Proceedings of Machine Learning Research 123:232–241, 2020 NeurIPS 2019 Competition and Demonstration Track

The surprising efficiency of framing geo-spatial time series forecasting as a video prediction task – Insights from the IARAI *Traffic4cast* Competition at NeurIPS 2019

David P Kreil David.Kreil@iarai.org
Institute of Advanced Research in Artificial Intelligence (IARAI), Landstraßer Hauptstraße 5, 1030 Vienna, Austria

Michael K Kopp Michael.Kopp@iarai.org
Institute of Advanced Research in Artificial Intelligence (IARAI) & HERE Technologies Zurich

David Jonietz David.Jonietz@here.com
HERE Technologies Zurich

Moritz Neun Moritz.Neun@here.com
HERE Technologies Zurich

Aleksandra Gruca aleksandra.gruca@polsl.pl
Department of Computer Networks and System, Silesian University of Technology, Gliwice, Poland

Pedro Herruzo* pherruzo@ac.upc.edu
Dept. of Computer Architecture, Polytechnic University of Catalonia, North Campus, C6 building, 08034 Barcelona, Spain

Henry Martin martinhe@ethz.ch
Institute of Cartography and Geoinformation, ETH Zurich, Stefano-Franscini-Platz 5, 8093 Zurich, Switzerland

Ali Soleymani ali.soleymani@here.com
HERE Technologies Zurich

Sepp Hochreiter Sepp.Hochreiter@iarai.org
Institute of Advanced Research in Artificial Intelligence (IARAI) & Johannes-Kepler University Linz

Editors: Hugo Jair Escalante and Raia Hadsell

Abstract

Deep Neural Networks models are state-of-the-art solutions in accurately forecasting future video frames in a movie. A successful video prediction model needs to extract and encode semantic features that describe the complex spatio-temporal correlations within image sequences of the real world. The *IARAI Traffic4cast* Challenge of the NeurIPS Competition Track 2019 for the first time introduced the novel argument that this is also highly relevant for urban traffic. By framing traffic prediction as a movie completion task, the challenge requires models to take advantage of complex geo-spatial and temporal patterns of the underlying process. We here report on the success and insights obtained in a first Traffic Map Movie forecasting challenge. Although short-term traffic prediction is considered hard, this novel approach allowed several research groups to successfully predict future traffic states in a purely data-driven manner from pixel space. We here expand on the original ratio-

* Work performed at SEAT, S.A.

nale, summarize key findings, and discuss promising future directions of the *Traffic4cast* competition at NeurIPS.

1. Introduction

The emergence of interconnected devices, low-cost sensors and the ability to gather data in large volumes has provided us with global-scale environmental data sets allowing us to reason about and understand the underlying geo-spatial processes in new ways, including precipitation (Agrawal et al., 2019) and weather, pollution, and urban traffic dynamics (Nittel, 2009). This provides exciting new opportunities for machine learning (ML) to tackle real-world problems of high societal impact.

Our *Traffic4cast* Challenge at the NeurIPS Competition Track 2019 provides a unique large real-world data set in order to advance modelling the spatially-explicit task of traffic prediction, which is still considered as largely unsolved due to the highly complex underlying systems (Guo et al., 2019). We encouraged contestants to predicting traffic flow volumes, velocities, and the dominating flow directions. Such forecasts form the basis for building and managing our cities to provide efficient and sustainable mobility (Bucher et al., 2019; Jonietz et al., 2018). In this competition, we focus exclusively on short-term traffic prediction, *i. e.*, with a temporal prediction horizon of 15 minutes (see *e. g.* Ermagun and Levinson, 2018b; Dunne and Ghosh, 2011; Ermagun and Levinson, 2018a; Lana et al., 2018)

Moreover, we propose a novel approach to traffic forecasting: we aggregate our traffic data from individual sensor measurements in space and time bins and thus forego, say, the need for information about the underlying road network, which are unavailable in many emerging economies. We thus represent the dynamics of a complex traffic system as a sequence of changing views of a map, with each pixel corresponding to a fixed area of space and each frame summarizing a discrete time bin.

