Traffic Accident Risk Forecasting using Contextual Vision Transformers with Static Map Generation and Coarse-Fine-Coarse Transformers

Artur Grigorev[1,*], Khaled Saleh[2] and Adriana-Simona Mihaita[1]
Faculty of Engineering and IT
[1] University of Technology Sydney, Australia
[2] University of Newcastle, Australia
Email: Artur.Grigorev@student.uts.edu.au

Abstract— **We propose an enhancement to our previously proposed novel model called Contextual Vision Transformer (ViT) to address the problem of traffic accident risk forecasting. This framework combines spatial and temporal information using a data-driven approach. By treating the problem as a computer vision task, we can predict traffic accident risk as the next frame in a video sequence. Specificaly, we extend the ViT network with a Static Map generation (named XViT) for even better results on the Chicago dataset. Furthermore, we propose a Coarse-Fine-Coarse transformer architecture as an alternative approach to enhance traffic accident risk prediction.**

Keywords: traffic accident risk; risk prediction; vision transformers; deep learning

I. Introduction

Traffic accidents pose a significant impact on global health and economics, with an upward trend in incidents particularly notable in developing countries 16. The issue persists with over 5 million accidents annually in the United States alone 10, and 1.35 million fatalities worldwide in 2016 17.

Traditionally, traffic accident risk forecasting is viewed as a time-series prediction task, requiring separate models to handle spatial and temporal aspects. Despite initial Deep Learning attempts to predict traffic accident risks, some studies didn't consider traffic flow or time-related factors 2. Subsequent research 11, 21, 22, 20, 15, 14 offered enhancements by incorporating additional contextual data like air quality, weather, and the condition of roads.

This research on traffic accident risk prediction can offer several benefits to different stakeholders involved in urban planning and traffic management:

Traffic Management Authorities: The predictive insights offered by this research can aid traffic management authorities in deploying resources effectively. If certain areas are predicted to have a high risk of accidents at particular times, they can arrange for additional traffic police deployment or emergency medical services in those areas in advance. This can lead to faster response times and potentially save lives in the event of accidents.

Emergency Services: Predicting high-risk scenarios and their potential locations can significantly enhance emergency services' readiness. Knowing when and where accidents are likely to happen means ambulances, fire services, and police can be strategically located to respond rapidly when needed.

This paper introduces a series of enhancements to our previously proposed novel approach that relies on Vision Transformers 4, 18 to forecast traffic accident risk. Our approach leverages the spatio-temporal nature of the problem and the influence of contextual information in a unified end-to-end model. Specifically, we introduce the Coarse-Fine-Coarse Transformer architecture and static map incorporation into ViT architecture.

The code for the paper can be found by the following link: `https://github.com/Future-Mobility-Lab/ViT-traffic-accident-risk`

A. Related works

In this study, we compare a set of diverse models to predict traffic accident risk. These models have shown effectiveness in capturing spatial-temporal patterns on the task of traffic accident risk prediction:

RNN-GRU 3: This model utilizes a variant of deep recurrent neural networks (RNN) known as the gated recurrent unit (GRU). It approaches the traffic accident risk forecasting problem by treating it as a time-series prediction task. The model includes a hidden state that allows it to keep track of long-term dependencies, making it particularly suited for time-series prediction tasks such as traffic accident risk forecasting.

SDCAE 1: The model is based on the stacked denoised convolutional auto-encoder architecture. This model is able to extract local spatial features from a city grid and, with the autoencoder structure, it can learn a compressed representation that captures the essential spatial patterns related to traffic accident risks.

H-ConvLSTM 20: The model that combines deep convolution layers with RNN-based LSTM layers. It extracts spatio-temporal features by using a sliding window over the city's grid cells. It relies on a sliding window approach over the grid cells of the city to understand the variations in spatial patterns over time.

GCN 19: The GCN model is a deep learning approach that leverages graph convolutional neural networks. It represents historical traffic accident data as a graph, allowing it to capture long-term spatio-temporal dependencies. Nodes represent different locations and edges indicate spatial proximity

Manuscript 1694 submitted to 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC). Received May 28, 2023.

or similarity, the model can uncover long-term connectivity-based patterns.

GSNet 14: a recent model that incorporates GCN, LSTM, and attention mechanisms to learn complex spatial-temporal correlations in traffic accident risk. It combines the strengths of graph convolution, recurrent modeling, and attention-based mechanisms. Currently, GSNet is considered the state-of-the-art method for the NYC and Chicago datasets.

