TrajViViT: A Trajectory Video Vision Transformer Network for Trajectory Forecasting

Gauthier Rotsart de Hertaing[a], Dani Manjah[b] and Benoit Macq[c]
Institute of Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM), UCLouvain, Place de l'Université 1, 1348 Louvain-la-Neuve, Belgium

Keywords: Vision Transformers, Transformers Networks, Noise Robustness, Trajectory Forecasting, Multi-Modality.

Abstract: Forecasting trajectory is a complex task relying on the accuracy of past positions, a correct model of the agent's motion and an understanding of the social context, which are often challenging to acquire. Deep Neural Networks (DNNs), especially Transformer networks (TFs), have recently evolved as state-of-the-art tools in tackling these challenges. This paper presents TrajViViT (Trajectory Video Vision Transformer), a novel multimodal Transformer Network combining images of the scene and positional information. We show that such approach enhances the accuracy of trajectory forecasting and improves the network's robustness against inconsistencies and noise in positional data. Our contributions are the design and comprehensive implementation of TrajViViT. A public Github repository will be provided.

1 INTRODUCTION

Trajectory forecasting has high-impact applications such as autonomous vehicles for collision avoidance (Liu et al., 2021), tumor motion prediction for efficient proton-therapy (Romaguera et al., 2023; Lombardo et al., 2022) or pedestrian and vehicles motion forecasting for smart-cities (Giuliari et al., 2021; Alahi et al., 2016; Liu et al., 2023). The forecasting of an agent's (i.e., the object being tracked) future motion relies on the agent's past positions (i.e., object's center of mass) and a model of motion, which can be challenging to obtain. More precisely, accurate forecasting remains challenging due to issues like noisy detection of past positions (Zhang et al., 2023; Cheng et al., 2023) and the stochastic nature of agents' movements.

In recent years, Transformer networks (TFs) (Vaswani et al., 2017) have shown promise in trajectory forecasting. Initially, they became state-of-the-art for sequence modelling thanks to their attention mechanism (Giuliari et al., 2021; Franco et al., 2023; Quintanar et al., 2021). The latter leads the network to look at all available observations and to estimate which part of the input trajectory to focus on, in contrast to Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997), which process the past observations sequentially before predicting the future trajectory auto-regressively. Consequently, these networks have a greater ability to deal with missing data, thus capturing the non-linear dynamics of agents more effectively.

However, existing models (Giuliari et al., 2021) operate on the assumption of perfect detection of agents' past positions and primarily focus on encoding past observations, leaving room for improvement in real-world applicability and noise robustness. Our work claims that combining positional and semantic information in the input of the transformer network results in more noise-robust networks, as has been previously demonstrated in LSTM-based models such as Social-BiGAT and Trajectron++ (Kosaraju et al., 2019; Salzmann et al., 2020).

Our work proposes TrajViViT, a **TRAJ**ectory **VI**deo **VI**sion **T**ransformer network, for trajectory forecasting. The main contributions are:

- **TrajViViT.** Implementation of a multimodal Video Vision Transformer network for robust trajectory forecasting.
- **Noise Robustness.** Evaluation of the noise robustness of Transformer networks for trajectory forecasting over various prediction horizon sizes.
- A code to reproduce the experiments and utilize

[a] https://orcid.org/0009-0003-3753-177X
[b] https://orcid.org/0000-0001-9034-0794
[c] https://orcid.org/0000-0002-7243-4778

Rotsart de Hertaing, G., Manjah, D. and Macq, B.
TrajViViT: A Trajectory Video Vision Transformer Network for Trajectory Forecasting.
DOI: 10.5220/0012372000003654

In *Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2024)*, pages 753-760
ISBN: 978-989-758-684-2; ISSN: 2184-4313

the framework will be available at https://github.com/GauthierRotsart/TrajViViT.

