Deep Learning for Spatio-Temporal Forecasting: Benchmarks, Methods, and Insights from Mobility and Weather Predictions

By

Pedro Herruzo Sanchez

Departament d'Arquitectura de Computadors
Universitat Politècnica de Catalunya

Supervisors:
Josep Lluís Larriba-Pey

A thesis submitted for the degree of

Doctor of Philosophy

Barcelona, July 2023

This thesis is submitted to the Computer Science Department, Universitat Politècnica de Catalunya in fulfilment of the requirements for the degree in Doctor of Philosophy. This thesis is entirely my own work, and except where otherwise stated, describes my own research.

Pedro Herruzo Sanchez, July 2023

Acknowledgements

Este trabajo no existiría sin las personas que me han mostrado amor y me han aceptado tal y como soy.

En primer lugar, a mi abuela Guadalupe Balsera Peñas. Ella siempre me insistía en que tenía que trabajar duro y ayudar a mi familia. Ella cumplió con ese ejemplo hasta el día que se fue. En segundo lugar, a mi padre, Francisco Herruzo Balsera. No sé por qué se fue tan pronto, pero ahí lo siento siempre conmigo, gracias papá. A mi madre, María Cruz Sánchez Bernabé, que me ha enseñado con su ejemplo lo que significa amar incondicionalmente. Y a mi hermano, Aitor Herruzo Sánchez, que me sigue orgulleciendo, todos los días. Hemos pasado momentos muy difíciles, pero ahora con el tiempo, está claro que los hemos superado y hemos salido más fuertes. Te quiero tete, os quiero familia.

A la meva dona, Laia Llorens Pol, t'estimo. Gràcies per sempre estar allà. Gràcies per mostrar-me tot el que em falta i gràcies per aguantar el que em sobra. El meu ésser es sent complet quan et miro i respiro. Gràcies als teus pares i a la teva família. Per acceptar-me, voler-me com un més i per ser el bon exemple de família que sou.

Gràcies al meu director de tesi, Larri. Per la teva saviesa, per el teu esforç i el teu lideratge. A més, agraeixo profundament la teva confiança i la teva capacitat per donar-me la llibertat d'explorar independentment i sempre estar allà per guiar i donar suport quan era necessari. El teu equilibri entre autonomia i assessorament ha estat un factor clau en el meu creixement personal i acadèmic.

I would also like to express my gratitude to the Institute of Advanced Research of Artificial Intelligence (IARAI) for providing me with a fellowship and the opportunity to conduct research at their facility. This environment allowed me to interact with some of the most brilliant researchers I've ever known. I owe a special thanks to David for enlightening me on the significance of science from your perspective, and to

Aleksandra for your unwavering coordination throughout all of our collaborative projects.

También quiero agradecer al grupo de investigadores de la Agencia Española de Meteorología (AEMet). Gracias por desde el primer momento ser tan profesionales y mostrar tanta experiencia en vuestro campo. Sin vuestra ayuda no me hubiese atrevido a tratar tal semejante complicado tema. Gracias a vosotros, Pilar, Xavier y Llorenç.

Finalment, volia agrair al grup automobilístic SEAT, S.A., pel seu programa de doctorats industrials. Gràcies per col·laborar perquè la investigació continuï desenvolupant-se en aquest país i gràcies pels contactes que permet fer el pertànyer a una organització com aquesta.

Pedro Herruzo agraeix el suport de la Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya, sota la beca de Doctorat Industrial 2017 DI 52.

Abstract

This thesis explores the intersection of deep learning and spatio-temporal forecasting, focusing on the challenges and opportunities present in applying machine learning methods to predict complex geospatial and temporal phenomena. Specifically, we focus on two critical domains: weather prediction and traffic forecasting.

Firstly, we delve into the nuances of encoding traffic data optimally for deep learning models, recognizing the potential of these methods to revolutionize mobility patterns, city planning, and freight delivery services. Our work aims to provide a clear pathway to effectively harness GPS data and utilize deep learning models for accurate traffic prediction, thereby influencing real-world decision-making significantly.

Next, we turn our attention to weather forecasting. Given the substantial impact of weather on human activities and the environment, our aim is to establish best practices for encoding weather data for deep learning applications. We explore various machine learning models, assessing their performance, and identifying the most efficient and accurate approach for weather prediction.

Throughout the thesis, we emphasize the urgent need for a robust benchmark in the field of spatio-temporal forecasting, to enable systematic comparison of methods and accelerate research advancements. We discuss the essential components of such a benchmark, including open data of free access, specific tasks, relevant metrics, viable baselines, and comprehensive evaluation methodologies.

To further illustrate the practical application of these principles, we have contributed to the scientific community by publishing two novel benchmarks. The first is a high-resolution, multimodal weather forecasting benchmark, derived from satellite data, which provides comprehensive insights into the complexities of meteorological prediction.

The second is a ground-breaking high-resolution precipitation benchmark, which innovatively utilizes satellite to radar data at the surface level. This latter benchmark promotes a deeper understanding of rainfall patterns and their potential implications.

Our exploration culminates in the organization of the Weather4cast competition at the competitive venues NeurIPS, IEEE Big Data and CIKM. This sets unprecedental benchmarks for spatio-temporal prediction in weather domains, promoting innovative solutions in this intricate field. By bridging the gap between deep learning and spatio-temporal forecasting, this thesis makes a significant contribution to both machine learning methodologies and the accuracy of weather and traffic predictions. The findings promise to inspire further advancements in the application of deep learning to complex spatio-temporal processes.

Contents

List of Figures

Chapter 1

Introduction

The rapid advancement of computational capabilities and the ever-increasing availability of data have dramatically transformed our world, revolutionizing numerous fields of study. One area of particular interest is the field of spatio-temporal forecasting, which holds enormous potential for various applications, from urban planning and transportation logistics to environmental studies and disaster management.

Despite this potential, the application of state-of-the-art deep learning methods to spatio-temporal processes is still in its early stages. A significant challenge is the absence of robust and standardized benchmarks that facilitate methodological comparisons and foster accelerated research advancements. Furthermore, our understanding of how to effectively leverage deep learning for spatio-temporal processes remains limited, especially when it comes to encoding and predicting complex phenomena like traffic and weather patterns.

This thesis is motivated by the need to bridge these gaps. We aim to establish the foundation for effectively employing deep learning methods to tackle spatio-temporal problems. Furthermore, we seek to foster a community-wide benchmarking culture by providing datasets and methodologies that researchers can use to compare and improve their methods systematically. This endeavor is not only critical to advancing our scientific understanding but also crucial in harnessing the transformative potential of spatio-temporal forecasting for societal benefit.

In the following sections, we pose a series of research questions that guide our exploration and offer a detailed account of the work undertaken to answer them. The journey that ensues provides a comprehensive study on the application of deep learning in spatio-temporal forecasting, highlighting the successes, challenges, and future potential of this exciting field.

1.1 Research Questions

In this era, computation and data have become fundamental resources, propelling deep learning methods to state-of-the-art status in nearly all tasks they approach. However, in the context of spatio-temporal processes, standard benchmarks for the at-scale application of deep learning are still lacking. Hence, the central research question that frames the scope of this thesis is: **How can deep learning be effectively applied for spatio-temporal forecasting in diverse domains, such as weather and traffic prediction?**

Recognizing the breadth of this inquiry, we dissect it into more focused, specific research questions:

The interconnectedness of geographical and temporal processes significantly influences human activities, particularly in mobility. GPS-enabled devices collect and monitor these mobility patterns, forming a rich dataset for traffic prediction. Yet, the question remains: **Q1: How can traffic data be optimally encoded to exploit the capabilities of deep learning algorithms for forecasting?**

While data availability and appropriate formatting are crucial, they alone are insufficient for real-world decision-making. Stakeholders —ranging from city planners to freight delivery services and citizens— rely heavily on estimated time of arrival (ETA) provided by GPS systems. Consequently, this motivates our second research question: **Q2: How can deep learning be leveraged to accurately predict traffic patterns and conditions?**

Weather, as a geospatial and temporal process, also exerts a considerable influence on human mobility, city planning, and risk prevention. Recognizing its criticality, we ask: **Q3: What are the best practices for encoding weather data for deep learning applications, and which models yield superior performance in weather forecasting?**

To galvanize the advancement of deep learning in spatio-temporal processes —akin to progress in other fields— more than just data and models are required. Establishing a common benchmark for systematic comparison of various methods is essential. This inspires our final research question: **Q4: What are the fundamental components —including tasks, metrics, baselines, and evaluation methodologies— necessary to assess the generalization and adaptability of deep learning models in the context of spatio-temporal forecasting?**

Through these research questions, this thesis aims to explore and highlight the

potential of deep learning in dealing with complex spatio-temporal processes, contributing towards the advancement of weather and traffic forecasting.

1.2 List of Contributions

The investigation articulated in this dissertation was guided by the research questions delineated in the previous section. The exploration of these inquiries led to a series of impactful contributions to the field of spatio-temporal deep learning, culminating in various peer-reviewed publications that underpin the body of this dissertation. The contributions are organized and summarized as follows:

Part I: Deep Learning for Traffic Forecasting

[C1] We devised a recurrent autoencoder model for traffic prediction that uniquely employs multimodal input data, specifically integrating traffic and weather inputs. This model utilizes a bifurcated loss function, operating in both the input and an embedding space. This novel approach marked a significant advancement in traffic prediction methodologies [Herruzo and Larriba-Pey, NeurIPS 2019 competitions].

Part II: Open-Data Benchmarks: Multimodality, Adaptation, and Generalization in Deep Learning

[C2] We pioneered a comprehensive open-data benchmark for weather forecasting. This benchmark, which encompasses the development of a dataset, the definition of tasks, the determination of metrics, and the construction of baseline models, demonstrates that models with superior performance in training locations also show enhanced generalizability to novel geospatial locations. This underscores the benchmark's potential for promoting transfer learning in the realm of weather forecasting [Herruzo et al., IEEE Big Data 2021].

[C3] Extending from this foundation, we developed an advanced multimodal weather benchmark, also open-data, designed to assess model performance under spatio-temporal shifts and necessitating the application of super-resolution techniques. This benchmark sets a high bar for the development of models capable of handling real-world variability and refining the precision of weather predictions [SUBMITTED, NeurIPS 2022 competitions].

These groundbreaking contributions not only serve as a catalyst for further advancements in spatio-temporal deep learning, but also underscore the importance of open-data practices in fostering collaborative and inclusive scientific progress.

In the reach of this dissertation, we also contributed to other articles in the topic of traffic forecasting and benchmarks for deep learning:

[C4] We show that traffic can be tackled as a movie completion task and compare different models against the baseline [Kreil et al., NeurIPS 2019 competitions]

[C5] We show that an accurate traffic prediction horizon can span 60 minutes and U-Net models are a good foundational model [Kopp et al., NeurIPS 2020 competitions]

[C6] We constructed a benchmark for traffic forecasting focused on model robustness and generalizability across multiple new cities and pre/post covid mobility dynamics. [Eichenberger et al., NeurIPS 2021 competitions]

Also in the reach of this dissertation, and previous to the aforementioned contributions, we first explored multi-task models and the creation of benchmarks in a different topic:

[C7] We proposed a multi-task model that can summarize a set of images into different categories. We created the dataset, the labeling tool, defined the metrics, and explored explainability methods. [Herruzo et al., BMVC Workshop 2019]

Other research contributions done while pursuing this thesis

[C8] Advances on the Spanish Meteorological Agency (AEMet) [Agudo et al., XIX Congreso de la Asociación Española de Teledetección]

[C9] Co-organized *The Landslide4Sense Competition* Ghamisi et al. (2022)

[C10] Co-organized the *First Workshop on Complex Data Challenges in Earth Observation* [Gruca et al., CIKM 2021]

1.3 Dissertation Outline

The remainder of this thesis is structured as follows:

Chapter 3 addresses research questions Q1 and Q2. It elucidates the methods for encoding traffic data and creating models for predicting traffic volumes and speeds. This chapter specifically emphasizes the consideration of multimodal data, such as weather and time of the day, to enhance predictive capabilities.