C-ViT12: an our previously proposed baseline SoTA model 12, an example of application of Computer Vision model to non-vision task of accident risk prediction. The model utilizes a transformer-based architecture to predict traffic accident risk. It consists of three components: historical risk map encoding, historical contextual information encoding, and a transformer encoder. The historical risk maps are divided into image patches, which are individually passed through a linear embedding layer. The contextual information is encoded using a linear embedding layer. The transformer encoder, with its self-attention mechanism, captures global contextual dependencies across different patches, thereby enhancing the accuracy of future accident risk prediction. Compared to the existing state-of-the-art GSNet model, C-ViT demonstrated competitive performance while offering a more computationally efficient solution.

II. Methodology

Grid Representation: We model a specified city region, determined by latitude and longitude bounds, as a uniform grid. This grid comprises I rows and J columns, with each cell being identical in size.

Traffic Accident Risk: The risk of traffic accidents at time t for a specific grid cell i, designated Y_t^i, is quantified as the cumulative weighted sum of various types of traffic incidents that have transpired in that cell. Based on the classification provided in 14, traffic accidents are divided into three types, each assigned a specific weight: minor accidents are given a weight of 1, accidents causing injuries have a weight of 2, and fatal accidents receive a weight of 3. As an example, let's consider a grid cell that has seen two fatal accidents, one accident causing injuries, and four minor accidents. The cumulative traffic accident risk for this grid cell would then be calculated as $(2\times 3)+(1\times 2)+(4\times 1)=10$.

Problem Formulation: We redefine the traffic accident prediction problem from a standard time-series prediction task to an image regression task. We interpret the series of historical traffic accident risk maps, $\mathbf{Z}_{1:T}$ where $\mathbf{Z}\in\mathbb{R}^{I\times J}$ spans the time frame $[1:T]$, as an image X with a resolution of $I\times J$ and T channels. This image, combined with historical contextual data $C_{1:T}$, is input into our C-ViT model to generate a forecast of the accident risk map for the next hour, $\hat{\mathbf{Y}}_{T+1}$, with $\mathbf{Y}\in\mathbb{R}^{I\times J}$.

The Contextual Vision Transformer (C-ViT) - is optimized for traffic accident risk forecasting and consists of the following steps:

Historical Risk Map Encoding: This component takes in historical risk maps and encodes them into a sequence of 'patch embeddings' (as illustrated in Fig. 1), equally-sized image patches, each of which is processed individually through a linear patch embedding layer. These patches are created by dividing a unified single image of the city's grid into sub-spatial regions. To each of these patch embeddings, an additional 'regression token' and position embeddings are added. The regression token acts as an image representation while position embeddings provide sequence order information, both crucial for processing by the transformer encoder.

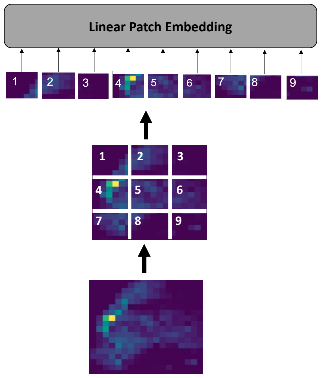

Fig. 1. Historical risk map encoding: we take a unified image X, divide it into equally-sized patches, and feed these patches individually into a linear patch embedding layer.

Historical Contextual Information Encoding: This module processes key contextual features, such as the time of day, the day of the week, whether it's a holiday, weather conditions, temperature, and traffic inflow/outflow. These features are encoded using a linear embedding layer of dimension 'D'. The output from this module is then fused with the output from the transformer encoder via a concatenation operation 14.

Transformer Encoder: This core unit of the C-ViT model features six layers, each composed of self-attention heads 13 and feed-forward fully connected sub-layers. Inside the encoder, multi-head self-attention processes the query, key, and value vectors based on their dot product, followed by softmax function application for determining their weights. This operation enables the model to handle complex spatial-temporal correlations within the data, improving accident risk predictions. In this paper, we propose two modifications to this architecture: 1) Addition of learnable Static Maps each of which concatenated either to accident risk map or attention layer to improve prediction performance.

A. Datasets

In our study, we utilize two publicly accessible real-world datasets for forecasting traffic accident risk: NYC[1] and Chicago[2].

The NYC dataset, reported from January 1, 2013, to December 31, 2013, consists of about 147K accidents, 173,179K taxi trips, 15,625 points of interest (POIs), 8,760 weather reports, and data about a road network consisting of

[1] https://opendata.cityofnewyork.us/

[2] https://data.cityofchicago.org/

103K segments. An additional feature unique to this dataset is its Point of Interest (POI) data, which provides information about specific locations like residences, schools, cultural facilities, recreation spots, social services, transportation hubs, and commercial centers.