2 STATE-OF-THE-ART

Many works have been done on trajectory forecasting during the last decades. Initially, traditional predictors were based on linear, gaussian or Kalman filter (Kalman, 1960) methods. With the growth of deep learning, novel approach were developed based on Artificial Neural Network. Trajectory forecasting has been surveyed by (Huang et al., 2022a) in the context of autonomous vehicles. Therefore, in this paper, we will only review the state-of-the-art with transformer-based techniques.

2.1 Computational Time-Series Trackers

The earliest work on tracking can be traced back to the development of radar technology during World War II where they were used to detect and track the positions of enemy aircraft and ships (Norbert, 1949). In the 1960s, Rudolf Kalman extended (Norbert, 1949) by adding a system state model (e.g., physical law of motion) to sequential state measurements, such as from sensors. The Kalman Filter (KF) (Kalman, 1960) is a recursive algorithm and MSE-optimal estimator for linear system driven by white Gaussian and uncorrelated noise. KF cannot process images as inputs and has to rely on image processing techniques detecting the object and providing the coordinates of the bounding box (Bewley et al., 2016). Moreover, KF is unable to integrate social context and demands a motion model, which can be challenging to procure and may necessitate tuning upon new scenes. Yet, it remains in the state-of-the-art as it provides explainable results and is computationally low-cost algorithm.

2.2 Transformer-Based Tracker

Nowadays, transformer networks (TF) are considered as a state-of-the-art technique in trajectory forecasting. The work of (Giuliari et al., 2021) is the first one using transformers to predict agent's motion and reached the best score on the TrajNet dataset (Sadeghian et al., 2018). Although a lot of works propose to model the social interactions between agents, TF's are simple models because only the positional information is feed into the network. The work of (Franco et al., 2023) extends (Giuliari et al., 2021) by providing a better study of transformers models in trajectory forecasting. They showed the superiority of TF with respect to LSTM (Hochreiter and Schmidhuber, 1997) whatever the forecasting horizon. However, even if the attention mechanism in TF is more effective than the memory mechanism in LSTM, techniques using social context and semantic maps are still challenging transformers (Kosaraju et al., 2019; Salzmann et al., 2020). Finally, (Yao et al., 2022) improves (Giuliari et al., 2021) by adding random deviation in the decoder's input.

Traditionally, trajectory forecasting is formulated as a Deterministic Trajectory Prediction (DTP) task where the predictor is expected to provide only one trajectory, usually the one minimising the L_2 distance with respect to the ground truth (Huang et al., 2023). However, agent's trajectories are generally multi-modal and multiple plausible paths are socially acceptable. For example, there are usually several future possible trajectories for a car at a roundabout and a model cannot be perfectly sure which one is going to be the most plausible one. Hence, (Gupta et al., 2018) formulated the trajectory forecasting problem as a Multi-modal Trajectory Prediction (MTP) task. This has been surveyed by (Huang et al., 2023). Works such as (Liu et al., 2021; Shi et al., 2022; Huang et al., 2022b; Geng et al., 2023) use transformer networks to predict multi-modal predictions. In the MTP formulation, model's performances are usually measured by taking the minimum error between the K plausible predicted trajectories and the ground truth. However, in situations such as video surveillance, there is a need to regress only one future trajectory because the MTP formulation only provides an upper-bound reachable. In that way, we developed TrajViViT, a transformer-based network that uses positional and semantic information to predict trajectories in a deterministic way.

3 PROBLEM FORMULATION

We assume the camera images are preprocessed by a detector such as YOLO (Redmon et al., 2016) or a human, leading to noisy ground truth positions. Then, based on the raw images (i.e., the social and semantic information) as well as the agent's past positions, a transformer-based predictor is trained to predict the future agent's positions. More formally, given an observation window T_{obs} of the agent's positions $X_{obs} = \{u_t \in \mathcal{R}^2 \text{ and } v_t \in \mathcal{R}^{224\times224} | t \in [0, T_{obs}-1]\}$, the model predicts the future agent's positions $X_{pred} = \{x_t \in \mathcal{R}^2 | t \in [T_{obs}, T_{pred}+T_{obs}-1]\}$.