With this data representation the prediction task resembles a video frame prediction. Video forecasting is a highly active field with many distinct approaches (see *e. g.* Srivastava et al., 2015; Lee et al., 2018; Kwon and Park, 2019; Walker et al., 2016; Xue et al., 2016; Han et al., 2019). Given, though, that motion dynamics in movies hold substantial redundancies, with large areas of correlated pixels in a frame, it was unclear how well these techniques might generalize to this new setting or to other geo-spatial processes. In particular, geo-spatial time series are known to be governed by somewhat more complex underlying spatio-temporal dependencies, a matter often referred to as 'spatial is special' (Anselin, 1989). Nevertheless, the result of this competition suggest that this approach has merit and should be thoroughly explored. This is corroborated by more recent independent work on precipitation prediction that used both a similar discretized geo-temporal data representation and a similar model architecture to our winning approaches (Agrawal et al., 2019).

We set out to further elaborate on these ideas and critically evaluate the learnings from our *Traffic4cast* competition: after a brief review of selected prior work from traffic forecasting and video prediction, we will recapitulate the motivation for and setting of our *Traffic4cast* competition. In the following, selected competition submissions will be reviewed and discussed. Finally, we will provide an outlook to potential next steps to build on the successful first year of the competition and outline planned advances.

2. Competition Overview

2.1. Data

We provide a unique real-world industrial-scale data set derived from trajectories of raw GPS position fixes, consisting of a latitude, a longitude, a time stamp, as well as the vehicle speed and driving direction recorded at the time. The data is made available by HERE Technologies (www.here.com) and originates from a large fleet of probe vehicles which recorded their movements in multiple culturally and socially diverse metropolitan areas around the world (Berlin, Moscow, and Istanbul) throughout the course of the entire year 2018. We shared over 300,000 frames with the scientific community based on over 10^{11} probe-points. Playing these frames at a rate of 24 frames/s this would exceed 3h of data movie footage.

A simplified schema of the encoding process is shown in Fig. 1 and an illustrative example of a frame sequence can be seen in Fig. 2. Specifically, the aggregation procedure involves the following steps (which are carried out for each city):

- Spatial tessellation of the study area: the study area is tessellated into regular grid cells. The size of the grid cells in the provided data set is roughly 100*100 meters.
- Aggregation of probe points: Probe points (tuples of location, time stamp, measured speed and heading) are grouped based on their spatial and temporal attributes, *i. e.*, the grid cell their location falls into and the 5-minute time bin within which their time stamp belongs.
- Core channels: In each grid cell and time bin, we compute the following three features:
  - Volume: The number of probes points recorded from the collection of HERE sources. The values are first filtered to remove noisy data by, respectively, flooring and capping at a minimum and a maximum value.
  - Mean speed: The average speed from the collected probe points. The values are capped at a maximum level to remove noisy/false recordings.
  - Main heading direction: Each probe point records the heading direction (values from 0 to 359). The points are binned in four heading directions of North-East (0-90), South-East (90-180), South-West (180-270) and North-West (270-0) and the major heading bin, *i. e.*the bin with the highest number of points is selected.

  After calculation of the above three features, they are normalized over the range of [1-255] to map to the RGB channels, where Volume represents R channel, Speed G channel and Heading B channel. Volume values are normalized from [Vol_{min} -Vol_{max}] to [1-255], Speed values from [0-Speed_{max}] to [0-255] and Heading from [0-359] to the four distinct values of [1,85,170,255] in the order of [NW,NE,SW,SE]. To indicate *NoData* in a cell, all channels are filled with *0*, therefore [*0,0,0*] in the three-channel representation indicates no data. This can clearly be distinguished from cases with non-zero volume, where average speed or average heading may nevertheless be zero.
- Generation of video frames: Finally, the encoded values are stored in a tensor of the form (t, h, w, c) where t is the number of individual 5 minute time bins, *i. e.*, the