Chapter 4 shifts focus to weather forecasting as a standalone task, addressing research questions Q3 and Q4. In this chapter, we detail the process of establishing dedicated methods and benchmarks for multimodal weather forecasting.

Chapter 5 further deepens the investigation into research questions Q3 and Q4. In this chapter, we introduce the first high-resolution satellite to radar benchmark and share insights gleaned from the development of these models.

Both chapters 4 and 5 also provide detailed information about the competitions we organized as part of this thesis. The competitions validated the utility and usability of the datasets, tasks, models, and metrics developed within this thesis. Furthermore, they provided valuable insights from the research community and demonstrated the effective application of our work.

This comprehensive approach ensures that our research findings contribute substantially to the field of spatio-temporal forecasting, paving the way for further advancements in this domain.

Chapter 2

Background

This chapter provides an essential background and introduces the key concepts that will be foundational to the research presented in this thesis. As we embark on an exploration of the interplay between deep learning and spatio-temporal forecasting, particularly in the context of traffic and weather prediction, it is important to establish a common understanding of the terminology, methodologies, and frameworks that underpin this field of study.

The chapter is structured to provide a comprehensive overview of the fundamental concepts, ranging from the principles of deep learning and forecasting techniques to the evaluation of model performance using benchmarks, tasks, datasets, and metrics. Additionally, we delve into advanced topics such as multitask models, multimodal data integration, and transfer learning, which play a crucial role in enhancing the predictive capabilities of our models. By elucidating these concepts, this chapter aims to equip readers with the necessary background knowledge to fully appreciate the contributions and findings of the thesis.

2.1 Spatio-temporal processes

Spatio-temporal processes are a fundamental concept in various scientific disciplines, including geography, ecology, meteorology, and epidemiology. These processes describe phenomena that evolve over both space and time, capturing the dynamic interplay between spatial patterns and temporal changes. Understanding spatio-temporal processes is essential for modeling and predicting the behavior of complex systems, such as the spread of diseases, the movement of wildlife populations, urban mobility, and the impact of climate change on natural resources.

The importance of studying spatio-temporal processes can be illustrated through several real-world examples. For instance, in epidemiology, the spread of infec-

tious diseases such as COVID-19 is a spatio-temporal process, as the transmission of the virus occurs in specific geographic locations and evolves over time. Similarly, in ecology, the migration patterns of bird species are spatio-temporal processes, as they involve the movement of birds across different regions during specific seasons. In meteorology, weather patterns such as hurricanes and tornadoes are also spatio-temporal processes, as they develop and move across geographic areas over time. Understanding these processes is crucial for developing effective interventions, conservation strategies, and disaster response plans.

Definition: A spatio-temporal process is a stochastic process that describes the evolution of a variable or set of variables over both space and time. It is characterized by a **spatial domain**, a **temporal domain**, and a **set of values or states associated with each point in the spatio-temporal domain**. The process captures the dependencies and interactions between spatial and temporal components, allowing for the analysis of patterns, trends, and dynamics in the data.

To illustrate this concept, let's consider the example of temperature variation across Barcelona city over the course of a year. In this case, the temperature variation can be described as a spatio-temporal process, where the variable of interest is temperature, and its values change over both space and time.

In this example, the spatial domain is the city of Barcelona, which can be represented as a subset of a two-dimensional Euclidean space. The spatial domain includes all the locations within the city where temperature measurements are taken, such as different neighborhoods, parks, and landmarks. The granularity of these locations is known as the **spatial resolution**. A higher spatial resolution means more detailed or finer spatial sampling (e.g., measurements taken at every street corner), while a lower spatial resolution means coarser or broader spatial sampling (e.g., measurements taken at the center of each city). The choice of spatial resolution depends on the scale of analysis and the research objectives.

The temporal domain in this example is a one-year interval, starting on January 1st and ending on December 31st. The temporal domain includes all the time points at which temperature measurements are recorded, such as hourly, daily, or monthly intervals. The frequency or granularity of observations over time is known as **temporal resolution**. A higher temporal resolution means more frequent or detailed observations (e.g., samples every minute), while a lower temporal resolution means less frequent or coarser observations (e.g., samples every day).

Let us also explicitly enumerate the dependencies and interactions in this particular spatio-temporal process. Values at different locations in Barcelona city may

be spatially correlated, meaning that nearby locations are likely to have similar temperatures. Similarly, the temperature values at different times may be temporally correlated, meaning that the temperature at a given location is likely to be similar from one hour to the next. The spatio-temporal process allows us to analyze and model these correlations to understand the patterns and dynamics of temperature variation across the city over time.

We finish this section with a mathematical formalization. A spatio-temporal process can be represented as a function $Z(s,t)$, where s denotes a point in the spatial domain S, t denotes a point in the temporal domain T, and $Z(s,t)$ represents the value or state of the process at location s and time t. The spatial domain S is typically a subset of a Euclidean space (e.g., a two-dimensional geographic area), and the temporal domain T is usually an interval of the real line (e.g., a time period). The function $Z(s,t)$ is often modeled as a random field, with specific statistical properties and dependencies that capture the spatio-temporal structure of the process.

2.1.1 Traffic Forecasting

Traffic forecasting is the process of predicting future traffic conditions, such as traffic flow, speed, and density, on transportation networks. It is a critical component of transportation planning, traffic management, and intelligent transportation systems (ITS). Traffic forecasting models leverage historical and real-time data to estimate traffic states at future time intervals and specific locations on the road network. These predictions are essential for mitigating congestion, optimizing traffic signal timings, and informing travelers about expected travel times.

The key terms and definitions in traffic forecasting that are used in this thesis are:

Traffic Flow: The number of vehicles passing a given point on a roadway per unit of time, typically measured in vehicles per hour (vph).

Traffic Speed: The rate at which vehicles are moving on a roadway, typically measured in kilometers per hour (km/h) or miles per hour (mph).

Traffic Density: The number of vehicles occupying a unit length of roadway, typically measured in vehicles per kilometer (veh/km) or vehicles per mile (veh/mi).

Origin-Destination (OD) Matrix: A matrix that quantifies the demand for travel between different origin and destination pairs in a transportation network.

Traffic Simulation: A computational model that mimics the behavior of traffic on a road network to analyze and predict traffic conditions.

2.1.2 Weather Forecasting

Weather forecasting is the application of scientific principles and techniques to predict atmospheric conditions at a specific location and time. It involves the use of numerical weather prediction (NWP) models, satellite and radar observations, and meteorological data to generate forecasts of temperature, precipitation, wind speed, and other atmospheric variables. Weather forecasts are vital for public safety, agriculture, aviation, and various other sectors that are sensitive to weather conditions.

Key terms and definitions in weather forecasting include:

Numerical Weather Prediction (NWP): The use of mathematical models to simulate the behavior of the atmosphere and predict future weather conditions based on initial observations.

Precipitation: Any form of water, such as rain, snow, sleet, or hail, that falls from the atmosphere to the Earth's surface.

Temperature: Temperature is a measure of the average kinetic energy of the molecules in a substance or system. In the context of weather forecasting, temperature refers to the air temperature, which is a key atmospheric variable that affects various weather phenomena. Air temperature is typically measured in degrees Celsius (°C) or degrees Fahrenheit (°F) and is a fundamental component of weather forecasts. Accurate temperature predictions are essential for agriculture, energy consumption planning, and public health.

Atmospheric Pressure: The force per unit area exerted by the weight of the atmosphere, typically measured in hectopascals (hPa) or millibars (mb).

Spectral Bands of Satellites: College (2020) Spectral bands refer to specific wavelength ranges of the electromagnetic spectrum that are captured by satellite sensors. Different spectral bands are sensitive to different features of the Earth's surface and atmosphere.

Visual or visible bands refers to the portion of the electromagnetic spectrum that is detectable by the human eye. It includes wavelengths ranging from approximately 380 nanometers (nm) to 740 nm. In this range, light is perceived as different colors, with violet at the shorter wavelengths, red at the longer wavelengths, and other colors (blue, green, yellow, and orange) in between. Visible band imagery from satellites is used to observe features such as clouds, land surfaces, bodies of

water, and vegetation. Visible imagery is similar to what we see with our eyes and is often used in true-color satellite images.

Infrared bands refer to the portion of the electromagnetic spectrum with wavelengths longer than visible light but shorter than microwaves. Infrared wavelengths range from approximately 700 nm to 1 millimeter (mm). Infrared radiation is primarily associated with thermal energy or heat emitted by objects. Infrared bands are further divided into categories, including near-infrared (NIR), shortwave infrared (SWIR), mid-infrared (MIR), and thermal infrared (TIR). In weather satellite imagery, thermal infrared bands are commonly used to measure the temperature of clouds, land surfaces, and bodies of water. Infrared imagery is also used to detect cloud heights, atmospheric moisture, and sea surface temperatures.

Microwave bands refer to the portion of the electromagnetic spectrum with wavelengths longer than infrared radiation but shorter than radio waves. Microwave wavelengths range from approximately 1 mm to 1 meter. Microwaves can penetrate clouds, smoke, and precipitation, making them valuable for observing the Earth's surface and atmosphere under various conditions. Passive microwave sensors on satellites measure the natural microwave emissions from the Earth's surface and atmosphere, while active microwave sensors (such as radar) transmit microwave pulses and measure the returned signals. Microwave imagery is used for measuring precipitation, soil moisture, snow cover, sea ice, and ocean surface wind speed.

Geostationary satellites are a type of Earth-orbiting satellite that maintains a fixed position relative to the Earth's surface. These satellites are positioned in a geostationary orbit, which is a circular orbit located approximately 35,786 kilometers (22,236 miles) above the Earth's equator. At this altitude, the satellite's orbital period matches the Earth's rotation period, allowing the satellite to remain stationary with respect to a specific geographic location on the Earth's surface. As a result, geostationary satellites provide continuous monitoring of the same region, making them ideal for weather observation, telecommunications, and broadcasting. Geostationary weather satellites are used to monitor large-scale weather patterns, track the development of storms, and provide real-time imagery of atmospheric conditions over a specific geographic area.

Polar-orbiting satellites are a type of Earth-orbiting satellite that travels in a near-polar orbit, passing close to both the North and South Poles during each orbit. These satellites typically operate at much lower altitudes than geostationary satellites, with typical orbital altitudes ranging from 700 to 800 kilometers (435 to

497 miles) above the Earth's surface. Due to their lower altitude and polar orbit, polar-orbiting satellites provide global coverage and higher spatial resolution than geostationary satellites. Each orbit takes approximately 90 to 100 minutes, and the satellite passes over different regions of the Earth with each orbit, allowing for comprehensive global observations. Polar-orbiting weather satellites are used for a wide range of applications, including monitoring atmospheric temperature and moisture profiles, measuring sea surface temperatures, observing ice and snow cover, and tracking global weather patterns.