3.1 Transformer Network

Transformer network (Vaswani et al., 2017) is a promising neural network in sequence modelling and its success comes from the use of the attention mechanism which estimates what are the most useful part of the input sequence. The network is divided into two parts: an encoder, to encode the input sequence, and a decoder to process auto-regressively the predicted sequence. The input sequence X_{obs} is first encoded into the embedding space of dimension D through a linear projection layer: $E_{obs} = x_t^T W_x$ where W_x is a matrix of weights. Although the attention module enables to capture the input sequence non-linearities, this mechanism is position invariant. Therefore, time is encoded through a "positional encoding layer" as defined in (Vaswani et al., 2017). This encoding vector p has the same dimension than the latent space. Hence, at time t, the input is encoded as $\varepsilon_{obs}^t = E_{obs}^t + p^t$ and processed by the encoder. Finally, the decoder uses the encoder's output to predict auto-regressively the output sequence.

4 MATERIALS AND METHODS

In this work, images come from the Stanford Drone Dataset and targets are annotated with bounding boxes (Robicquet et al., 2016). Position information corresponds to the center of mass of the bbox. During inference, random noise is added on the positional data. We then evaluate the robustness to noise of the transformer network depending on the input modality. In that way, three transformer-based architectures are developed as depicted on Figure 1.

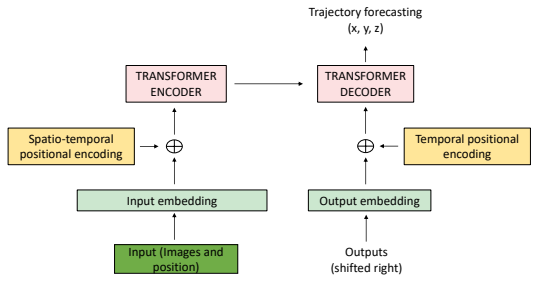

Figure 1: TrajViViT's network is built on a vanilla transformer network (Vaswani et al., 2017). The input is the semantic map and the positional information.

4.1 Implementation Details

The network is trained from scratch and weights are initialised using the Xavier Glorot's initialisation (Glorot and Bengio, 2010). A stochastic gradient descent (SGD) with Adam optimizer (Kingma and Ba, 2014) is used for the training and we arbitrarily chose 100 epochs. The batch size is equal to 4 and a constant learning rate of 5.10^{-5} is used, without scheduling or warm-up. Images coming from SDD contain multiple objects and networks are deterministic. Therefore, we indicate which agent to follow by coloring the bbox in black. The coloured images are then fed to the network which has to predict the future trajectory of the chosen agent. Those contain all the semantic information of the scene and images are reduced to a size of 224×224. While in (Dosovitskiy et al., 2020) the images are first split into patches and then encoded into an embedding space, we use three 3D convolutional layers to encode the spatio-temporal information. The latent space is 16. As in (Giuliari et al., 2021), a linear projection layer is used to embed positional information into a latent space of 16. Then, positional and semantic information are concatenated. A vanilla transformer architecture (Vaswani et al., 2017) is then used with 4 attention heads. The encoder and the decoder have 6 layers. All the process is seeded for reproducibility reasons[1].

4.2 Training Process

The training dataset is composed of N tracks, each containing P_n data points. The method consists of drawing positions randomly among all the tracks in order to create batches with more variability (i.e, pedestrian's track and biker's track inside the batch). Moreover, the network is trained using teacher forcing. Since the transformer is composed of an encoder, it could lead to leakages during inference because the network gets future information when updating weights. However, it saves a lot of time during training.

4.3 Evaluation Process

The detection of an agent in an image is challenging. For instance, an image-based object detector may be subject to domain-shift, domain-drift or occlusion. Alternatively, a radar has a range resolution of several tens of centimetres and is unable to distinguish metal objects that are too close. Consequently, methods cannot assume perfect observations. Therefore, while the network is trained using the ground truth information, TrajViViT is evaluated on different noise levels and different scenes. Those represent localisation errors in coordinates (u, v). In that way, the robustness to noise and the generalisation of the

[1] Supplementary materials for results reproducibility via the Github repository.

network are evaluated. Noise is modelled as Additive White Gaussian Noise (AWGN) distributed with $e_u, e_v \sim \mathcal{N}(\mu = 0, \sigma^2)$ where $\sigma^2 \in [0, 20]$ by step of 5.

