improving model performance.

## 5 Conclusion

In this paper, we have proposed a novel approach called *CMATP*, which utilizes an attention-based Transformer Network for pedestrian trajectory prediction. Our framework employs attention mechanisms on dynamic scene context and a cross-attention mechanism to capture complex relationships between the inputs (positions and context), resulting in improved overall performance. The model can produce future-conditional predictions that respect dynamic constraints and full probability distributions, making it suitable for robotic tasks. Our study demonstrates the effectiveness of the Cross Attention mechanism in enhancing the model’s performance. Despite discrepancies between predicted and ground-truth trajectories that may be attributed to the multi-modal nature of pedestrian paths in diverse environments, *CMATP* has significant potential to advance the field of pedestrian trajectory prediction and contribute to the development of safer and more efficient transportation systems. Future work will focus on exploring more sophisticated attention mechanisms, larger training datasets, multi-class settings, and additional contextual information (such as weather and time of day) to enhance our model’s prediction capabilities. We also plan to leverage hierarchical modeling techniques to improve the model’s accuracy further.

## References

- [1] Alexandre Alahi, Vignesh Ramanathan, and Li Fei-Fei. Socially-aware large-scale crowd forecasting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2203–2210, 2014.
- [2] Alahi, Alexandre and Goel, Kratharth and Ramanathan, Vignesh and Robicquet, Alexandre and Fei-Fei, Li and Savarese, Silvio. Social LSTM: Human Trajectory Prediction in Crowded Spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–971, June 2016.
- [3] Javad Amirian, Jean-Bernard Hayet, and Julien Pettré. Social Ways: Learning Multi-Modal Distributions of Pedestrian Trajectories with GANs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.
- [4] Stefan Becker, Ronny Hug, Wolfgang Hubner, and Michael Arens. Red: A simple but effective baseline predictor for the trajnet benchmark. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018.
- [5] Hao Cheng, Wentong Liao, Xuejiao Tang, Michael Ying Yang, Monika Sester, and Bodo Rosenhahn. Exploring Dynamic Context for Multi-path Trajectory Prediction. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12795–12801. IEEE, 2021.
- [6] Chiho Choi and Behzad Dariush. Looking to relations for future trajectory forecast. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 921–930, 2019.
- [7] Francesco Giuliari, Irtiza Hasan, Marco Cristani, and Fabio Galasso. Transformer Networks for Trajectory Forecasting. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 10335–10342. IEEE, 2021.
- [8] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2255–2264, 2018.
- [9] Dirk Helbing and Peter Molnar. Social Force Model for Pedestrian Dynamics. *Physical Review E*, 51(5):4282–4286, May 1995. doi: 10.1103/PhysRevE.51.4282. URL <http://arxiv.org/abs/cond-mat/9805244>.
- [10] B. Ivanovic and Marco Pavone. The Trajectron: Probabilistic Multi-Agent Trajectory Modeling With Dynamic Spatiotemporal Graphs. *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2375–2384, 2019.
- [11] Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian D. Reid, Hamid Rezatofghi, and Silvio Savarese. Social-BiGAT: Multimodal Trajectory Forecasting using Bicycle-GAN and Graph Attention Networks. In *Advances in Neural Information Processing Systems*. Neural Information Processing Systems (NIPS), 2019.
- [12] Parth Kothari and Alexandre Alahi. Human trajectory prediction using adversarial loss. In *Proceedings of the 19th Swiss Transport Research Conference, Ascona, Switzerland*, pages 15–17, 2019.

- [13] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher Bongsoo Choy, Philip H. S. Torr, and Manmohan Chandraker. DESIRE: Distant Future Prediction in Dynamic Scenes with Interacting Agents. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2165–2174, 2017.
- [14] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by Example. *Computer Graphics Forum*, 26(3):655–664, 2007. ISSN 1467-8659. doi: 10.1111/j.1467-8659.2007.01089.x.
- [15] Jiachen Li, Hengbo Ma, and Masayoshi Tomizuka. Conditional Generative Neural System for Probabilistic Trajectory Prediction. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6150–6156. IEEE, 2019.
- [16] Junwei Liang, Lu Jiang, Juan Carlos Nieves, Alexander G Hauptmann, and Li Fei-Fei. Peeking into the future: Predicting future person activities and locations in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5725–5734, 2019.
- [17] Yuexin Ma, Xinge Zhu, Sibozhang, Ruigang Yang, Wenping Wang, and Dinesh Manocha. Trafficpredict: Trajectory prediction for heterogeneous traffic-agents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6120–6127, 2019.
- [18] Karttikeya Mangalam, Harshayu Girase, Shreya Agarwal, Kuan-Hui Lee, Ehsan Adeli, Jitendra Malik, and Adrien Gaidon. It Is Not the Journey but the Destination: Endpoint Conditioned Trajectory Prediction. In *ECCV*, 2020.
- [19] Abdullallah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. 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*, pages 14424–14432, 2020.
- [20] S Pellegrini, A Ess, K Schindler, and L van Gool. You’ll never walk alone: Modeling social behavior for multi-target tracking. In *IEEE 12th International Conference on Computer Vision*, pages 261–268, Kyoto, September 2009. IEEE. ISBN 978-1-4244-4420-5. URL <http://ieeexplore.ieee.org/document/5459260/>.
- [21] Atanas Poibrenski, Matthias Klusch, Igor Vozniak, and Christian Müller. M2P3: multimodal multi-pedestrian path prediction by self-driving cars with egocentric vision. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, pages 190–197, Brno Czech Republic, 2020. ACM. ISBN 978-1-4503-6866-7. URL <https://dl.acm.org/doi/10.1145/3341105.3373877>.
- [22] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning Social Etiquette: Human Trajectory Understanding In Crowded Scenes. In *Computer Vision – ECCV 2016*, volume 9912, pages 549–565. Springer International Publishing, 2016. Series Title: Lecture Notes in Computer Science.
- [23] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, S. Hamid Rezaatofighi, and Silvio Savarese. SoPhie: An Attentive GAN for Predicting Paths Compliant to Social and Physical Constraints. In *Proceedings of the IEEE Computer*