2.2 Benchmarks

In the context of scientific research and particularly in machine learning and deep learning, a benchmark is a standard or point of reference against which things may be compared or assessed. A benchmark often consists of a dataset and a set of tasks or problems, along with evaluation metrics and sometimes baseline results or models. Benchmarks are used to evaluate and compare the performance of different algorithms, models, or systems under the same conditions, providing a fair and objective measure of their capabilities.

In the field of deep learning, benchmarks have played a crucial role in driving progress by providing researchers with common platforms for comparison and competition. Here are a few examples of well-known benchmarks:

- ImageNet Deng et al. (2009) is a benchmark in the field of computer vision that consists of a large dataset of annotated images and a competition (the ImageNet Large Scale Visual Recognition Challenge, or ILSVRC). ImageNet has been successful because it provides a large-scale, diverse, and challenging dataset for image classification and other tasks. The annual competition has spurred many advances in deep learning, including the development of convolutional neural networks (CNNs) like AlexNet Krizhevsky et al. (2012), VGG Simonyan and Zisserman (2015), and ResNet He et al. (2015).

- Common Objects in Context (COCO) Lin et al. (2015) is another benchmark in computer vision that provides a dataset and tasks for object detection, segmentation, and captioning. COCO has been successful due to its focus on the detection of objects in complex scenes with context, which is a more realistic and challenging problem than classifying isolated images.

- GLUE (General Language Understanding Evaluation) and SuperGLUE Wang et al. (2019) are benchmarks for natural language processing (NLP) tasks. They provide multiple datasets and tasks to evaluate the generalization ability of NLP models across different types of linguistic understanding tasks. The success of GLUE and SuperGLUE comes from their comprehensive and diverse set of tasks, which has driven the development of powerful language models like BERT Devlin et al. (2019), GPT Radford et al. (2018) (precursor of the well-known ChatGPT), and RoBERTa Liu et al. (2019).

These benchmarks have been successful because they provide **large, diverse, and challenging datasets**; **clear and meaningful tasks**; and **objective evaluation metrics**. They also **foster competition and collaboration** in the research community, which **drives innovation and progress**.

2.2.1 Tasks

In machine learning, a "task" refers to a specific problem or type of problem that a machine learning system is designed to solve. This could include classification, regression, clustering, anomaly detection, reinforcement learning, and others. Each task is defined by a particular set of inputs and outputs, and the goal of the machine learning system is to learn a function that maps the inputs to the outputs based on patterns in the training data.

In this thesis, we focus on tasks that involve using historical and real-time data to make predictions about future conditions. These tasks are typically approached as regression problems (predicting a continuous value, like traffic speed or temperature) or classification problems (predicting a discrete value, like the occurrence of a traffic incident or a severe weather event).

2.2.2 Datasets

A dataset is a collection of structured data that is used for analysis or to train, validate, and test machine learning models. Each entry in the dataset, often called an instance or example, typically consists of a set of features (or inputs) and, in supervised learning tasks, a corresponding label (or output). The features represent the characteristics of the data, while the labels represent the outcome or target variable that the machine learning model aims to predict.

2.2.3 Evaluation Metrics

Evaluation metrics are measures used to quantify the performance or quality of a machine learning model's predictions. These metrics provide an objective way to assess the model's accuracy, precision, recall, F1 score, area under the ROC curve (AUC-ROC), mean squared error (MSE), log loss, and many others. The choice of evaluation metric depends on the specific task, the nature of the data, and the business or research objectives. Evaluation metrics are essential for model selection, tuning, and validation.

For **traffic speed** and **volume prediction**, which are typically regression tasks, the following evaluation metrics are commonly used:

- **Mean Absolute Error** (MAE): Measures the average magnitude of the errors in a set of predictions, without considering their direction. It's the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.
- **Root Mean Square Error** (RMSE): It is a quadratic scoring rule that also measures the average magnitude of the error. It's the square root of the average of squared differences between prediction and actual observation. The RMSE gives a relatively high weight to large errors.
- **Mean Absolute Percentage Error** (MAPE): Measures the size of the error in percentage terms. It is calculated as the average of the unsigned percentage error.
- **R-squared** (Coefficient of Determination): It is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

For **weather forecasting** the above metrics are also found in the literature together with the following additional metrics:

- **Critical Success Index** (CSI): The CSI measures the correctly predicted events relative to the total number of events that were either forecasted or observed. It's calculated as the number of true positives divided by the sum of true positives, false positives, and false negatives. The CSI ranges from 0 to 1, with 1 indicating a perfect forecast. This metric is particularly useful for rare events, as it does not consider true negatives (i.e., correctly predicting the non-occurrence of an event) in its calculation.

- **Probability of Detection** (POD) or hit rate, the POD is the proportion of actual positive events (e.g., instances of precipitation above a certain threshold) that were correctly identified by the model. It's calculated as the number of true positives divided by the sum of true positives and false negatives. A POD of 1 indicates that all actual positive events were correctly identified, while a POD of 0 indicates that no actual positive events were correctly identified.

- **False Alarm Ratio** (FAR) is the proportion of predicted positive events that were actually negative (e.g., instances where the model predicted precipitation above a certain threshold, but it did not occur). It's calculated as the number of false positives divided by the sum of true positives and false positives. A FAR of 0 indicates that all predicted positive events were correct, while a FAR of 1 indicates that all predicted positive events were incorrect.

2.2.4 Baseline

A baseline refers to a method, model, or metric score that serves as a comparison point for other methods, models, or experiments.

In machine learning, a baseline model is often a simple or well-established model that is easy to implement and understand. For example, in a classification task, a baseline model might predict the most common class for all instances, while in a regression task, a baseline model might predict the mean or median outcome for all instances.

The performance of the baseline model, as measured by appropriate evaluation metrics, provides a minimum threshold that more complex or novel models should aim to exceed. If a complex model does not perform significantly better than the baseline, it may not be worth the additional complexity and computational cost.

2.2.5 Benchmarks Presented in this Thesis

This thesis introduces comprehensive benchmarks constituted by data related to traffic and weather dynamics.

These sets of data, intrinsically rich in multi-variate, spatio-temporal characteristics, are transformed into a **sequence of multi-channel images**. For instance, in the context of traffic, each pixel corresponds to the volume of vehicles and their average speeds within a defined area. In the weather dataset, a pixel might encapsulate variables like temperature, precipitation, cloud coverage, among others.

Detailed descriptions of these open datasets, along with specific tasks, evaluation methods, and deep learning baseline models will be thoroughly discussed in the corresponding chapters.

2.3 Deep Learning

Deep learning, a subfield of machine learning, focuses on algorithms inspired by the structure and function of the brain called artificial neural networks. They contain multiple layers (hence "deep") to model and understand complex patterns in data. In this section, we focus on three specific types of neural networks that are critical to our research: Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Autoencoders. Furthermore, we shed light on essential practices such as batch sampling and one-hot encoding, which are also concepts required to know for our research.

2.3.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a class of deep learning models primarily used for processing grid-like data such as images. CNNs utilize convolutional layers, where each neuron is connected to a small, local region of the input, and all neurons in the layer share weights. This weight-sharing scheme helps CNNs process high-dimensional data efficiently, making them highly effective for tasks like image and video recognition Lecun et al. (1998).

2.3.2 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are designed for processing sequential data. Unlike feedforward neural networks, RNNs maintain hidden states that capture historical information about the sequence. At each time step, an RNN takes the current input and the previous hidden state as input and updates the hidden state. This allows RNNs to exhibit dynamic temporal behavior, making them suitable for tasks such as language modeling, speech recognition, and time series prediction Hochreiter and Schmidhuber (1997a).

2.3.3 Autoencoders

Autoencoders are a type of neural network used for learning efficient codings of input data. They consist of two main components: an encoder that maps the input

data to a lower-dimensional representation, and a decoder that attempts to reconstruct the original data from this representation. The learning process minimizes the difference (or distance) between the original input and the reconstructed output, making autoencoders useful for tasks like dimensionality reduction, anomaly detection, and learning generative models of data Ballard (1987), Goodfellow et al. (2016).

2.3.4 Batch Sampling and One-hot Encoding

In the training of deep learning models, it is common practice to use batch sampling. This involves dividing the dataset into several subsets, or batches, and updating the model's parameters based on one batch at a time. This process makes training more manageable and efficient, especially for large datasets. Moreover, shuffling the batches ensures that the model is not biased by the order of the data, which can lead to more robust performance Bengio (2012), Goodfellow et al. (2016).

One-hot encoding is another essential technique in data preprocessing. It is a method of converting categorical data into a format that can be provided to machine learning algorithms to improve prediction. This technique transforms each category value into a new column and assigns a binary value of 1 or 0. Each integer value is represented as a binary vector, making this method suitable for dealing with discrete categorical values where no ordinal relationship exists Goodfellow et al. (2016).

By understanding these deep learning concepts, we can effectively design and train neural network models for our specific forecasting tasks.

2.4 Multitask Models

Multitask Learning (MTL) is a subfield of machine learning where multiple learning tasks are solved at the same time while exploiting commonalities and differences across tasks. This is in contrast to traditional machine learning approaches that treat each task independently. The key idea is that by learning tasks in parallel, the model can leverage the information contained in multiple related tasks to improve generalization. This is particularly beneficial when the tasks are related in some way, as the learning for one task can inform the learning for the others. MTL can lead to improved learning efficiency and prediction accuracy for each task, especially when the amount of data for each task is limited.

As an example, consider a city with a network of roads, and suppose we have historical data on both traffic speed and traffic volume for each road segment at different times of day. We could train a multitask learning model to predict both traffic speed and traffic volume for each road segment at a future time, such as the next hour.

In this case, the model would have two tasks: speed prediction and volume prediction. The model would be trained on both tasks simultaneously, with a shared representation learning component (such as a shared neural network layer) that captures common features relevant to both tasks, and task-specific components (such as separate output layers) that capture features unique to each task.

The idea is that by learning to predict both speed and volume together, the model can leverage the correlation between these two traffic parameters to improve its predictions. For example, the model might learn that when traffic volume is high, traffic speed tends to be low, and vice versa. This shared knowledge can help the model make more accurate predictions for both tasks, compared to if it were trained on each task separately.

2.5 Multimodal Data

Multimodal data refers to data that comes from different sources or formats, or data that represents different types of information. In the context of machine learning, multimodal learning involves building models that can process and relate information from multiple types of data.

For example, in a traffic forecasting scenario, multimodal data could include data from road sensors (like traffic volume and speed), weather data (like temperature and precipitation), and event data (like road accidents or sporting events). Each of these data types provides a different ”mode” of information that can contribute to the overall prediction task.

Another recent example a joint embedding is learned from six different modalities Girdhar et al. (2023) - images, text, audio, depth, thermal, and IMU data. It enables new generative cross-modal prompts like text and sound or image and IMU.

The key challenge in learning from multimodal data i**s to effectively integrate the different types of information**, which may require sophisticated model architectures and training methods. When done effectively, multimodal learning can

lead to more robust and accurate models by leveraging the complementary information available in different data modes.

2.6 Generalization and Domain Adaptation

In the field of machine learning, the ultimate goal of a model is not just to perform well on the training data, but to make accurate predictions on new, unseen data. This ability to perform well on unseen data is known as "**generalization**". Generalization is a fundamental aspect of machine learning models, and assessing a model's generalization performance is crucial for understanding its effectiveness and reliability.