4.4 Dataset

The training dataset contains 90 % of all the available tracks of the scene. The validation and test datasets are both equal to 5 % of the remaining tracks. We evaluated the three transformers networks over the Stanford Drone Dataset (SDD) (Robicquet et al., 2016), more specifically on four different scenes (bookstore, coupa, little and nexus). A total of twenty-seven cameras were used for the comparison, each camera constituting a single dataset. Then, TrajViViT is evaluated on three remaining scenes (gates, hyang and deadcircles). Those include around 1,250 (1 million data points) and 2,000 different trajectories (two million data points) for gates and hyang respectively. The deadcircle scene is composed of a bit less than 3,000 trajectories (more than 2 million data points). The quad scene is rejected from the test because it contains not enough trajectories and data points.

4.5 Evaluation Metrics

As in prior works (Ivanovic and Pavone, 2019; Salzmann et al., 2020; Giuliari et al., 2021; Franco et al., 2023), we use the Average Displacement Error (ADE), equivalently Mean Average Displacement (MAD) and the Final Displacement Error (FDE), equivalently Final Average Displacement (FAD), to evaluate the networks performances:

$$ADE = \frac{\sum_{i=t_{obs}}^{T_{pred}+T_{obs}-1} (u_i - \hat{u}_i)^2 + (v_i - \hat{v}_i)^2}{T_{pred}} \quad (1)$$

$$FDE = (u_{t_{end}} - \hat{u}_{t_{end}})^2 + (v_{t_{end}} - \hat{v}_{t_{end}})^2 \quad (2)$$

where (u_i, v_i) are the predicted position in the pixel space at the horizon i and $(\hat{u}_i, \hat{v}_i)$ the respective ground truth position. The timestamp t_{end} corresponds to the last one.

The lower the ADE/FDE, the better the perfomances. We chose to assess networks by using the average of the two metrics.

5 RESULTS

The evaluation protocol of (Alahi et al., 2016) is adopted: each track is sampled at 0.4s (2.5 fps) so that the network observes 8 inputs and predicts 12 outputs. First, networks are trained and tested on the same scene, leading to scene-specific models. Performances are measured with ADE and FDE and the Mean Square Error (MSE) is analysed over the forecasting horizon. We then analyse the robustness to noise of the proposed TrajViViT network. Finally, models are tested on three specific scenes, enabling to assess their generalisation performances on unseen scenes.

5.1 Multimodal TrajViViT

Our first experiment compares the performance of each modality (image or position) and their fusion (image and position). They were evaluated on twenty-five videos coming from 4 domains - coupa, little, bookstore and nexus. Each network is trained on 90% of the available tracks. The 10% remaining tracks are used to validate (5%) and test (5%). We denote by *Img*, *Pos* or *Img+Pos* the input modality when networks are trained and tested on the same scene. Additionally, we evaluated the contribution of each modality when TrajViViT is trained on all the videos of a scene, keeping the same split for the training/validation/test dataset for each video. Those are denoted by *Img multi*, *Pos multi* and *Img+Pos multi*, depending on the input modality. The analysis of Table 1 shows an improvement of the performances when the network uses both modalities, regardless of the number of training data. It results in more accurate trajectories over time and the prediction is better at large horizon.

5.2 Forecasting Horizon and Noisy Measures

In this section, we consider the forecasting performances along the prediction horizon. First, Figure 2 shows the MSE evolution. Since models are tested auto-regressively, the error propagates over time and predictions get worse. We show the use of two modalities reduces MSE. This is especially true when models are trained over all the available videos of the scene.