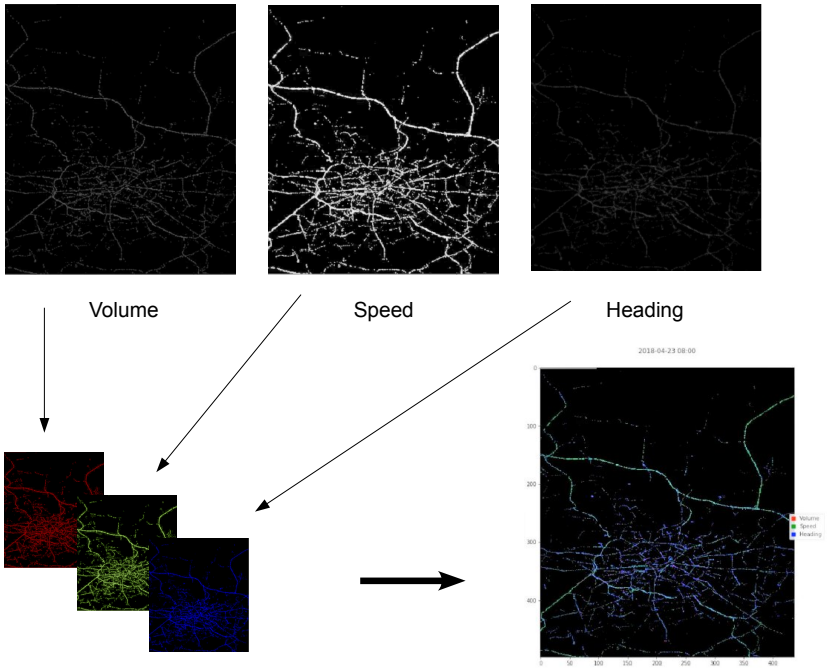

Figure 1: Encoding Volume, Speed, and Heading in an RGB frame

number of frames, while h and w denote the height and width of the frame grid cells, and c stands for the number of data channels, with $c = 3$ when using only a single channel for each traffic state feature (as in Fig. 1), or $c > 3$ when using complementary data. Note that the simple $c = 3$ case is sufficient for the core competition task.

For the prediction challenge, 72 days are randomly chosen as test days, 285 days for training and 7 for validation. For each of the test days, we zero out all frames except for 5 one-hour long time intervals roughly set at time intervals where one could expect rush hour- as well as free-flow situations. These one-hour intervals will serve as inputs for the predictive models, the task of which is to predict the immediately preceding future states.

2.2. Task and Metric

The main task of the *Traffic4cast* 2019 core competition then was the large-scale short-term traffic state forecasting, *i. e.*, the prediction all three traffic state variables (speed, direction, and volume) for the missing 3 frames representing the traffic state 5, 10, and 15 minutes into the future for the whole city. Predictions were scored by the mean squared error (MSE) of the predicted pixel values compared to the ground truth. The MSE per frame was summarized by city, and the three scores averaged for the leaderboard ranking. We purposefully selected the MSE metric in spite of the obvious drawbacks this might have on the heading channel where, maybe, a discrete metric might seem more appropriate.

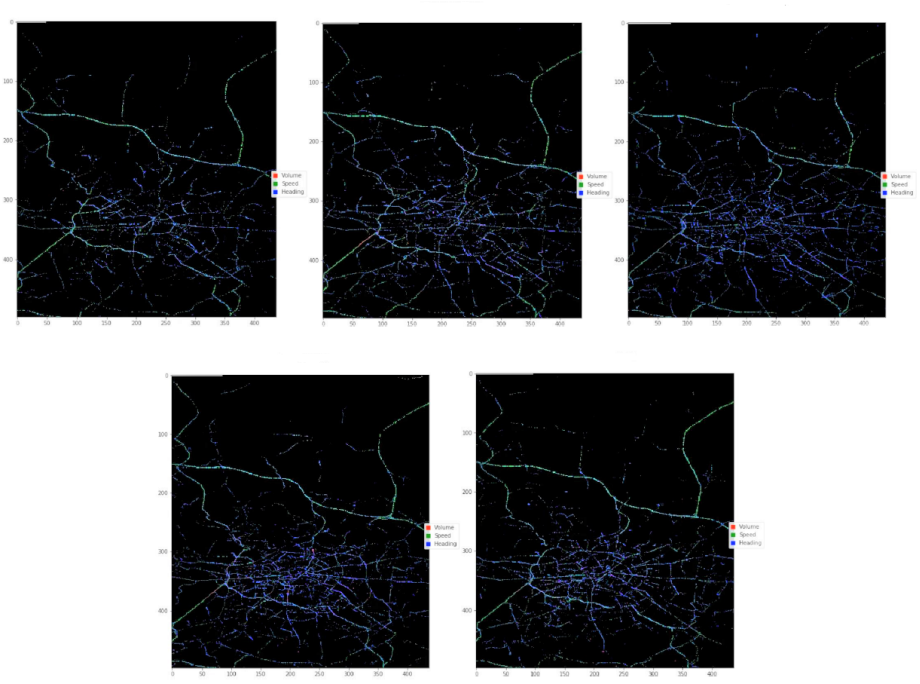

Figure 2: Exemplary frames extracted for Berlin, Germany

2.3. Winning submissions

In total, over 40 teams of international researchers submitted more than 500 submissions to the competition. The final leaderboard[1] reveals both the drastic performance differences between the baseline solutions provided by the organizers (such as a simple moving average predictor) and the submitted models, as well as the small score differences between the top submissions. From the perspective of a leaderboard, the question arises if these differences are still significant, and this was studied in depth by Gruca et al. (2020). While the specific differences between the summary scores may indeed appear small, a statistical analysis of the multiple independent tests compounding the overall score confirmed consistent differences between the leading submissions (Fig. 3).