Generalization is typically evaluated by training a model on a subset of the available data (the training set) and then testing the model on a different subset of the data (the test set) that the model has not seen during training. The performance of the model on the test set provides an estimate of its generalization ability.

However, in many real-world scenarios, we are interested not just in a model's ability to generalize to unseen samples from the same distribution, but also its ability to adapt to different but related distributions. This is known as "**domain adaptation**". Domain adaptation is particularly relevant in spatio-temporal forecasting tasks, where the underlying data distributions can change over time or across locations.

In this thesis, we introduce benchmarks that are designed to assess both the generalization and domain adaptation capabilities of models for spatio-temporal forecasting tasks. We train models on one spatio-temporal domain and test them on both unseen samples from the same domain (to assess generalization) and on samples from a different spatio-temporal domain (to assess domain adaptation). Through these benchmarks, we aim to provide a comprehensive evaluation of model performance in spatio-temporal forecasting tasks, taking into account both the ability to generalize from seen to unseen data and the ability to adapt to new spatio-temporal domains.

Part I

Deep Learning for Traffic Forecasting

Chapter 3

Multimodal Traffic Autoencoder

This chapter is based on materials from the following peer-reviewed paper:

Herruzo and Larriba-Pey (2020). Recurrent Autoencoder with Skip Connections and Exogenous Variables for Traffic Forecasting. In NeurIPS 2019 Competition and Demonstration Track. PMLR.

3.1 Introduction

Mobility plays a vital role in our society, affecting various aspects of our daily lives, including transportation efficiency, urban planning, and environmental sustainability. With the increasing availability of data sources and advancements in machine learning, particularly deep learning, there is a growing opportunity to leverage these technologies to tackle mobility challenges more effectively. In this context, machine learning offers the potential to extract valuable insights and make accurate predictions from the vast amounts of mobility data generated by various sources.

Accurately forecasting traffic conditions in complex urban environments poses a significant challenge. This chapter presents a novel approach that makes core contributions to address this task by leveraging an innovative multimodal method for predicting speed, volume, and main traffic direction. Our approach harnesses the power of an aggregated representation of traffic data presented as videos, allowing for more accurate and comprehensive mobility forecasting.

The core contributions of this work lie in three key areas. First, we exploit the temporal continuity and dynamics within a sequence of frames, enabling the prediction of evolving traffic patterns in a lower-dimensional space. This approach

captures the inherent relationships between past and future traffic states, enhancing the accuracy of predictions. Second, our model incorporates multimodal data, including weather conditions, time, and seasonal information, to provide a more comprehensive understanding of the factors influencing traffic behavior. By integrating these additional variables, our approach enhances the predictive capabilities of the model, resulting in more accurate and context-aware forecasts. Finally, we introduce a novel sampling approach for sequences that ensures batch diversity while parallel optimization during the training process.

Through extensive experimentation and evaluation, our approach demonstrates significant improvements in accuracy and efficiency, providing valuable insights for effective mobility planning and management. By leveraging the power of aggregated traffic data presented as videos and incorporating multimodal information, our method offers a promising solution for accurate and comprehensive mobility forecasting in complex urban environments.

3.2 Related Work

Traffic forecasting, a crucial component of efficient transportation management, endeavors to predict future traffic flow on various infrastructures, including roads, bridges, railway lines, and airports Wikipedia contributors (2023). The formulation of these forecasts necessitates the amalgamation of myriad data sources, encompassing current traffic patterns, the physical characteristics of the infrastructure, and demographic information. A diverse array of devices, including loop detectors, Bluetooth Mac scanners, mobile phones, and connected cars, are instrumental in the acquisition of such data Respati et al. (2018).

This wealth of information is harnessed to build models capable of generating predictions that illuminate the trajectory of future traffic flows. The surge in accessibility to dynamic and big data, facilitated by modern technologies, has unlocked new avenues for augmenting the precision and predictability of these traffic estimations.

However, traffic forecasting is not devoid of challenges. The task has to navigate substantial complexities, primarily resulting from the high dimensionality of the large data volumes and the range of dynamics at play, such as unforeseen incidents like traffic accidents Jiang and Luo (2022). Furthermore, the traffic state at a particular location presents both spatial and temporal dependencies. Traditional

linear time series models, including the Auto-Regressive Integrated Moving Average (ARIMA), encounter limitations in addressing such spatiotemporal forecasting conundrums. This scenario has prompted the integration of machine learning and deep learning methodologies, which offer significant enhancements in forecasting accuracy.

Deep learning has emerged as a powerful tool for tackling traffic forecasting problems, offering sophisticated mechanisms to model spatial and temporal dependencies within traffic data.

Specific deep learning architectures, such as Convolutional Neural Networks (CNNs) LeCun et al. (1989) and Recurrent Neural Networks (RNNs) or Long Short Term Memory (LSTM) Hochreiter and Schmidhuber (1997b), have seen extensive use in this context for spatial and temporal domains respectively. CNNs, with their spatial invariance and ability to extract local patterns, have proven effective in learning the spatial dependencies that exist within traffic data. On the other hand, RNNs and, in particular, LSTMs, with their memory cells and gate mechanisms, are adept at modeling the temporal dependencies and long-term patterns that are inherent in time-series data such as traffic flow. Together, these models have significantly enhanced the state-of-the-art in traffic forecasting, with improvements in both accuracy and robustness Du et al. (2019), Chen et al. (2019).

Traffic forecasting involves the evaluation of historical traffic data, alongside optional external factors such as weather and holidays, to predict future traffic states. Thus, requiring multimodal data to be included in traffic models.

In the Time and Weather Aware Deep Neural Network (TW-DNN) model Ryu et al. (2020), the authors use a multi-module DNN framework to generate reliable long-range traffic data. It extracts features from various inputs such as weather forecast data, time, road network information, and traffic speed and flow data. The performance of this model has been tested in different traffic situations, including rush hours, holidays, and heavy rains, increasing the accuracy of the systems. However, the model predicts speed and traffic flow on specific links (road segments) in highways, as opposed to urban traffic where interaction across multiple streets happens.

A seminal piece of research that has significantly influenced our work is "Hierarchical Long-term Video Prediction without Supervision" [Wichers et al. 2018]. In this study, an encoder-decoder structure is employed for predicting future video frames. The process involves predicting within an embedded space and subsequently decoding to regain the original dimensionality. This dual-fold approach

aims at minimizing both the subsequent frame output and the prediction within the embedded space. The strategies and methodologies delineated in this work have provided considerable inspiration for our own approach to the novel image-like encoding of traffic forecasting.

For the development and evaluation of traffic forecasting models, the availability of robust traffic datasets is paramount. The Traffic Flow Forecasting Dataset, curated by the University of California Zhao et al. (2019), serves as a notable example. This resource aids in forecasting spatio-temporal traffic volume, basing predictions on the historical traffic volume and associated features of neighboring locales. The dataset is rich in features, encapsulating 47 distinct variables such as historical traffic volume sequences, day of the week, hour of the day, road direction, the number of lanes, and the road's name. The designated task within this dataset involves forecasting the traffic volume across all sensor locations 15 minutes into the future.

Traffic forecasting research typically focuses on individual road segments or clusters of adjacent segments within a highway. However, recent methodologies have emerged that process traffic data in a format compatible with deep learning models originally designed for image processing. One such approach employs Convolutional Neural Networks (CNNs) to model an entire city as a grid Jiang and Zhang (2019). This representation facilitates traffic forecasting by efficiently managing the high dimensionality inherent in the data and addressing spatial dependencies. These attributes have proven the efficacy of CNNs in the context of traffic forecasting, paving the way for more sophisticated, image-inspired forecasting models.

3.3 The Traffic4cast Formulation

The Traffic4cast challenge frames the problem of traffic forecasting as a scene completion task over the course of a year in three distinct cities: Berlin, Istanbul, and Moscow. This task involves projecting trajectories of raw GPS positions for each city onto a graphical representation of the city. This representation takes the shape of a three-channel image with dimensions corresponding to the height (495), width (436), and channels (3). The channels represent speed, volume, and heading respectively.

In the image, each pixel corresponds to a square region measuring 100m by 100m. These regions are aggregated over a time bin of 5 minutes. To represent

a full hour of data, we stack 12 of these 5-minute time bins together. Consequently, a complete day can be represented by 288 time bins, each spanning 5 minutes (24x12). Hence, the data for each city and day can be represented as a four-dimensional tensor $T^{day}_{city}[t, h, w, c]$ with shape (288, 495, 436, 3). Here the values are integers acting as indices for the tensor: $t \in [0, 287]$ (time bin), $h \in [0, 494]$ (height), $w \in [0, 435]$ (width), and $c \in [0, 2]$ (channel). As a consequence, this representation allows for this task to be framed as a video-to-video prediction.

The respective domains for the speed and volume channels are integer values in the range of 0 to 255, while the heading channel takes on one of the values in the set 0, 1, 85, 170, 255. The volume channel represents the count of vehicles in the given interval and region (t, h, w), with the count being capped at certain levels to filter out noise. These capped values are then proportionally mapped to the range [1, 255] and rounded to the nearest integer. A value of 0 in this channel signifies that no data is available for the given time bin.

The speed channel is calculated in a similar manner, with the only difference being that the aggregation method is averaging as opposed to counting. Here, a value of 0 denotes stationary vehicles, provided the volume at this location is greater than zero.

The heading channel, on the other hand, is calculated differently. Each GPS probe point records the heading direction in degrees (from 0 to 359), which is then divided into four distinct heading directions; North-East (from 0 to 90, represented as heading=85), South-East (from 90 to 180, as heading=255), South-West (from 180 to 270, as heading=170), and North-West (from 270 to 359, as heading=1). The selected value for the heading channel is the bin with the highest count of points, with a value of 0 being assigned when it's impossible to determine the maximum due to equal number of points in all directions. It's important to note that there is no data if and only if the volume is 0. Also note that even minor variations in car counts, such as a difference of one vehicle, could result in a change in the value of the heading channel. This potentially introduces noise into the data.

Figure 3.1 presents an example of conditional distributions observed within this dataset. Notably, the *heading* appears to be uniformly distributed across various *speed* or *volume* values. Conversely, the *volume* exhibits a narrower range as the *speed* increases. This pattern is not unique to this instance; similar behaviors can be observed across different days and in different cities.

In Figure 3.2, we illustrate the daily evolution of *volume*. It's interesting to note that the mean *volume* on Tuesdays in Istanbul is consistently higher than the overall

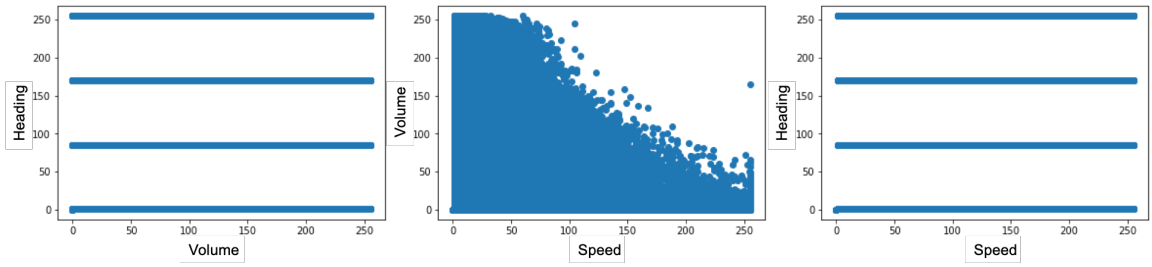

Figure 3.1: Conditional distributions for the city of Berlin on 2018.03.14 in the validation set. The figure illustrates the relationships between heading, speed, and volume. Note the uniform distribution of heading across speed and volume, and the narrowing range of volume values as speed increases.