Although trajectory forecasting using transformer networks was already studied in (Giuliari et al., 2021; Franco et al., 2023; Quintanar et al., 2021; Liu et al., 2021; Liu et al., 2023), these works do not take into account the noise in the agent's past positions, which could lead to a drop of performance during inference. Comparison with respect to the noise's variance is depicted on Figure 3 and notations of Table 1 are used. Performances are represented as the average of ADE and FDE at a noise level σ. Lower is the average, better are the performances. Networks *img* and *img*

Table 1: The contribution of each modality is shown on the Table. Networks are trained and tested on the same scene. In average, results show the combination of the semantic and positional information gives the best performances.

Domain name	Video	Img *ADE/FDE*	Pos *ADE/FDE*	Img + Pos *ADE/FDE*
	0	19,67/22,30	16,51/20,61	17,20/22,20
	1	17,21/22,46	22,32/23,67	12,98/15,06
coupa	2	18,57/28,13	19,14/29,60	15,64/21,26
	3	16,04/18,80	11,82/14,42	9,82/12,68
	ALL	13,92/18,64	7,45/12,16	**5,52/8,86**
AVG coupa		17,87/22,92	17,45/22,08	**13,91/17,80**
	0	18,27/23,10	17,08/23,45	21,94/28,52
	1	17,08/27,10	14,52/27,10	11,50/19,78
little	2	40,61/48,10	17,58/27,02	14,67/22,13
	3	14,99/22,48	13,42/18,55	12,50/16,40
	ALL	21,94/29,43	13,74/18,92	**11,09/16,79**
AVG little		22,74/30,20	15,65/**19,22**	**15,15**/21,71
	0	12,54/17,52	11,45/17,53	10,04/13,54
	1	12,22/16,01	7,15/10,88	10,00/11,19
	2	14,59/21,93	11,83/20,08	10,74/16,61
	3	11,31/16,89	15,71/21,42	12,48/16,02
bookstore	4	14,81/22,13	16,34/21,64	14,92/18,42
	5	9,06/13,01	14,76/18,50	10,25/11,28
	6	33,13/38,70	18,81/26,59	8,84/15,44
	ALL	10,56/14,94	6,34/10,76	**5,42/9,14**
AVG bookstore		15,38/20,88	13,72/19,52	**10,04/14,64**

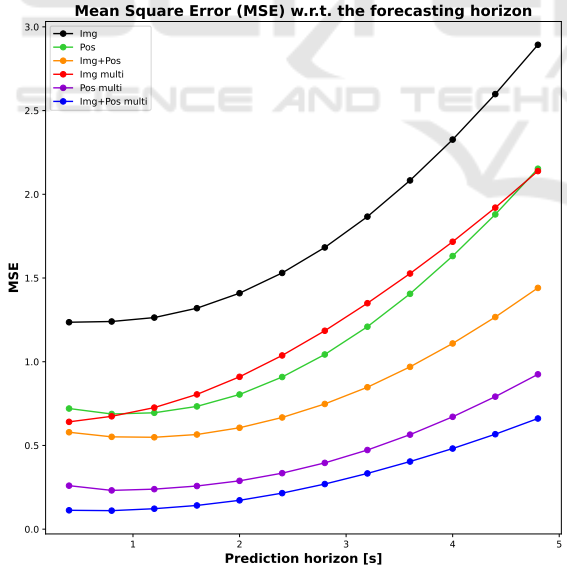

Figure 2: The evolution of the MSE of TrajViViT is depicted with respect to the input modality. The figure shows the combination of the two input modalities improves the performances at larger horizon.

multi are considered as baselines because they are not affected by noise. Figure 3 shows that multimodality significantly improves robustness to noise.