The Chicago dataset, reported from February 1, 2016, to September 30, 2016, contains approximately 44K accidents, 1,744K taxi trips, 5,832 weather reports, and data about a road network comprising 56K segments.

Both datasets include historical traffic accidents and taxi trips data. The traffic accident data provides details about time, date, location (latitude and longitude), the number of causalities, weather condition (clear, cloudy, rainy, snowy, or mist), temperature, and road segment data (i.e., road length, width, and type). The taxi trip data, which includes location and times of pick-up and drop-offs, is utilized to compute the inflow/outflow of traffic condition in each area.

B. Baselines and Experiment Setup

In our study, we have followed a set of comprehensive steps for data pre-processing and have detailed the implementation of our proposed enahncement to C-ViT model. These steps ensure the validity of our research while facilitating comparison with previous studies. The specific steps of data preprocessing, and the parameters of the C-ViT model, are presented in the following table (Table I).

To assess the our model's performance, we used three key metrics commonly applied in traffic accident risk prediction: root mean squared error (RMSE), Recall, and mean average precision (MAP) 9, 14. RMSE measures the square root of the average of squared differences between the predicted and actual risk values. Recall calculates the fraction of actual accident-prone cells that were correctly identified by the model. MAP is the average of Precision achieved at the grid cell level. Metrics defined as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(Y_n - \hat{Y}_n\right)^2}, \tag{1}$$

$$\text{Recall } = \frac{1}{N}\sum_{n=1}^{N}\frac{H_n \cap A_n}{|A_t|}, \tag{2}$$

$$\mathrm{MAP} = \frac{1}{N}\sum_{n=1}^{N}\frac{\sum_{j=1}^{|A_t|}\mathrm{PR}(j)\times\mathrm{REC}(j)}{|A_n|}, \tag{3}$$

where N is the total number of samples to be evaluated, $Y_n, \hat{Y}_n$ are the ground truth and the predicted risk values for all grid cells of sample n respectively. A_n corresponds to the set of grid cells of sample n that have an actual/true traffic accident risk values. H_n corresponds to the set of grid cells within A_n with the highest traffic accident risk values.

TABLE I

SETUP FOR X-VIT MODEL AND PREPROCESSING PROCEDURES

Processing Step / Implementation Detail	**Specification**
Datasets	NYC, Chicago
Grid Representation	Each city map divided into grid cells of $2KM \times 2KM$
Accident Grouping	All accidents in each grid cell grouped based on location and duration time
Data Split	Training: 60%, Validation: 20%, Testing: 20%
Overlapping Accident Control	No overlapping accidents based on time
Data Standardization	Mean and standard deviation normalization
Traffic Accidents Periodicity	1 hour
Historical Traffic Risk Map Size	$7 \times 20 \times 20$
Historical Accident Risks	7 (most recent accident risks in past 3 hours + past accident risks in the last 4 weeks)
Prediction Horizon of Traffic Accident Risk	1 (next hour)
Dimension (D) of Linear Patch Embedding	64
Dimension (D) of Position Embedding Layer	64
Dimension (D) of Linear Embedding Layer of the Historical Contextual Encoder	64
Resolution of Input Patches (P) to the Patch Embedding Layer	5
Number of Self-Attention Heads	8
Dimension of the Final Output Fully Connected Layer	128
Optimization Function	Weighted Mean-Squared Error (MSE)
Loss Weighting Procedure	Focal loss
Risk Value Classes	0, 1, 2, ≥3
Loss Function Weights	0.05, 0.2, 0.25, 0.5
Training Epochs	200
Optimizer	Adam
Learning Rate	0.003
Batch Size	32

III. COARSE-FINE-COARSE VISUAL TRANSFORMER (CFC-VIT)

One of the issues to solve in the topic of accident risk prediction is the zero-inflated issue - the imbalance between the amount of non-zero and zero accident risk cells. This issue can be resolved by using a comparison mask or variations of focal loss 14. Another issue, which is usually ingnored is the fine granularity of accident risk map. For example, in the grid representation, cells can be separated and of minimal 1x1 cell size (see Fig. 2).