Focusing on the top 5 models, it is striking that several were based on a Unet architecture with skip connections as proposed by Ronneberger et al. (2015), including the top-scoring model by Choi (2019) and the runner-up by Martin et al. (2019). While the latter abide by the original architecture, Choi (2019) applies several changes. First, he replaces the blocks of convolutional layers, RELU activation and max pooling operations by the densely connected convolutional blocks of Huang et al. (2017), where additional skip connections are introduced between all convolutional layers. Secondly, he trains a total of 4 separate models as an ensemble. Interestingly, Martin et al. (2019) report that alternative approaches which incorporated rich additional geographical information such as the road network failed to outperform the vanilla Unet. The team of Yu et al. (2019) achieved rank 3 with a

1. https://www.iarai.ac.at/traffic4cast/competitions/traffic4cast-2019-core/?leaderboard

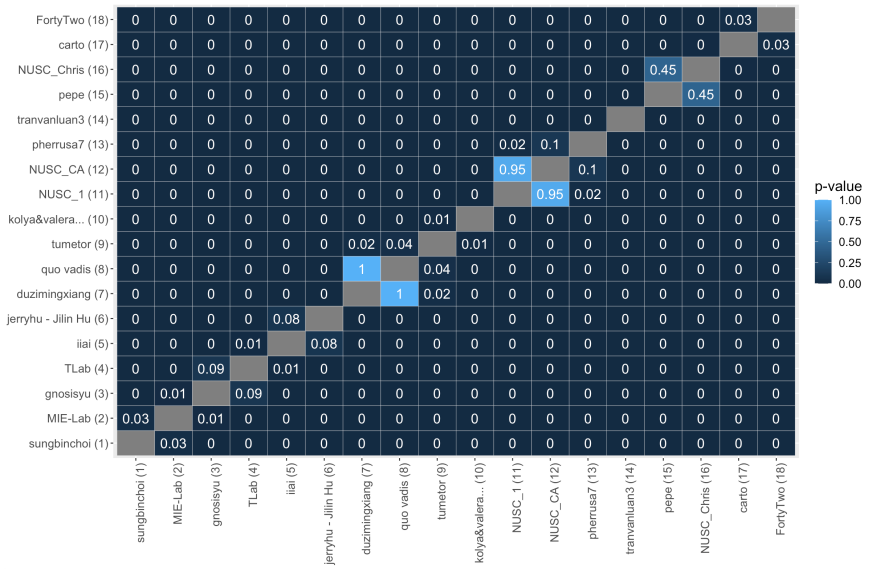

Figure 3: Significance of pairwise compound score rank differences between teams. For each team, their respective best submission was included. The 18 teams with above median performance were considered. Significance analysis considered the rank concordance of the independent test scores compounding the leaderboard score. A Friedman rank test confirmed significance of the overall ranking ($p < 10^{-15}$). The pairwise mean rank *post-hoc* test proposed by Demšar (2006) indicated significant differences for most algorithm pairs after Benjamini and Hochberg (1995) correction for multiple testing when considering high-resolution scores, see Gruca et al. (2020). Note that although ranks 1–3 were significantly different from one another ($p < 3\%$), the performance differences between teams ranked third and fourth were only of marginal significance ($p < 9\%$).

bijective two way autoencoder and predictor that preserves information flow through the convolutional layers.

In summary, Unet-type encoder-decoder model architectures based on multiple stacked convolutional layers constantly outperformed other popular approaches such as convLSTM or attention-based models. The impressive performance of the top-scoring Traffic Map Movie based models is in line with current state-of-the-art in traffic prediction, where DL models have constantly demonstrated better capability of handling the non-linearity and stochastic character of traffic dynamics.

3. Conclusion and Outlook

The *Traffic4cast* competition aimed to bring together scientists from the fields of ML, GISc, and transportation science to explore the following research questions:

- Even if deliberately ignoring additional geographical information such as the underlying road graph, does our aggregated map representation still preserve the main spatio-temporal patterns of traffic to be extracted and learned by neural networks?
- Do we trivialize the predictive task by aggregating the data in space and time? How applicable are state-of-the-art DL models from computer vision (especially video prediction) to this kind of data?
- Is there a one-solves-all model architecture such that it can be expected to learn the underlying rules of a broad range of geo-spatial processes from such video-like data?