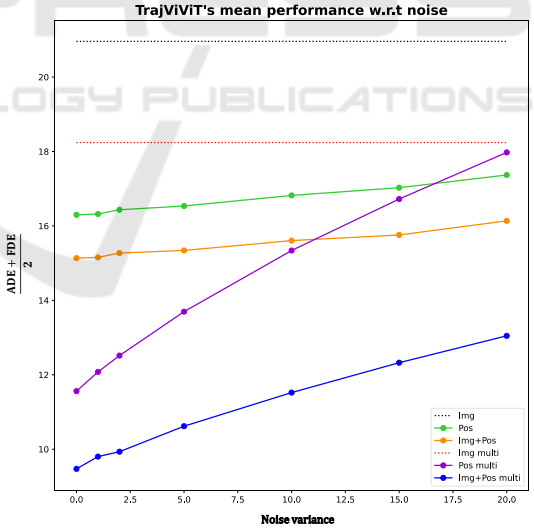

Figure 3: Evaluation of TrajViViT at different noise levels. Networks trained with the image modality are considered as the baseline.

5.3 Generalisation

Finally, we evaluate the generalisation performances of TrajViViT when trained and tested on different domains. Since Figures 2 and 3 show that the more training data there is, the better the performance, we

Table 2: The table shows the generalisation of TrajViViT on three scenes, each containing million of data points. In average, results show the combination of the semantic and positional information gives the best performances.

Domain name	**Modality**	**Gates** *ADE/FDE*	**Hyang** *ADE/FDE*	**Deadcircle** *ADE/FDE*
coupa	img	75,80/78,06	74,48/75,05	76,11/78,10
	pos	10,66/17,45	7,88/11,96	10,03/16,06
	img+pos	8,65/15,38	6,19/9,75	8,20/14,02
little	img	77,50/80,59	69,74/73,02	59,36/63,40
	pos	15,09/20,16	12,05/14,58	13,73/18,35
	img+pos	13,41/20,19	14,66/21,38	11,48/16,56
bookstore	img	87,87/89,05	71,69/71,23	76,37/78,21
	pos	7,75/14,05	5,50/9,21	7,29/13,57
	img+pos	8,27/15,03	5,97/10,74	8,41/14,73
AVG	img	80,39/82,57	71,97/73,10	70,61/73,24
	pos	11,17/17,22	**8,48/11,92**	10,35/15,99
	img+pos	**10,11/16,87**	8,94/13,96	**9,36/15,10**

Table 3: Cross-domain performances of models. The results highlight that combining positional and image inputs increases networks transferability in average.

Domain	**Modality**	**Coupa** *ADE / FDE*	**Little** *ADE / FDE*	**Bookstore** *ADE / FDE*	**Nexus** *ADE / FDE*	**AVG** *ADE / FDE*
coupa	*Img*	13.74 / 18.50	79.33 / 82.58	87.06 / 88.89	58.85 / 61.22	59.74 / 62.80
	Pos	7.33 / 12.09	13.49 / 23.02	8.90 / 14.44	8.44 / 12.80	9.54 / 15.59
	Img+Pos	5.55 / 8.90	11.90 / 20.96	7.45 / 12.58	5.85 / 9.72	**7.68 / 13.04**
little	*Img*	101.60 / 105.10	21.90 / 29.45	85.94 / 86.72	59.20 / 62.83	67.16 / 71.02
	Pos	13.14 / 17.25	13.74 / 18.97	19.73 / 23.41	10.25 / 14.54	14.21 / 18.54
	Img+Pos	15.09 / 18.74	11.05 / 16.82	15.80 / 20.27	9.69 / 14.72	**12.91 / 17.64**
book-store	*Img*	123.51 / 125.63	92.90 / 97.90	11.15 / 15.50	74.70 / 78.02	52.34 / 79.26
	Pos	5.83 / 10.14	9.55 / 17.96	6.42 / 10.79	4.76 / 8.08	**6.64 / 11.74**
	Img+Pos	6.73 / 11.71	10.98 / 20.55	5.41 / 9.12	5.95 / 10.25	7.27 / 12.91
AVG	*Img*	79.62 / 83.08	64.71 / 69.98	61.38 / 63.70	64.25 / 67.36	59.75 / 71.03
	Pos	8.77 / 13.16	12.26 / 19.98	11.68 / 16.21	7.82 / 11.81	10.13 / 15.29
	Img+Pos	9.12 / 13.12	11.31 / 19.44	9.55 / 13.99	7.16 / 11.50	**9.29 / 14.53**

only use networks trained on a whole scene. Table 2 shows the generalisation performances on three scenes (gates, hyang, deadcircle) where models are tested on all the available tracks. The combination of the modalities results in an improvement of the performances.