Current computer vision methods applied to the task of accident risk prediction produce 'blurred' results due to intrinisic limitations of convolution network architectures 7, 5. To resolve this issue we propose an alternative approach which consists of up-scaling the patches before the embedding to allow fine-grained processing by internal layers of transformer, and then down-scaling embedding to match the original output shape 3. Up-scaling may be a necessary step in the of use of asymmetric convolutional networks

TABLE II

PERFORMANCE EVALUATION OF OUR C-VIT MODEL AGAINST A NUMBER OF BASELINE APPROACHES FROM THE LITERATURE OVER THE NYC AND CHICAGO DATASETS.

Dataset	NYC			Chicago		
Model	RMSE ↓	Recall ↑	MAP ↑	RMSE ↓	Recall ↑	MAP ↑
RNN-GRU 3	8.3375	28.09%	0.1228	12.6482	17.83%	0.0664
SDCAE 1	7.9774	30.81%	0.1594	11.3382	18.78%	0.0753
H-ConvLSTM 20	7.9731	30.42%	0.1454	11.3033	18.43%	0.0716
GCN 19	7.7358	31.78%	0.1623	11.0835	18.95%	0.0805
GSNet 14	7.6151	33.16%	0.1787	11.3726	19.92%	0.0822
C-ViT (previous)	**7.0053**	**33.86%**	**0.1875**	**9.4456**	**20.93%**	**0.0980**
SARM-ViT (new)	**7.05**	**33.72%**	**0.19**	**9.01**	**22.24%**	**0.11**
CFC-ViT (new)	**6.81**	**32.15%**	**0.1838**	**8.62**	**19.38%**	**0.08**

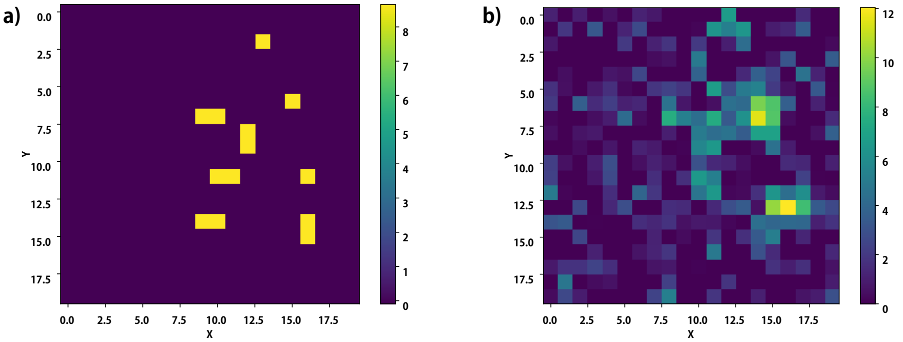

Fig. 2. Example of GSNet predictions (after training for 2 epochs, when the best performance is observed): a) Actual map of the accident occurrence b) Predicted map of the accident occurrence.

for segmentation to increase the detailization of results 8 since there are no upscaling layers at the final part of the asymmetric network. We upscale patches before the embedding to perform fine-grained processing of these patches. The dimensionality of embedding is also increased proportionally to the patch size; processed embedding then downscaled by the same rate. This allows the network to form intermediate results of higher dimensionality, which when down-scaled, will produce more fine-grained image.

Results for the Chicago data set show a significant improvement in the RMSE metric results both for 2x and 4x scale factors (see Table III where the RMSE is 8.62 as compared to 9.45 translating in a 8.78% improvement). There is an inverse dependence observed between the scale factor and the Recall or MAP metrics: the increase in the scale factor lowers RMSE but MAP and recall also decrease. However, given the robustness of the RMSE metric, the improvement is consistent.

Results for the NYC data set show that the prediction performance can increase at a specific scale factor (2x) and decrease at different scale factor (4x) (see Table IV. These results suggest that the optimal scale factor for each data set can exist, which leads to a deciated optimization task of finding the optimal scale factor value. Results both for NYC and Chicago data sets show a non-linear dependency between the RMSE, the MAP (mean average precision) or Recall metrics. These metrics are intended for different purposes (RMSE for the regression, MAP and Recall for the classification results) and therefore can produce different results based on the characteristics of the predicted values.

Overall, our new proposed CFC-ViT approach shows an improvement in the RMSE results, but these results and other metrics depend on the scale factor parameter. The optimal scale factor can vary for each data set and can be found using other optimization techniques.