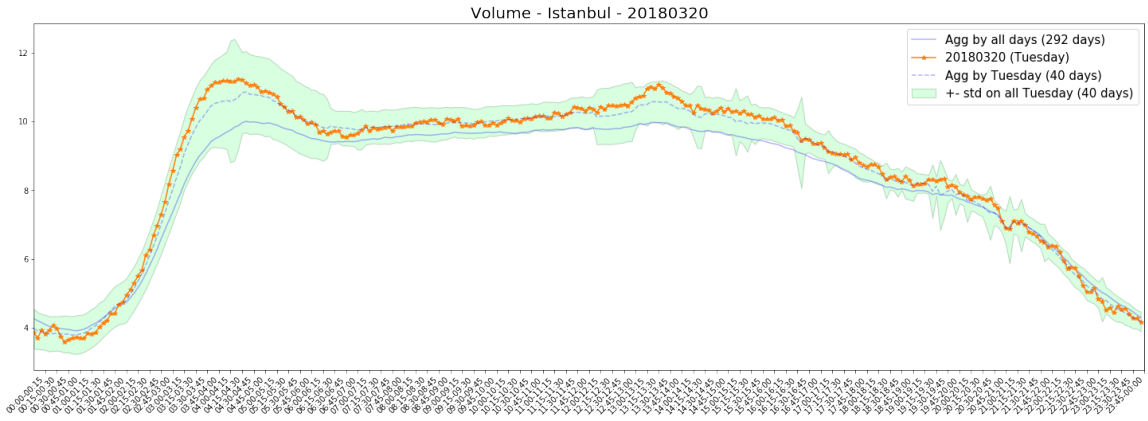

Figure 3.2: Daily evolution of *volume* in Istanbul (represented by orange asterisks), compared with the distribution for the same days of the week (blue dashed line for mean and shadow for standard deviation), and with all days (continuous blue line). The figure highlights the higher mean *volume* observed on Tuesdays as compared to the overall daily mean. The number of days used in each aggregation is indicated in brackets. Best viewed in electronic form.

daily mean.

The Traffic4cast dataset Kreil et al. (2020) provides 285 days for training (not to be confused with the 288 time bins per day) , 7 for validation and 72 for testing for each city. The training and validation sets contain information for each time bin, but the test set only has information in 5 blocks of 12 bins (1 hour of information for each block). **The challenge's goal is to predict the traffic for 5, 10, and 15 minutes ahead, given the information about the previous hour, five times per day**. The start times for these 15-minute prediction blocks differ slightly for the cities of Istanbul, Moscow, and Berlin.

The problem can thus be defined as finding a function f that minimizes the loss function L, which measures the error between the ground truth and the prediction.

$$f = \min_{\widetilde{f} \in \Theta} L\left(\widetilde{f}\left(T_{city}^{day}[s-12:s,h,w,c]\right),\ T_{city}^{day}[s:s+3,h,w,c]\right) \tag{3.1}$$

The function operates on a tensor that represents a sequence of $q = 12$ time bins as inputs $(s-12:s)$ leading up to a given time bin (s) and aims to predict the state of the next three time bins $(s:s+3)$. The parameter q could be fixed to any integer from 1 to 12 given that the provided test sets contain 1 hour in time-bins of 5 minutes. In this work we use the whole input length $q = 12$.

3.4 Sampling Strategies for Video to Video

We propose three distinct sampling strategies, all of which vary based on the desired sequence input length, represented as q, where q can take any value from 1 to 12. It is important to note that the output length remains constant at 3 frames across all strategies. As a reminder, each day corresponds of 288 bins of 5 minutes, starting from bin 0 and ending on bin 287. Figure 3.3 provides a visualization of the methodologies.

Non-overlapping: This technique involves partitioning each day into a specific number of non-repeated sequences. The number of sequences, T_q, is determined by taking the ceiling of $288/(q+3)$, where the ceiling function gives the smallest integer greater than or equal to the given value. For instance, with an input length q of 3, a day can be segmented into $T_3 = 48$ sequences each of length 6 (3 frames for training and 3 for testing).

Sliding Window: This approach employs every conceivable sequence of length $q+3$, initiating from frame 0, then frame 1, and continues until the last sequence commences at frame $288-(q+3)$. For example, with an input length q of 3, the last sequence will start at 288-(6)=282. That means that input will include times 282, 283, and 284. Then, the predicted sequence would correspond to time bins 285,286, and 287. As a comparison, the non-overlapping strategy would yield $285\ \textit{days} \times 48\ \textit{sequences / day} = 13680$ sequences to train per city. In contrast, the sliding window strategy produces $285\ \textit{days} \times (288-(3+3))\ \textit{sequences / day} = 80370$ sequences, equating to nearly 6 times more sequences to train.

Like-test: This method restricts training to sequences with output time bins that align with those in the test set. This would result in only a total of $285 \times 5 = 1425$ sequences to train.

To instill diversity of patterns across times, days of the week or even across cities, we use the following method. At the time of constructing each batch, our dataset is defined as a list of pairs, each pair consisting of a 'day' and a 'time bin'. These pairs are shuffled at the start of every epoch, leading to batches that amalgamate sequences from various days, time bins and cities. This approach is expected to expedite convergence, as batches are always unique after each epoch, a concept elucidated by Yoshua Bengio in [Bengio 2012]. Additionally, all batch preparation and preprocessing are executed parallel to the optimization process, which heightens the diversity and efficiency of our training. During preprocessing, data is converted to float numbers and all values are normalized to fall within the range $[0, 1]$.

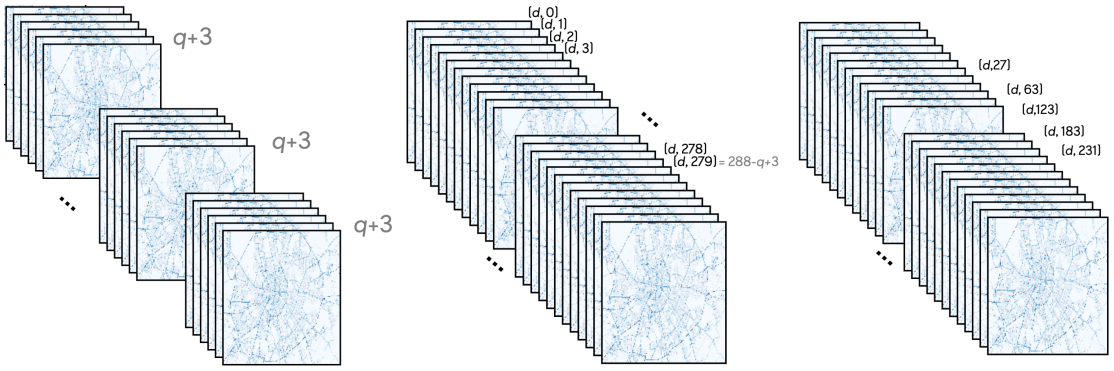

Figure 3.3: Sampling clips: The left image delineates the process of partitioning a video into non-overlapping segments, each containing $q + 3$ frames. Despite its simplicity and preservation of temporal continuity within each segment, this method does not facilitate the mixing of frames from disparate segments during training. The center image provides an overview of the sliding window methodology, which involves indexing day and time period pairs. Subsequent $q + 3$ frames are then sampled sequentially from each selected pair within a batch, enabling a flexible combination of sequential frames from varying segments. The rightmost image exclusively illustrates the retrieval of pairs from the test set at their specific time periods.

In practice, these sampling strategies proved invaluable in streamlining the model testing and training process. The *non-overlapping* method was instrumental in swiftly testing different architectures using a sub-sample of the dataset. The models that showed promising performance were then trained more extensively using the *sliding window* approach, which utilized all available data. Finally, the

like-test strategy was employed to fine-tune the top-performing model, focusing specifically on the test time bins required by the challenge.

3.5 Multimodal Model with Weather and Seasonal Encoding

Our proposed model incorporates multimodal data, extending beyond traffic videos to include exogenous information that could impact traffic patterns. Specifically, we incorporate the time of day, day of the week, and both current weather conditions and weather predictions for the next three time bins corresponding to the traffic prediction horizon. These exogenous variables are matched to each frame in the video sequence.

The weather data is sourced from World Weather Online [Weather 2019]. Importantly, we note that the weather data utilized for predicting traffic is not the a posteriori known weather, but rather, the predicted weather conditions at the time of the input sequence. We note that the granularity of the weather data is hourly, whereas the traffic data is recorded every 5 minutes. Therefore, we match the weather data to the nearest time bin.

Figure 3.4 depicts the encodings used for each exogenous variable, which include both one-hot encodings and continuous values. Specifically, the day of the week, day of the month, and day of the year are all represented as one-hot encoded vectors of dimensions 7, 31, and 365, respectively.

The time of day is represented through cyclical continuous features. This approach efficiently captures the cyclical nature of daily patterns.

As for weather variables, we incorporate different types of data. We construct a one-hot encoding vector that represents one of 28 possible weather states as defined in [Weather 2019]. These states include, but are not limited to, conditions such as cloudy, foggy, rainy, drizzly, and sunny.

In addition to these categorical representations, we also include continuous variables, each normalized by its range. These variables encompass temperature (in Celsius), 'feels like' temperature (also in Celsius), wind speed (in km/h), precipitation (in mm), and visibility (in km).

It's important to note that the spatial resolution for weather variables is one data point per city per time bin. While we acknowledge that weather patterns can vary within a city, our model uses the average conditions for simplicity. Future work may consider higher-resolution weather data to capture intra-city variations.

Since the resolution of the traffic data and these variables are different, they require data fusion. This process is explained in detail in the following section.

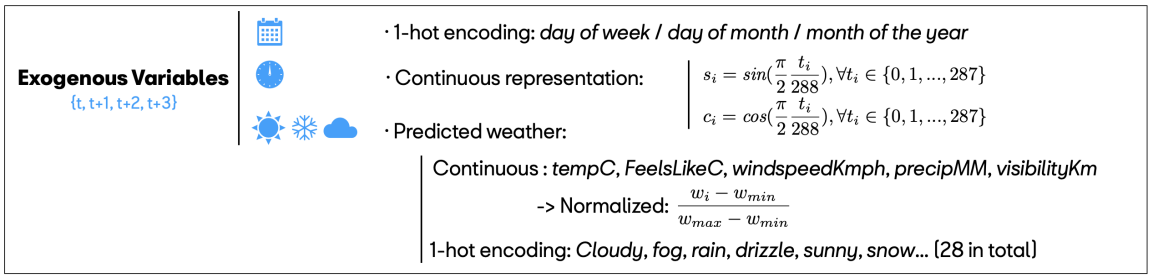

Figure 3.4: This figure visualizes our multimodal model's architecture, showcasing how a diverse set of exogenous variables are encoded as inputs. This is alongside the three types of traffic data. The included multimodal data comprise both one-hot encodings and continuous values.

3.6 Recurrent Autoencoder with Skip Connections

In this section, we present the proposed multimodal architecture along with its variants that we explore in our ablation studies.