6 DISCUSSION

First of all, Figure 3 shows that TrajViViT is underfitted when only trained on the image modality. This results in high MSE over the prediction horizon and this is surely due to the fact that input and output data do not belong to the same domain, leading to a too complex task. However, combining the positional and semantic information gives the best performances and shows robustness to noise. We can also notice that training on all the videos of a scene gives a bigger improvement when combining both modalities. Therefore, we think giving more training data enables TrajViViT to use better the semantic information.

6.1 Scene-Specific Models

Table 1 shows the performances of TrajViViT when trained and tested on the same domain. The multi-modal network always gives the best performances, except when training on a single video on the *little* scene. This comes to its poor performances on video 0. Therefore, this is represented by a lower MSE, ADE and FDE over the forecasting horizon. Moreover, as shown on Figure 3, the multi-modality shows robustness to noise for both the ADE and FDE. This results in improvements with respect to models trained only with the positional information.

6.2 Generalisation to Other Domains

In this section, we only consider models trained on a whole scene, due to their superior performances. Table 3 shows the drop of performances when training and testing on the same domain or not. As we can see, the network trained only with the semantic modality has poor generalisation performances. However, on average, the multimodal TrajViViT network gives the best performances both in ADE and FDE. This results in more accurate predicted trajectories at larger horizon.

6.3 Limitations

Firstly, TrajViVit has only been tested on *SDD* data for single-object forecasting in a 2-D context. Consequently, extending it to multi-object tracking and adapting it to depth-presenting images, such as video-surveillance (Naphade et al., 2021) and tumor tracking, is not straightforward and constitutes a future research direction. Secondly, the computational resources required for both training and data can hinder application deployment. To mitigate costs, deployment strategies based on active learning and knowledge distillation, as presented in (Manjah et al., 2023), could reduce the training complexity and data needs of Transformers while ensuring domain adaptation.

7 CONCLUSIONS

In this paper, we proposed TrajViViT, a Trajectory Video Vision Transformer network, for trajectory forecasting. We showed the improvement of performances when combining semantic and positional information, with respect to transformer networks only trained with the agent's past positions. In particular, we have shown the mean square error was the lowest during the forecasting horizon. Moreover, we showed the robustness of the multimodal TrajViViT with respect to noise. Finally, the network was tested on a dataset containing thousands of different trajectories (millions of data points) and the contribution of each modality (semantic and/or positional information) was assessed. The combination of the modalities results in an improvement in the generalisation of the network.

ACKNOWLEDGEMENTS

Gauthier Rotsart de Hertaing and Dani Manjah are respectively supported by the Walloon region under grant n°2010235 – ARIAC by DIGITALWALLONIA4.AI and grant n°2010149 - ARIES. We also wish to give a special thanks to the OpenHub team of UCLouvain for the use of their computational resources that have contributed to the results presented in this paper.

REFERENCES

Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., and Savarese, S. (2016). Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971.

Bewley, A., Ge, Z., Ott, L., Ramos, F., and Upcroft, B. (2016). Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3464–3468.

Cheng, H., Chen, L., et al. (2023). An end-to-end framework of road user detection, tracking, and prediction from monocular images. *arXiv preprint arXiv:2308.05026*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Franco, L., Placidi, L., Giuliari, F., Hasan, I., Cristani, M., and Galasso, F. (2023). Under the hood of transformer networks for trajectory forecasting. *Pattern Recognition*, 138:109372.

Geng, M., Cai, Z., Zhu, Y., Chen, X., and Lee, D.-H. (2023). Multimodal vehicular trajectory prediction with inverse reinforcement learning and risk aversion at urban unsignalized intersections. *IEEE Transactions on Intelligent Transportation Systems*.