IV. APPLICATION OF THE STATIC MAP GENERATION

The use of Attention layers is a computer vision technique which implies an estimation of attention maps from different images. Since each image may have different areas of attention, the attention map is generated for every case of prediction (which we can call the dynamic attention estimation). But in the case of accident risk prediction, we predict on the same area each time. Therefore, we can use the statically generated attention map (static attention estimation). We evaluate multiple scenarios of combining dynamic (DA) and static attention (SA) estimations using varying combination operations. To further utilise the advantage of a non-volatile area, we also try to generate the Static Accident Risk Map (S-ARM) so our network needs to predict the offset of the accident risk (relative accident risk) from the statically generated risk map values instead of predicting the absolute accident risk values. Therefore, another contribution of this

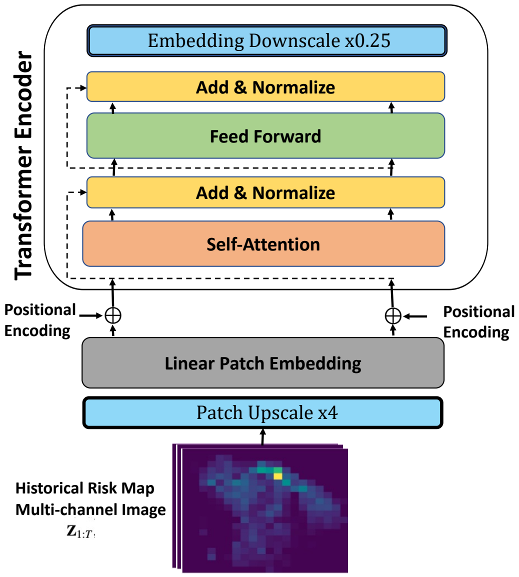

Fig. 3. Coarse-Fine-Coarse Transformer

TABLE III
PERFORMANCE EVALUATION OF OUR CFC-VIT MODEL ON CHICAGO DATA SET

	RMSE	Recall	MAP	HFT-RMSE	HFT-Recall	HFT-MAP	Data set	CFC Scale factors
0	**8.62**	19.38	0.08	**6.48**	19.89	0.09	chicago	4x, 0.25x
1	9.24	18.84	0.06	6.85	20.85	0.07	chicago	2x, 0.5x
–	–	–	–	–	–	–	–	–
-	9.45	20.93	0.098	7.035	21.95	0.125	chicago	1x,**baseline (multi-epoch)**

TABLE IV
PERFORMANCE EVALUATION OF OUR CFC-VIT MODEL ON NYC DATA SET

	RMSE	Recall	MAP	HFT-RMSE	HFT-Recall	HFT-MAP	Data set	CFC Scale factors
0	7.09	33.17	0.1808	6.45	33.90	0.1751	NYC	4x, 0.25x
1	**6.81**	32.15	0.1838	**6.14**	33.24	0.176	NYC	2x, 0.5x
–	–	–	–	–	–	–	–	–
-	7.0053	33.86	0.1875	6.2658	34.46	0.1802	NYC	1x, **baseline (multi-epoch)**

work is to further combine the Predicted Offset Accident Risk Map (PO-ARM) with the Static Accident Risk Map (S-ARM) (see Fig. 4).

A. *Pipeline description*

The generalisation performance of the Transformer model can be greatly improved by using one-epoch training 6. Therefore we use results obtained from one-epoch training in further scenarios. Other parameters of the setup are the same as in Section II-B.

B. *Description of combination operations*

The use of static map generation at the beginning of the attention layer as well as near the network output can remove the necessity for the network to predict the absolute risk values (static map is assumed to act as a static image and network is required to predict the relative risk from the one in a static map). We test multiple different approaches to achieve the benefit of using the static map generation. Different constraint functions can be used to limit the range of values observed from the static map. Also, the actual static map can be combined differently with the final and the intermediate network values.

Combination operations for the attention layer that we have considered:

1) None - using only the original pipeline structure with no static map,
2) tanh(map)+x - the static map is bounded by the tanh function in order to obtain the static map values distribution between (-1,+1), combined with layer input values,
3) tanh(map)*x - the same as above, but combined using a multiplication operation,
4) map - we use the static map instead of the attention