With regards to the first question, the results achieved by the top-tier model submissions clearly demonstrate outstanding predictive accuracy, and therefore implies the existence of non-random patterns in our data to be exploited by the models. Both the spatio-temporal aggregation of the available sensor data and the deliberate ignorance of the road network – a highly unusual approach in this domain – did not obfuscate the spatio-temporal regularities and dependencies underlying traffic dynamics. The fact that at least one participating team reported no performance benefits of adding additional geographical information further supports that hypothesis. In our view, the presence of clear patterns in the aggregated data further indicates that this representation could also be useful for other analytical tasks besides prediction, such as anomaly detection or missing data imputation. Even if subsequent applications would require traffic state variables estimated on the level of road segments, a model could potentially be used to translate between these representation formats.

At the same time, however, the competition outcomes demonstrate that simpler baselines are clearly outperformed by more complex DL models with several tens of millions of parameters. In our view, this indicates that our novel approach did not trivialize the task of traffic prediction, and highlights a continuing need for high-capacity models. The success of the Unet architecture – a state-of-the-art approach for many tasks related to computer vision – can be explained by the interplay of deep encoding of spatio-temporal regularities using the convolutional layers, while the skip connections still allow high-frequency information to be preserved in the model. Therefore, and in view of the more recent success of a similar architecture for precipitation prediction (Agrawal et al., 2019), the Unet and related encoder-decoder models might be promising candidates for a one-solves-all architecture, although it is clear that more research is needed.

Since the *Traffic4cast* conference session at the NeurIPS 2019 Competition Track additional interesting results have been reported and further work is under way. The solution of Herruzo and Larriba-Pey (2020) uses a recurrent autoencoder with skip connections that allowed for exogenous variables to be added. They found that, in this setting, summarial weather or time information yielded little performance increase, notably suggesting that this information is indirectly already encoded in the traffic itself. Similarly, Martin et al. (2020) implemented several graph neural network (GNN) approaches and compared them to their Unet solution that reached second place in the *Traffic4cast* leaderboard. These GNNs did not manage to perform on par with their Unet model. Yet, tantalizingly, they indicate an ability to support transfer learning from on city to another, by beating the provided baseline predictors for Berlin and Istanbul having only trained their ResNet based GNN

on Moscow. In Yu et al. (2020), our third-ranked team employed their method on other well known data challenges (KITTI, Moving MNIST) and obtained state-of-the-art results. The fact that their two-way autoencoder using a reversible architecture is self-supervised and memory efficient while still maintaining no-information loss during feature extraction makes it a serious candidate for being used as a generative pre-training strategy in other downstream tasks.

Taking the competition into the next year, for *Traffic4cast* 2020 we will substantially increase data extent and types to allow both deeper analyses and new questions of interest to be studied. Building on the now freshly established approach of Traffic Map Movies, additional cities can add both diversity in scale, covering smaller and larger urban settings, as well as diversity in cultural / geographic background, including cities from other continents and emerging economies. This will let us ask interesting questions about the similarities and differences in urban traffic, as well as explore master models trained on multiple cities in parallel. The latest results from a companion paper by Martin et al. (2020) suggest that this should be a feasible avenue of research as they could demonstrate that traffic in different cities shared some internal structures that can be learned and exploited by Deep Neural Network models with appropriate encoding.

Besides supporting studies of horizontal data integration the *Traffic4cast* Competition will grow to also challenge submissions to take advantage of vertical data integration. This will build on initiatives by companion papers by Herruzo and Larriba-Pey (2020) and Martin et al. (2020) to incorporate summarial weather information or street maps into models. Interestingly, the Traffic Map Movie concept lends itself naturally to the incorporation of additional *channels* to the map. We can therefore easily add information on local weather, pollution, or semantic information like street network properties or points of interest to the data set, with the data immediately becoming available to modern Deep Neural Networks and their efficient feature encoding. As an extra dimension we can also add information on special events in time and space. This opens up the competition to two kinds of questions: on the machine learning side we can explore and test both the efficiency of encoding and integrating additional information on different scales. For traffic forecasting we can learn about which information actually has the greatest impact on predicting and hence understanding traffic.

In this we can build on a lively community from our successful first year, which witnessed hundreds of downloads of the competition data and over 500 submissions by more than 40 teams of international researchers. With the organizational and computational infrastructure for the *Traffic4cast* competition already in place, we look forward to an early engagement in the forums to define and announce the core and side challenges of *Traffic4cast* 2020.