3.6.1 The Model

Our model expands on the architecture proposed by [Wichers et al. 2018], which presents a neural network capable of predicting the subsequent frame (F_{t+1}) in a video sequence, given the previous frame (F_t). Utilizing an encoder followed by a recurrent layer, the model predicts the future frame's embedding, which is subsequently compared to F_{t+1} processed via the same encoder, leveraging an L2 loss. The embedding is then upsampled back to the original spatial resolution using a decoder. This process involves skip connections between different layers in the encoder-decoder tandem to fine-tune the output.

As the nature of our problem pertains to sequence-to-sequence prediction, we generalize the above architecture to leverage the full input sequence of length q, denoted as X_q, when forecasting the next three frames Y_3. This is achieved by processing each frame of the input sequence with the encoder iteratively and concatenating the resultant embeddings. A recurrent encoder then aggregates this temporal information from the input sequence into a single representation. Following

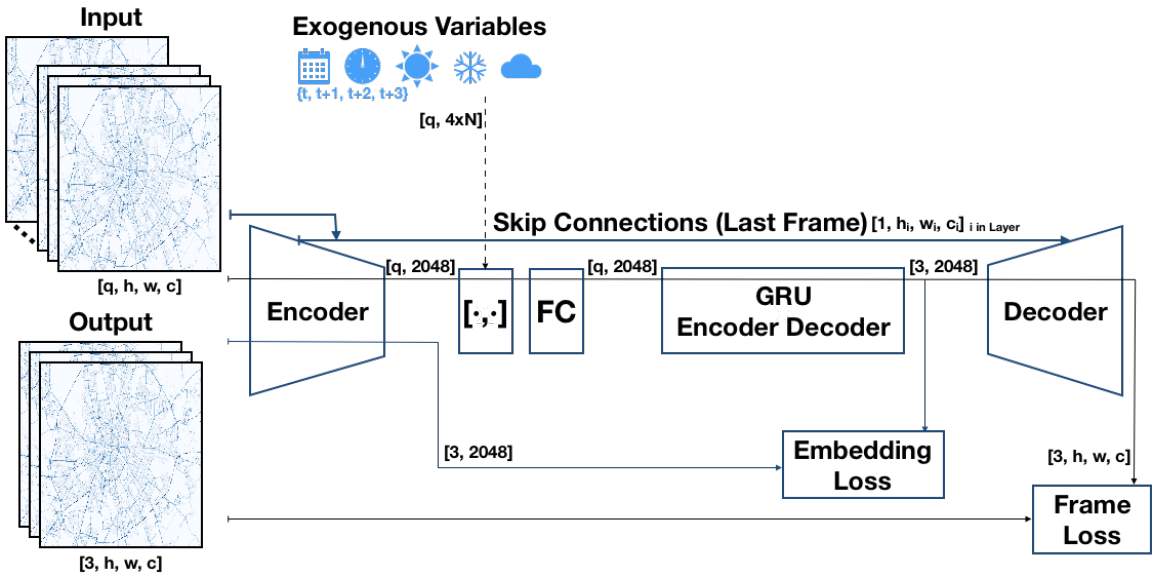

Figure 3.5: Multimodal Recurrent Autoencoder with Skip Connections. This model incorporates an *Embedding Loss* and GRU layers in the encoder to enhance low-dimensional predictions. By integrating these embeddings and skip connections (from sibling layers in the encoder), the model applies a loss function at the same resolution as the input images, significantly improving the quality of the decoder's outputs. Furthermore, exogenous variables are integrated alongside a fully connected layer, bolstering the model's predictive capacity. Please view it in electronic form for the best clarity. Our method code is available at `https://github.com/pherrusa7/Traffic4cast_NeurIPS_2019`.

this, a recurrent decoder generates three predictions in the embedded space, denoted as $\widetilde{e_3}$. These predictions are subsequently upsampled back to the original spatial resolution ($\widetilde{Y_3}$) using a decoder with skip connections. In addition, we include skip connections using only the final frame of the input video sequence (see Figure 3.6). We train both the predictions in the embedded and the original spatial resolutions using an L2 loss, weighted by $\alpha, \beta \in [0, 1]$:

$$L = \alpha L2, (Y_3, \widetilde{Y_3}) + \beta L2, (e_3, \widetilde{e_3}) \tag{3.2}$$

In our design, the encoder is composed of six distinct blocks. Each of these blocks integrates: i) two sub-blocks, each consisting of convolution, batch normalization, and a ReLU activation function; ii) Max pooling; iii) and a Dropout function with a value set at 0.5. The convolution count for each of the six blocks is sequentially [16, 16x2, 16x4, 16x8, 16x8, 16x2].

Moreover, we have incorporated multimodal exogenous data into our model. This data includes time-related factors such as the time of day and day of the week, as well as weather conditions, both current and predicted for the next three time intervals. These weather details are sourced from *World Weather Online* [Weather 2019]. Each frame encodes its corresponding exogenous variables along with the predicted data for the succeeding three time periods, ensuring that the most recent frame contains the forecast of the three target traffic frames. This exogenous information is then concatenated with the encoder's outputs in the embedding space.

The decoder is similarly structured with six blocks consisting of i) transposed convolutions [Dumoulin and Visin 2016]; ii) concatenation from the corresponding layer in the encoder; iii) Dropout set to 0.5; and iv) the same two sub-blocks as the encoder. The number of convolutions in the decoder block is [16x8, 16x8, 16x8, 16x4, 16x2, 16x1], consecutively.

The recurrent encoder-decoder employs layers of Gated Recurrent Units (GRUs) with unit sizes of (2048, 256, 128) and (128, 256, 2048) respectively. Frames are initially upsampled using bilinear interpolation to 512×512. The output is then adjusted to match the original size of 495×436 via cropping and a 3x3 convolution operation followed by ReLU activation.

Throughout the remainder of this paper, our proposed model will be referred to as the Recurrent Autoencoder All (*RAE_all*), which includes the skip connections (including the input), weather, and time data. The architecture of our model is illustrated in Figure 3.5.

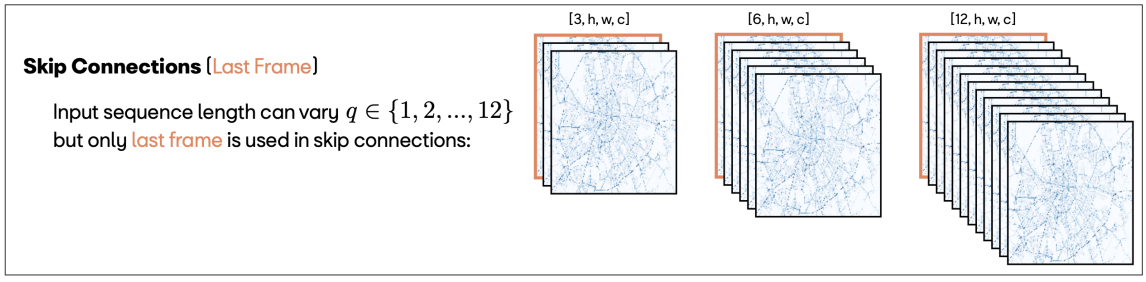

Figure 3.6: Illustration of skip connections utilizing solely the final frame from a sequence, irrespective of the sequence's length.

3.7 Model Variants and Baselines

In order to evaluate the importance of parts of the proposed model, we also evaluated several variations:

- *RAE_not_In*: This model is the same as the original one but it does not use skip connections.
- *RAE_not_Exo*: This model is the same as the original one but it does not incorporate exogenous variables
- *RAE_Clf*, This model is the same as the original one for the two regression outputs for *speed* and *volume*. However, instead of one output for the *heading* channel it produces five classification outputs, making this a multitask model in terms of two different types of objectives (regression and classification). In this model, the regression and classification tasks are minimized with an L2 loss and softmax cross-entropy, respectively. The seven outputs are then combined to generate the final three channels and minimized once more with an L2 loss. This approach is hypothesized to improve the accuracy of the *heading* channel predictions.

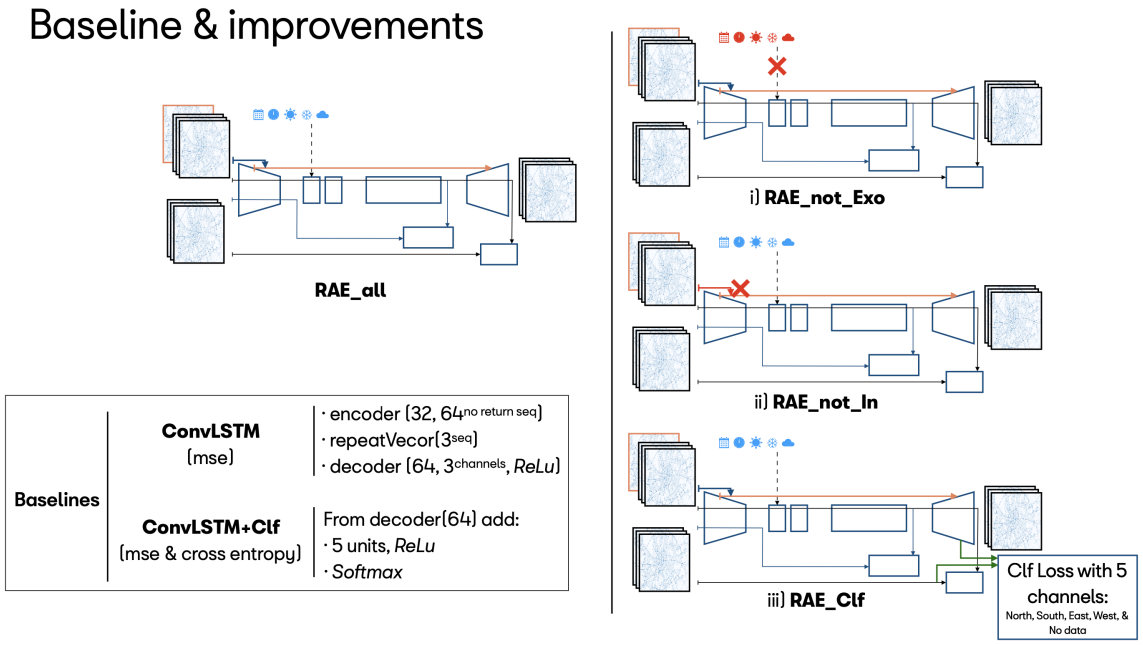

Figure 3.7: Depiction of the proposed architecture, its modifications, and the baseline models. The *RAE_not_In* and *RAE_not_Exo* variants highlight in red the components that they do not incorporate; namely, exogenous variables and skip connections, respectively. The *RAE_Clf* model introduces an additional classification layer for the heading channel, diverging from the regression approach to better match the channel's modality.

Two baseline models are established for comparison. The first, referred to as *ConvLSTM*, comprises three layers of Convolutional LSTM [SHI et al. 2015] with 32, 64, 64 units, respectively, followed by a *tanh* activation function. A final ConvLSTM layer with three units and ReLU activation is appended, and the entire model is trained using an L2 loss. Note that a final ReLU activation allows the model to predict a value of 1, which aligns with the *heading* channel's domain (see section 3.3), whereas *tanh* would not. The second baseline, *ConvLSTM+Clf*, augments the *ConvLSTM* model with an extra classification branch for the *heading* channel. The additional loss for this model is a softmax cross-entropy, just as in the variant *RAE_Clf*.