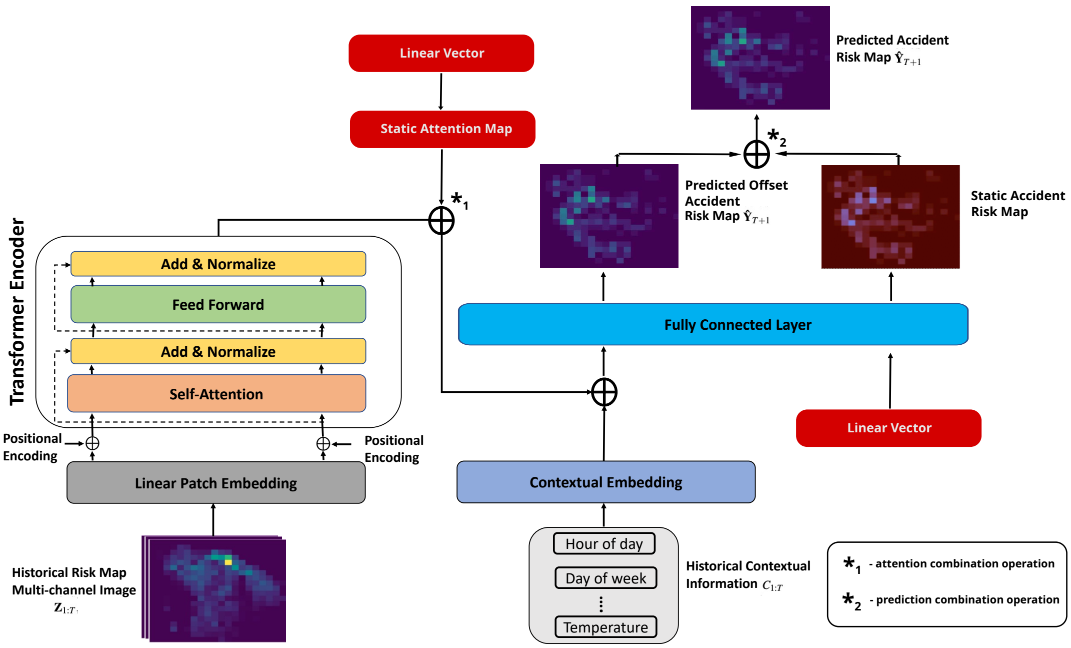

Fig. 4. The building blocks of our proposed XViT model with Static Map Generation

layer inputs,

5) map + x - we combine the static map with the attention layer inputs using the "plus (+)" operation,
6) sigmoid(map)+x - static map values are distributed between (0,+1) and are further combined with the attention layer inputs by using the "plus (+0)" operation (a linear offset combination),
7) sigmoid(map)*x - the same as above, but combined using the "multiplication (*)" operation.

Combination operations for the network output that we have implemented:

1) tanh(map)+x in which the static map values (map) are combined with the intermediate predictions (transformer output - x),
2) sigmoid(map)+x - same as above, but using sigmoid as a static map constraining function,
3) tanh(map)*x - tanh is used as a static map constraining function, and the static map values (map) are combined with the intermediate predictions by using the "multiplication (*)" operation,
4) sigmoid(map)*x -
5) head(map+non) - the static map values combined with the non-risk features and passed through the feed forward neural network,
6) head(map)+head2(x+non) - the static map is passed through a separate feed forward neural network, while other predictions together with the non-risk features are passed through the second network of the same structure,
7) head(map)*head(x) - the accident features are passed through the same network as the static map values and then combined using the "multiplication (*)" operation,
8) head(non)+head(x) - the non-risk features and the accident risk features are passed through the same network, and then combined using the plus operation,
9) None (head(x+non))- this is the original ViT implementation.

The constraint functions (tanh and sigmoid) are tested with the assumption that values close to the actual normalised accident risk values will be observed right after the network parameter initialisation. Due to the variation, we name this derivative model an XViT model.

C. S-ARM results

The results for the Chicago and NYC data sets are provided in Tables V-VI. The results are also provided for the high-frequency hours (HFT) - meaning the RMSE errors obtained only when using the HFT hours when more traffic is normally expected in the city. The use of the static map generation didn't show an improvement on the NYC data set. In fact, the results are a bit worse but closely realted to the baseline (7.05 RMSE for the best combination vs 7.00 RMSE using original baseline multi-epoch approach); however we observe that there is an improvement in the recall results for the high-frequency hours (34.84 for the best combination vs 34.25 when using the multi-epoch baseline). But results forthe Chicago data set show a very significant improvement across all the metrics (e.g. from 9.45 to 9.01 in RMSE, from 20.93 to 22.24 in Recall, from 21.95 to 23.46 in HFT-Recall). More than that, the 1-epoch training also shows a significant improvement in case of the baseline ViT structure (from 9.45 to 9.25 RMSE, from 20.93 to 21.77 Recall, from 7.035 to 6.93 in HFT-RMSE).

This slight reduction in the model performance in case of

TABLE V
PERFORMANCE EVALUATION OF OUR XVIT MODEL FOR A NUMBER OF COMBINATION OPERATIONS ON NYC DATA SET. TOP 7 RESULTS.