Giuliari, F., Hasan, I., Cristani, M., and Galasso, F. (2021). Transformer networks for trajectory forecasting. In *2020 25th international conference on pattern recognition (ICPR)*, pages 10335–10342. IEEE.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings.

Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., and Alahi, A. (2018). Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2255–2264.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Huang, R., Xue, H., Pagnucco, M., Salim, F., and Song, Y. (2023). Multimodal trajectory prediction: A survey. *arXiv preprint arXiv:2302.10463*.

Huang, Y., Du, J., Yang, Z., Zhou, Z., Zhang, L., and Chen, H. (2022a). A survey on trajectory-prediction meth-

ods for autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 7(3):652–674.

Huang, Z., Mo, X., and Lv, C. (2022b). Multi-modal motion prediction with transformer-based neural network for autonomous driving. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2605–2611. IEEE.

Ivanovic, B. and Pavone, M. (2019). The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2375–2384.

Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kosaraju, V., Sadeghian, A., Martín-Martín, R., Reid, I., Rezatofighi, H., and Savarese, S. (2019). Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. *Advances in Neural Information Processing Systems*, 32.

Liu, D., Li, Q., Li, S., Kong, J., and Qi, M. (2023). Non-autoregressive sparse transformer networks for pedestrian trajectory prediction. *Applied Sciences*, 13(5):3296.

Liu, Y., Zhang, J., Fang, L., Jiang, Q., and Zhou, B. (2021). Multimodal motion prediction with stacked transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7577–7586.

Lombardo, E., Rabe, M., Xiong, Y., Nierer, L., Cusumano, D., Placidi, L., Boldrini, L., Corradini, S., Niyazi, M., Belka, C., et al. (2022). Offline and online lstm networks for respiratory motion prediction in mr-guided radiotherapy. *Physics in Medicine & Biology*, 67(9):095006.

Manjah, D., Cacciarelli, D., Standaert, B., Benkedadra, M., de Hertaing, G. R., Macq, B., Galland, S., and De Vleeschouwer, C. (2023). Stream-based active distillation for scalable model deployment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4998–5006.

Naphade, M., Wang, S., Anastasiu, D. C., Tang, Z., Chang, M.-C., Yang, X., Yao, Y., Zheng, L., Chakraborty, P., Lopez, C. E., Sharma, A., Feng, Q., Ablavsky, V., and Sclaroff, S. (2021). The 5th ai city challenge. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Norbert, W. (1949). *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*, volume 113. MIT press Cambridge, MA.

Quintanar, A., Fernández-Llorca, D., Parra, I., Izquierdo, R., and Sotelo, M. (2021). Predicting vehicles trajectories in urban scenarios with transformer networks and augmented information. In *2021 IEEE Intelligent Vehicles Symposium (IV)*, pages 1051–1056. IEEE.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.

Robicquet, A., Sadeghian, A., Alahi, A., and Savarese, S. (2016). Learning social etiquette: Human trajectory understanding in crowded scenes. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 549–565. Springer.

Romaguera, L. V., Alley, S., Carrier, J.-F., and Kadoury, S. (2023). Conditional-based transformer network with learnable queries for 4d deformation forecasting and tracking. *IEEE Transactions on Medical Imaging*.

Sadeghian, A., Kosaraju, V., Gupta, A., Savarese, S., and Alahi, A. (2018). Trajnet: Towards a benchmark for human trajectory prediction. *arXiv preprint*.

Salzmann, T., Ivanovic, B., Chakravarty, P., and Pavone, M. (2020). Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 683–700. Springer.

Shi, S., Jiang, L., Dai, D., and Schiele, B. (2022). Motion transformer with global intention localization and local movement refinement. *Advances in Neural Information Processing Systems*, 35:6531–6543.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Yao, H.-Y., Wan, W.-G., and Li, X. (2022). End-to-end pedestrian trajectory forecasting with transformer network. *ISPRS International Journal of Geo-Information*, 11(1):44.

Zhang, P., Bai, L., Wang, Y., Fang, J., Xue, J., Zheng, N., and Ouyang, W. (2023). Towards trajectory forecasting from detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.