References

Shreya Agrawal, Luke Barrington, Carla Bromberg, John Burge, Cenk Gazen, and Jason Hickey. Machine learning for precipitation nowcasting from radar images. *arXiv preprint arXiv:1912.12132*, 2019.

Luc Anselin. What is special about spatial data? alternative perspectives on spatial data analysis. Technical Report 89-4, National Center for Geographic Information and Anal-

ysis, University of California, Santa Barbara, 1989.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57(1):289–300, 1995.

Dominik Bucher, Francesca Mangili, Francesca Cellina, Claudio Bonesana, David Jonietz, and Martin Raubal. From location tracking to personalized eco-feedback: A framework for geographic information collection, processing and visualization to promote sustainable mobility behaviors. *Travel behaviour and society*, 14:43–56, 2019.

Sungbin Choi. Traffic map prediction using unet based deep convolutional neural network. *arXiv preprint arXiv:1912.05288*, 2019.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(Jan):1–30, 2006.

Stephen Dunne and Bidisha Ghosh. Regime-based short-term multivariate traffic condition forecasting algorithm. *Journal of Transportation Engineering*, 138(4):455–466, 2011.

Alireza Ermagun and David Levinson. Spatio-temporal short-term traffic forecasting using the network weight matrix and systematic detrending. In *Compendium of papers of Transportation Research Board 97th Annual Meeting*, 2018a.

Alireza Ermagun and David Levinson. Spatiotemporal traffic forecasting: review and proposed directions. *Transport Reviews*, 38(6):786–814, 2018b.

Aleksandra Gruca, Michael K Kopp, and David P Kreil. Competitive score resolution and leaderboard rank significance – lessons learnt at NeurIPS *Traffic4cast* 2019, 2020. *in press.*

Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 922–929, 2019.

Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

Pedro Herruzo and Josep L. Larriba-Pey. Recurrent autoencoder with skip connections and exogenous variables for traffic forecasting. In Hugo Jair Escalante and Raia Hadsel, editors, *Proceedings of the NeurIPS 2019 Competition Track*, volume 123 of *Proceedings of Machine Learning Research*, pages 0–0. PMLR, 2020.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

David Jonietz, Dominik Bucher, Henry Martin, and Martin Raubal. Identifying and interpreting clusters of persons with similar mobility behaviour change processes. In *The Annual International Conference on Geographic Information Science*, pages 291–307. Springer, 2018.

Yong-Hoon Kwon and Min-Gyu Park. Predicting future frames using retrospective cycle gan. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1811–1820, 2019.

Ibai Lana, Javier Del Ser, Manuel Velez, and Eleni I. Vlahogianni. Road traffic forecasting: Recent advances and new challenges. *IEEE Intelligent Transportation Systems Magazine*, 10(2):93–109, 2018. doi: 10.1109/mits.2018.2806634.

Alex X Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, and Sergey Levine. Stochastic adversarial video prediction. *arXiv preprint arXiv:1804.01523*, 2018.

Henry Martin, Ye Hong, Dominik Bucher, Christian Rupprecht, and René Buffat. *Traffic4cast*-traffic map movie forecasting–Team MIE-Lab. *arXiv preprint arXiv:1910.13824*, 2019.

Henry Martin, Ye Hong, Dominik Bucher, Christian Rupprecht, and René Buffat. GraphResNets for short-term traffic forecasts in (almost) unknown cities. In Hugo Jair Escalante and Raia Hadsel, editors, *Proceedings of the NeurIPS 2019 Competition Track*, volume 123 of *Proceedings of Machine Learning Research*, pages 0–0. PMLR, 2020.

Silvia Nittel. A survey of geosensor networks: Advances in dynamic environmental monitoring. *Sensors*, 9(7):5664–5678, 2009.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852, 2015.

Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In *European Conference on Computer Vision*, pages 835–851. Springer, 2016.

Tianfan Xue, Jiajun Wu, Katherine Bouman, and Bill Freeman. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2016.

Wei Yu, Yichao Lu, Steve Easterbrook, and Sanja Fidler. Crevnet: Conditionally reversible video prediction. *arXiv preprint arXiv:1910.11577*, 2019.

Wei Yu, Yichao Lu, Steve Easterbrook, and Sanja Fidler. Efficient and information-preserving future frame prediction and beyond. In *ICLR (Poster)*. OpenReview.net, 2020.