	RMSE	Recall	MAP	HFT-RMSE	HFT-Recall	HFT-MAP	Data set	S-ARM Combination Operations
0	7.05	33.72	0.19	6.46	34.84	0.18	nyc	**tanh(map)*x, tanh(map)+x**
1	7.07	33.49	0.19	6.48	34.43	0.18	nyc	sigmoid(map)+x, sigmoid(map)+x
2	7.10	33.21	0.19	6.50	33.87	0.18	nyc	sigmoid(map)+x, head(map)+head(x)
3	7.10	33.57	0.19	6.50	34.15	0.18	nyc	sigmoid(map)+x, tanh(map)+x
4	7.11	33.26	0.19	6.49	33.76	0.18	nyc	none, tanh(map)*x
5	7.11	33.38	0.19	6.52	34.53	0.19	nyc	sigmoid(map)*x, tanh(map)+x
6	7.11	33.39	0.19	6.52	34.71	0.18	nyc	tanh(map)*x, sigmoid(map)+x
7	7.11	33.85	0.19	6.50	34.81	0.19	nyc	sigmoid(map)*x, tanh(map)*x
–	–	–	–	–	–	–	–	–
-	7.25	33.39	0.19	6.53	34.25	0.19	nyc	**baseline (1-epoch)**
-	7.0053	33.86	0.1875	6.2658	34.46	0.1802	nyc	**baseline (multi-epoch)**

the NYC data set and significant improvement in case of the Chicago data set can be interpreted through the concept of local optima and data set size. There may be multiple local optima for the accident risk approximation across historical accident risk records (e.g. multiple average risk maps for different months). This optima can have an ability to show a good approximation of the accident risk, but since the road networks and the city structures change over time, different local optima can appear over time as well. So finding just one static accident map may not be optimal for a large data set, but may show benefit in the case of small data set (Chicago has just 44K accident records in comparison to 147K for NYC attributing to 1 full year of records and these are mostly short-time accidents in Chicago - just 8 months). We conclude that there is evidence that the proposed method and the use of multiple static maps can be a topic of the future research which can bring improvement over large data sets.

Another important observation is that the same set of combination operations gives the best results in the case of the ViT network with a generated static map: "tanh(map)*x" in the attention layer and "tanh(map)+x" near the network output. This not only signifies the use of the constraint function tanh, but also shows where to use each combination operator (addition and multiplication). We also observe that the use of non-risk features is not present among the top 20 results for NYC data set (see Table V), while for the Chicago data set it is present in 9 combinations out of 20, which may indicate the difference in quality of these features in both data sets.

V. CONCLUSION

In this work, we propose a series of enhancements to our previously proposed approach for the task of traffic accident risk forecasting. In our approach we formulated the problem as an image regression problem and introduced a unique contextual vision transformer network (C-ViT) that can efficiently model the traffic accident risk forecasting task from both spatial and temporal perspectives. The proposed approach and its enhancements have been evaluated on two publicly available data sets for the traffic accident risk problem. The combination of static accident risk map with the ViT model (XVit) provides an even more significant improvement over the previous method in case of the New-York data set, thus establishing the new SoTA in the study area. The operation combination method has a potential for improvement (e.g. more different combination methods and constraint functions can be tested). Improvements in results obtained in the current research can also highlight the applicability of vision transformers for non-visual tasks. The Coarse-Fine-Coarse Visual Transformer (CFC-Vit) architecture allows for fine-grained processing of the accident risk map and introduces an additional scale factor parameter which affects (and may improve) the prediction performance. Overall, the use of visual transformers and its variations for traffic accident risk prediction outperforms previously used approaches. Further applications of image and video processing methods may provide further improved results and open alternative approaches for the task of accident risk prediction.

ACKNOWLEDGMENTS

This work has been done as part of the ARC Linkage Project LP180100114. This research is funded by iMOVE CRC (project 5-013) and supported by the Cooperative Research Centres program, an Australian Government initiative."

REFERENCES

[1] Chao Chen et al. "Sdcae: Stack denoising convolutional autoencoder model for accident risk prediction via traffic big data". In: *2018 Sixth International Conference on Advanced Cloud and Big Data (CBD)*. IEEE. 2018, pp. 328–333.

[2] Quanjun Chen et al. "Learning deep representation from big and heterogeneous data for traffic accident inference". In: *Thirtieth AAAI conference on artificial intelligence*. 2016.