A depiction of the model variants and the baselines can be seen in Figure 3.7.

3.8 Results

Our experimental results draw upon the validation set furnished by the challenge (refer to section 3.3). This set essentially acts as a test set, given that the models are encountering this data for the first time. We created our own validation set by randomly selecting seven days per city from the given training set, which served to identify the best-performing epoch snapshots. Since the test set in the challenge lacks ground truth, local computation of performance is not feasible. At the end of this section, we relay the score provided by the competition cloud system for our top-performing model.

Table 3.2 presents the outcomes for the baseline models across the three cities, with a focus on the mean square error (MSE) for all three target channels and the accuracy (ACC) only for the *heading* channel. Then, ACC is computed by comparing the only five possible values representing directions in the ground truth (see section 3.3) with the output of the regression task. The results indicate that the integration of the classification task (as shown in the *ConvLSTM+Clf* column in Table 3.2) significantly improves MSE in Moscow. However, the improvement is marginal in Berlin and the performance in Istanbul actually deteriorates. Nevertheless, the accuracy for the *heading* channel improves in two out of the three cities, suggesting that the regression in this channel benefits from the multitasking approach.

We experimented with using the output from the classification branch in the *ConvLSTM+Clf* model as the output for the *heading* channel, by taking its *argmax* values. This increased the ACC from 0.455 to 0.803, but also raised the MSE from

0.012 to 0.023, which worsened the crucial metric for the challenge. Our conjecture is that the gap between the incorrect prediction *heading* values is now larger when using the default prediction minimized with MSE, as softmax cross-entropy does not yield closer values when the correct one is not guessed (it does not take into account inter-class similarities). Consequently, we decided not to use the classifiers for the challenge. However, we plan to explore a classification loss with a distance for similar classes in future work, which could assist the model in predicting the most similar class when an incorrect prediction is made.

The baseline models were trained exclusively with the *non-overlapping* sampling strategy, and input sequence with length $q = 3$, as the training time for 1 epoch exceeded 7 hours on an NVIDIA Titan RTX with 24GB. In contrast, 1 epoch of our proposed method only required 35 minutes, primarily because the recurrent layers operate in a much lower dimensional space. In our experiments, the weights for the two terms in the loss function of eq. 3.2 were determined through a limited grid search, due to time constraints, in the set $0, 0.5, 0.9, 1$, with the best results achieved for $\alpha = 1$ and $\beta = 0.9$. See details of the machine and main libraries used in table 3.1.

Table 3.1: Machine and tools used for the training and evaluation of the models at the Traffic4cast 2019 challenge.

GPU	NVIDIA Titan RTX with 24GB
CPU	Intel(R) Core(TM) i9-9900K @ 3.60GHz
Cores/Threads	8 cores / 16 threads
RAM	64GiB
Tools	Keras, TensorFlow, NumPy

Given the time restrictions of the challenge, our remaining experiments concentrate solely on enhancing the performance for the city of Moscow, as the baseline models exhibited the poorest results there. Future work will extend these comparisons to other cities.

Table 3.3 presents the performance of our proposed method and its variants, as tested in Moscow. The standout model is *RAE_all*, highlighted in bold, which was fine-tuned from *RAE_not_In* by incorporating a new skip connection from the final frame in the input sequence to the decoder. This addition allows the model to be further refined. Notably, model *RAE_Clf* demonstrates effective learning in the *heading* channel, maintaining a mse of approximately 0.0144, a noteworthy improvement from the 0.0239 score achieved by the same approach in the baseline.

Table 3.2: Comparative Analysis Across Three Cities: Baseline Performance. This table illustrates the mean square error (mse) for three target channels and the accuracy of the heading channel. Each model displays the number of epochs prior to early stopping. A '+' symbol denotes that the *ConvLSTM+Clf* model commenced training from the final snapshot of the *ConvLSTM* weights, with its new classification branch initialized with random weights.

mse (acc *heading*), #epochs	*ConvLSTM*	*ConvLSTM+Clf*
Moscow	0.0126 (0.265), 3	0.**0120** (0.455), +3
Istanbul	0.**0090** (0.657), 3	0.0096 (0.686), +3
Berlin	0.0071 (0.536), 5	0.**0071** (0.418), +1
mean	0.009618 (0.4860)	0.**009603** (0.**5196**)

Table 3.3: Performance Comparison of Proposed Method and Variations Against Baseline in Moscow: This table presents the global mean square error (mse) and the accuracy of the *heading* channel for our proposed method and its variations, contrasted against the baseline in Moscow city. The number of training epochs is also displayed. A particularity is the epoch format "10+5", which signifies that the model was fine-tuned for an additional 5 epochs using the weights from a previously trained model (to the left) that underwent 10 epochs.

	ConvLSTM+Clf	*RAE_not_Exo*	*RAE_not_In*	*RAE_all*	*RAE_Clf*
mse	0.012037	0.011873	0.011875	**0.011816**	0.014442
heading acc	0.455	0.469	0.453	0.437	**0.508**
epochs	4	10	10+5	15+3	44

Table 3.4: Results for Our Top-Performing Model, *RAE_all*, in Moscow: This table showcases the complete mse results in Moscow city for our premier model. As anticipated, the MSE decreases as the predicted time bin approaches, indicating enhanced forecast accuracy for closer time periods.

	Moscow — mse: 0.011816756		
	volume	*speed*	*heading*
5 minutes	0.000095	0.005128	0.029793
10 minutes	0.000102	0.005173	0.030223
15 minutes	0.000104	0.005214	0.030514

As future work, we aim to enhance this model since our goal is for *heading* outputs to belong to the five possible values in the domain strictly, factoring in a metric for inter-class similarity.

In Table 3.4, we provide detailed metrics for each channel and target time for the Moscow dataset. It's evident that the prediction difficulty increases with the length of the forecast horizon across all channels, although the performance decre-

ment is only marginal. Interestingly, the volume channel proves to be the least challenging, a result that aligns with our expectation given that most pixels in the images do not cover roads, thereby frequently reflecting a volume of zero. This detail further underscores why Moscow, which boasts a denser road network compared to Istanbul or Berlin, poses a more significant challenge for accurate prediction.

Our proposed architecture, when used for Moscow, along with the best baseline for Istanbul and Berlin, achieved a mse score of 0.00981 in the challenge. We anticipate that a broader application of our method, with training implemented for all cities on the proposed model, will yield substantially improved performance. Unfortunately, this could not be tested due to time and hardware constraints.

3.9 Discussion

In this study, we introduce a Recurrent Autoencoder with Skip Connections and Exogenous Variables for Traffic Forecasting. This model is specifically designed for a unique representation of traffic data, as introduced by the Traffic4cast Challenge at NeurIPS 2019. This representation transmutes aggregated city traffic data into video-like sequences of images, thereby capturing the temporal evolution of traffic patterns.

Our method harnesses the sequential nature of the input data, effectively integrating spatio-temporal information within a compressed space and generating the output sequence through a single inference pass. The model employs dual loss functions to ensure robust predictions in the embedding space and high-resolution reconstructions in the original space. We additionally integrate exogenous variables such as day of the week, time of the day, and weather conditions. A sequence sampling strategy is proposed that operates concurrently with the optimization process, resulting in diverse and rich batches at each training epoch. The efficacy of our approach is demonstrated by achieving a mean squared error of 0.00981 on the Traffic4cast Challenge test set.

Looking ahead, we plan to refine our multi-task approach that incorporates the classification results of the *heading* channel as input for the regression-based prediction task, ensuring that inter-class similarity is accounted for. Further work will be conducted to explore the intrinsic properties of geographical models, with the goal of identifying which deep learning architectures can best leverage the inherent rules that characterize the data [Jonietz and Kopp 2019].

3.10 Author Contributions

The work presented in this chapter is primarily the result of the author's (P. Herruzo) individual research and effort. The author was responsible for the conception and design of the study, acquisition and interpretation of data, development of the methodology, and drafting and revising the manuscript.

The author's supervisor (J. L. Larriba-Pey) played a crucial role in reviewing the work and providing substantial intellectual contributions. They offered critical feedback, helped polish the text, and guided the research with their expertise and experience. Their support and advice significantly enhanced the quality and impact of the work.

Part II

Open-Data Benchmarks: Multimodality, Adaptation, and Generalization in Deep Learning

Chapter 4

Weather4cast 2021: A New Spatio-Temporal Benchmark

This chapter is based on materials from the following peer-reviewed paper:

Herruzo et al. (2021a). High-resolution multi-channel weather forecasting – First insights on transfer learning from the Weather4cast Competitions 2021. In IEEE International Conference on Big Data (Big Data), 2021.

4.1 Introduction

Understanding and monitoring the meteorological dynamics of our planet have become critical components of our global consciousness. They have far-reaching implications, not only for our immediate daily activities, such as agriculture and traffic, but also for our long-term survival as we grapple with the consequences of climate change Bauer et al. (2015). The modern age has seen significant strides in our ability to observe and predict the weather, with weather observation emerging as a key domain in assessing our planet's health.

Satellite technology has transformed the field of weather monitoring. High-resolution time series, which offer detailed observational data across time, can now be collected by geostationary satellites like the Meteosat Second Generation (MSG) constellation. These satellite constellations remain fixed over particular geographical regions, providing frequent updates and enabling continuous monitoring Schmetz et al. (2002). This capability is essential as it supplements terrestrial observations that are often sparse or completely absent in many regions across the globe.

Historically, satellite measurements have been employed in physics-based semi-empirical models for estimating weather variables of interest and formulating weather predictions. However, in recent years, with the surge of Big Data and the advent of advanced machine learning techniques, there has been a paradigm shift in short-term weather forecasting Ravuri et al. (2021b), Bauer et al. (2021), Sønderby et al. (2020), Agrawal et al. (2019), Rasp et al. (2020), Berthomier et al. (2020).

Machine learning approaches have emerged as competitive alternatives to traditional mechanistic models, transforming our methods of weather forecasting. While these innovative methods have garnered significant interest in meteorological circles, they are also fascinating in the broader machine learning sphere, specifically concerning spatio-temporal data. **Weather satellite data, with its complex and dynamic characteristics, serves as a unique application domain and experimental field for developing and refining deep neural networks**. These networks seek to identify and model functional patterns inherent in dynamic stateful systems.

The successful application of machine learning techniques in weather forecasting is dependent on the availability of comprehensive, high-quality reference data, along with robust benchmarks for comparison. Ideally, these benchmarks and datasets should capture the inherent complexity and diversity of weather patterns across the globe, reflecting the spatiotemporal distribution shifts across different spatial regions and time periods.

However, a closer look at the currently available shared datasets reveals some notable limitations. Some of these datasets lack high resolution, which can undermine the effectiveness of machine learning algorithms Rasp et al. (2020), Schroeder de Witt et al. (2021). Others may provide high resolution but are restricted in their geographical coverage, representing only specific regions like the United States of America or the European Union Veillette et al. (2020b), Saltikoff et al. (2019b). This limited diversity can hamper the development and testing of machine learning models intended for global application.

To address this gap, we introduce the Weather4cast competitions. These competitions are pioneering in that they provide common reference data and benchmarks, specifically designed for evaluating machine learning models on multi-channel, high-resolution weather satellite data. The challenge is constructed to test model predictions in previously unseen locations (spatial transfer learning) and in the year succeeding the training data period (temporal transfer learning).