- Society Conference on Computer Vision and Pattern Recognition, CVPR 2019*, pages 1349–1358, United States of America, 2019. IEEE, Institute of Electrical and Electronics Engineers. 506  
507  
508  
509
- [24] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-Feasible Trajectory Forecasting With Heterogeneous Data. *arXiv:2001.03093 [cs]*, January 2021. URL <http://arxiv.org/abs/2001.03093>. 510  
511  
512  
513
- [25] Christoph Schöller, Vincent Aravantinos, Florian Lay, and Alois Knoll. What the constant velocity model can teach us about pedestrian motion prediction. *IEEE Robotics and Automation Letters*, 5(2):1696–1703, 2020. 514  
515  
516  
517
- [26] Nasim Shafiee, Taskin Padir, and Ehsan Elhamifar. Introvert: Human Trajectory Prediction via Conditional 3D Attention. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16810–16820, Nashville, TN, USA, June 2021. IEEE. ISBN 978-1-66544-509-2. URL <https://ieeexplore.ieee.org/document/9577353/>. 518  
519  
520  
521  
522
- [27] Tong Su, Yu Meng, and Yan Xu. Pedestrian Trajectory Prediction via Spatial Interaction Transformer Network. In *2021 IEEE Intelligent Vehicles Symposium Workshops (IV Workshops)*, pages 154–159, 2021. 523  
524  
525  
526
- [28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 527  
528  
529
- [29] Anirudh Vemula, Katharina Muelling, and Jean Oh. Social Attention: Modeling Attention in Human Crowds. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4601–4607, October 2018. doi: 10.1109/ICRA.2018.8460504. 530  
531  
532  
533
- [30] Baoxin Wang. Disconnected recurrent neural networks for text categorization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2311–2320, 2018. 534  
535  
536
- [31] Chenxin Xu, Maosen Li, Zhenyang Ni, Ya Zhang, and Siheng Chen. GroupNet: Multiscale Hypergraph Neural Networks for Trajectory Prediction with Relational Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6498–6507, 2022. 537  
538  
539  
540
- [32] Hao Xue, Du Q. Huynh, and Mark Reynolds. SS-LSTM: A Hierarchical LSTM Model for Pedestrian Trajectory Prediction. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1186–1194, March 2018. 541  
542  
543  
544
- [33] Kota Yamaguchi, Alexander C. Berg, Luis E. Ortiz, and Tamara L. Berg. Who are you with and where are you going? In *CVPR*, pages 1345–1352, Colorado Springs, CO, USA, June 2011. IEEE. ISBN 978-1-4577-0394-2. 545  
546  
547
- [34] Cunjun Yu, Xiao Ma, Jiawei Ren, Haiyu Zhao, and Shuai Yi. Spatio-Temporal Graph Transformer Networks for Pedestrian Trajectory Prediction. In *European Conference on Computer Vision – ECCV 2020*, pages 507–523. Springer International Publishing, 2020. 548  
549  
550  
551

- [35] Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris M Kitani. AgentFormer: Agent-Aware Transformers for Socio-Temporal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9813–9823, 2021.
- [36] Lidan Zhang, Qi She, and Ping Guo. Stochastic trajectory prediction with social graph network. *CoRR*, abs/1907.10233, 2019. URL <http://arxiv.org/abs/1907.10233>.
- [37] Pu Zhang, Wanli Ouyang, Pengfei Zhang, Jianru Xue, and Nanning Zheng. SR-LSTM: State Refinement for LSTM Towards Pedestrian Trajectory Prediction. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12077–12086, 2019.
- [38] Tianyang Zhao, Yifei Xu, Mathew Monfort, Wongun Choi, Chris L. Baker, Yibiao Zhao, Yizhou Wang, and Ying Nian Wu. Multi-Agent Tensor Fusion for Contextual Trajectory Prediction. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 12126–12134, 2019.