[3] Junyoung Chung et al. "Empirical evaluation of gated recurrent neural networks on sequence modeling". In: *arXiv preprint arXiv:1412.3555* (2014).

[4] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

TABLE VI

PERFORMANCE EVALUATION OF OUR XVIT MODEL FOR A NUMBER OF COMBINATION OPERATIONS ON CHICAGO DATA SET. TOP 7 RESULTS.

	RMSE	Recall	MAP	HFT-RMSE	HFT-Recall	HFT-MAP	Data set	S-ARM Combination Operations
0	9.01	22.24	0.11	6.80	23.46	0.13	chicago	**tanh(map)*x, tanh(map)+x**
1	9.06	22.30	0.10	6.85	23.59	0.13	chicago	tanh(map)*x, head(map)+head(x)
2	9.09	22.30	0.10	6.81	23.18	0.11	chicago	none, head(non)+head(x)
3	9.10	21.77	0.10	6.80	22.63	0.13	chicago	none, head(map)+head2(x+non)
4	9.12	21.88	0.10	6.80	23.05	0.13	chicago	sigmoid(map)*x, head(map)+head2(x+non)
5	9.14	22.90	0.11	6.82	24.14	0.12	chicago	sigmoid(map)*x, head(non)+head(x)
6	9.15	22.12	0.11	6.91	22.63	0.13	chicago	none, sigmoid(map)+x
7	9.15	22.18	0.11	6.94	22.91	0.13	chicago	tanh(map)*x, sigmoid(map)+x
–	–	–	–	–	–	–	–	–
16	9.25	21.77	0.10	6.93	22.36	0.12	chicago	**baseline (1-epoch)**
-	9.45	20.93	0.098	7.035	21.95	0.125	chicago	**baseline (multi-epoch)**

[5] Jiuxiang Gu et al. “Recent advances in convolutional neural networks”. In: *Pattern recognition* 77 (2018), pp. 354–377.

[6] Aran Komatsuzaki. “One epoch is all you need”. In: *arXiv preprint arXiv:1906.06669* (2019).

[7] Zewen Li et al. “A survey of convolutional neural networks: analysis, applications, and prospects”. In: *IEEE transactions on neural networks and learning systems* (2021).

[8] Shao-Yuan Lo et al. “Efficient dense modules of asymmetric convolution for real-time semantic segmentation”. In: *Proceedings of the ACM Multimedia Asia*. 2019, pp. 1–6.

[9] Chen Ma et al. “Point-of-interest recommendation: Exploiting self-attentive autoencoders with neighbor-aware influence”. In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 2018, pp. 697–706.

[10] National Highway Traffic Safety Administration. *Traffic safety facts 2013*. U.S. department of transportation, 2013.

[11] Honglei Ren et al. “A deep learning approach to the prediction of short-term traffic accident risk”. In: *arXiv preprint arXiv:1710.09543* (2017).

[12] Khaled Saleh, Artur Grigorev, and Adriana-Simona Mihaita. “Traffic Accident Risk Forecasting using Contextual Vision Transformers”. In: *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE. 2022, pp. 2086–2092.

[13] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).

[14] Beibei Wang et al. “GSNet: Learning Spatial-Temporal Correlations from Geographical and Semantic Aspects for Traffic Accident Risk Forecasting”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 5. 2021, pp. 4402–4409.

[15] Senzhang Wang et al. “Traffic Accident Risk Prediction via Multi-View Multi-Task Spatio-Temporal Networks”. In: *IEEE Transactions on Knowledge and Data Engineering* (2021).

[16] World Health Organization. *Global status report on road safety 2015*. World Health Organization, 2015.

[17] World Health Organization. *Global status report on road safety 2018: summary*. World Health Organization, 2018.

[18] Bichen Wu et al. “Visual transformers: Token-based image representation and processing for computer vision”. In: *arXiv preprint arXiv:2006.03677* (2020).

[19] Zonghan Wu et al. “Graph wavenet for deep spatial-temporal graph modeling”. In: *arXiv preprint arXiv:1906.00121* (2019).

[20] Zhuoning Yuan, Xun Zhou, and Tianbao Yang. “Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data”. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2018, pp. 984–992.

[21] Zhengyang Zhou et al. “Foresee urban sparse traffic accidents: A spatiotemporal multi-granularity perspective”. In: *IEEE Transactions on Knowledge and Data Engineering* (2020).

[22] Zhengyang Zhou et al. “RiskOracle: a minute-level citywide traffic accident forecasting framework”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 01. 2020, pp. 1258–1265.