4.2 Related Work

We here outline early pioneering work in the field, related datasets, and remaining challenges. Deep learning models have for the first time achieved competitive results in short-term forecasting / now-casting Sønderby et al. (2020), Agrawal et al. (2019), Berthomier et al. (2020), Veillette et al. (2020b). Remarkably, the MetNet model Sønderby et al. (2020) obtained better results than operational numerical weather models up to 8 hours into the future. The underlying data incorporate high-resolution satellite and ground radar precipitation measurements but are limited to the USA. Moreover, the data unfortunately have not been publicly released.

Open weather datasets currently available include: **RainBench** Schroeder de Witt et al. (2021) and **WeatherBench** Rasp et al. (2020), with a coverage of the entire globe but at low resolution both in space and time. **EarthNet2021** Requena-Mesa et al. (2021), high resolution in space but very low resolution in time (5 days). **OPERA** radars Saltikoff et al. (2019b), with both high resolution in both space and time but only covering the European Union (EU). **SEVIR** Veillette et al. (2020b), at high temporal and spatial resolution but only covers the United States (US), and samples on the same location are not guaranteed to be retained for longer than 4h. This implies that a certain region can't be systematically compared with other regions for longer periods of times. For an overview of these datasets see Table 4.1.

Conversely, the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT) hosts an extensive database of satellite radiances and derivative products from the Meteosat Second Generation (MSG) on their website (archive.eumetsat.int). While these resources can theoretically be accessed without cost, acquisition of a license is subject to approval, contingent upon the intended use of the data and membership of the applicant's country in the EUMETSAT consortium, among other constraints. Furthermore, the process of bulk downloads is a substantial impediment due to its extensive duration, which can range from a week to several months. Cumulatively, these factors present significant barriers to data access, particularly for those without specialized knowledge or resources.

4.3 The Weather4cast Formulation

4.3.1 Datasets

The dataset utilized in the Weather4cast competition originated from the Meteosat geostationary meteorological satellites, managed by the European Organisation

Table 4.1: Comparison of key weather datasets. The table lists Dataset name, Source and sensor name, the number of Variables, spatial Resolution, Grid size and Coverage, Sampling time, and the total number of Data points. Spatial resolution for satellite measurements is the best resolution achieved, as given at the nadir, i.e., directly below the satellite. Private datasets are marked with *.

Dataset	Source	Variables	Resolution (Grid / Coverage)	Sampling	Data
WeatherBench	ERA5	110	156 km (128×256 world)	1h	1×10^{12}
RainBench	ECWMF ERA5 IMERG	3 110 1	156 km (128×256 world)	3 h 1 h 30 min	2×10^{9}
MetNet*	GOES-16 MRMS	16 1	1 km (1024x1024 US patches)	15 min 2 min	2.3×10^{13}
SEVIR	GOES-16 NEXRAD	5 + storm desc 1	2 km (192x192 US patches)	5 min	1×10^{11}
EarthNet2021	Sentinel-2 E-OBS	7 7	20 m (128x128 EU patches)	5 days daily	2.6×10^{11}
Opera	Opera radar	3	2 km (200 radars EU)	15 min	1×10^{9}
Weather4cast21	NWC SAF	23	3 km (256x256 EU, N. Africa, M. East patches)	15 min	2.3×10^{11}

for the Exploitation of Meteorological Satellites (EUMETSAT). This dataset encompasses a multitude of weather variables, meticulously derived from satellite data by dedicated EUMETSAT Satellite Application Facilities (SAF) units, specifically the Nowcasting (NWC SAF).[1] The information contained within the Meteosat's second-generation images was subsequently processed via NWC SAF software, created by a consortium of national meteorological services, which includes the Spanish State Meteorological Agency AEMET.[2]

The following weather products were delineated as target variables for prediction in the proposed benchmark: surface accessible temperature (which could either be the top of a cloud or the earth's surface), convective rainfall rate, probability of occurrence of tropopause folding, and the cloud mask. An example of these variables can be seen in Fig. 4.2. Moreover, the dataset included an additional 25 weather products comprising auxiliary variables like cloud type, accumulated rain rate, pressure, alongside a selection of quality flags. Although these variables were not required for prediction, they were available to be incorporated as supplementary model inputs by the participants.

[1] `http://nwc-saf.eumetsat.int/`nwc-saf.eumetsat.int

[2] Acknowledgement of the data source: The competition data contains modified AEMET/NWC SAF products from February 2019 to February 2021.

Each of the weather products was encoded as distinct channels in the weather images, with every image consisting of a 256×256 pixel grid. Every pixel represented an approximate area of $4,$ km $\times\ 4,$ km. The images were captured at regular 15-minute intervals over the span of a year. Subsequently, the competition data were divided at random, on a per-day basis, into training sets (constituting 80% of the days), validation sets (10%), and test sets (10%).

The **core dataset** included training, validation, and testing sets for five distinct geographical regions, as shown in Fig. 4.1:

- R1 – The Nile region, featuring Cairo,
- R2 – Eastern Europe, with Moscow,
- R3 – South West Europe, including Barcelona,
- R7 – The Bosphorus region, with Istanbul,
- R8 – Eastern Maghreb, including Marrakech.

On the other hand, the **spatial transfer learning dataset** solely included the testing sets for an additional six regions:

- R4 – Central Maghreb, featuring Timimoun,
- R5 – South Mediterranean, with Tripoli and Tunis,
- R6 – Central Europe, including Berlin,
- R9 – The Canarian Islands,
- R10 – The Azores Islands,
- R11 – North West Europe, including London, Paris, Brussels, and Amsterdam.

Furthermore, static geographical data, such as altitude, longitude, and elevation, were provided within separate channels for each pixel, ensuring comprehensive georeferencing for the entire dataset. This detailed data specification facilitates a thorough analysis and understanding of the unique weather patterns exhibited across these diverse geographical regions.

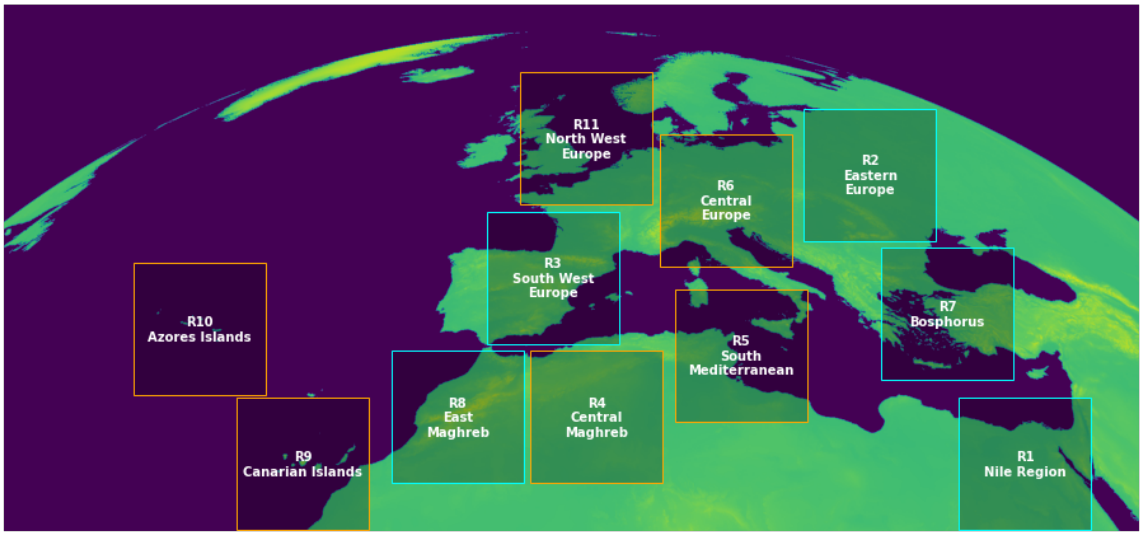

Figure 4.1: The dataset spans several selected regions marked in distinct colors. The blue-marked regions pertain to the core challenge, where participants have access to training data. Conversely, the orange-marked regions are exclusively for the spatial transfer learning challenge, in which no training data is offered. These regions are notably varied, encompassing a wide range of typical weather conditions and spanning across a diverse geographic domain in terms of latitude and longitude.

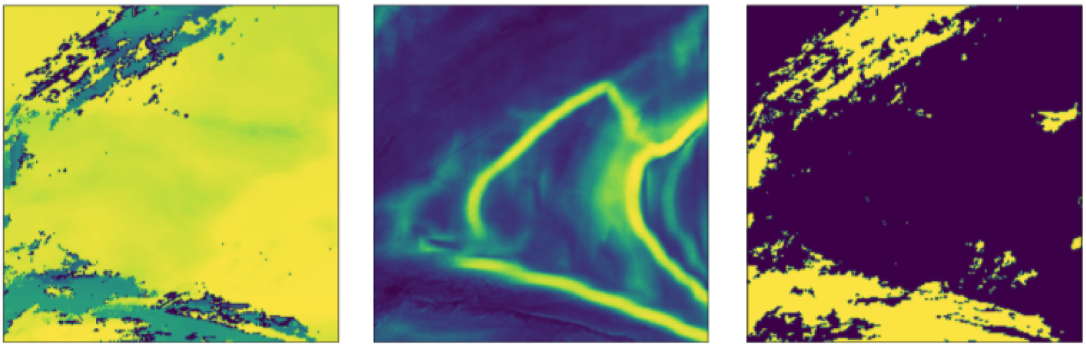

Figure 4.2: The illustrated examples represent target variables. Commencing from the left, typical frames are displayed for temperature, tropopause turbulence probability, and cloud mask, specifically for region R3, which covers South West Europe.

4.3.1.1 Processing of the Temperature Variable

EUMETSAT provides two distinct products for the variable of temperature, each with its unique data collection methodology and limitations. The first product presents temperature measurements taken at the surface level, although these measurements are exclusively available for regions devoid of cloud cover. In instances where clouds are present, obstructing the surface-level readings, the dataset marks these instances as a missing value.

Conversely, the second product measures the temperature at the tops of clouds, providing a valuable counterpoint to the surface-level data. This product similarly assigns a missing value to pixels corresponding to cloud-free areas. This dual dataset structure poses a challenge for deep learning models operating on images, as such models typically do not handle missing values effectively.

To mitigate this issue and enable efficient utilization of the available data, we have unified these two distinct variables into a single, cohesive dataset. This process entails integrating surface-level temperature measurements, top-of-cloud temperature readings, and instances where missing values are assigned due to the presence of clouds or the absence thereof.

Figure 4.3 visualizes this process of integration, demonstrating the methodology used to amalgamate these two disparate datasets into a unified, contiguous temperature dataset. This fusion of data ensures that the deep learning models can effectively operate on the entire dataset, circumventing potential difficulties associated with missing values. Furthermore, this approach maximizes the informational value of the EUMETSAT products by leveraging both surface-level and top-of-cloud temperature readings, providing a more comprehensive picture of the temperature variable.

4.3.2 Tasks

The Weather4cast benchmark, building on the precedent established by our NeurIPS 2019 competition in the traffic domain Kreil et al. (2020), and also drawing inspiration from Google Research's rainfall prediction approach Sønderby et al. (2020), Agrawal et al. (2019), revolutionizes weather forecasting by treating it as a video frame prediction task. This approach is further enriched by incorporating two distinct yet interconnected tasks, each representing different aspects of weather prediction and providing diverse challenges for the participating models. These tasks are:

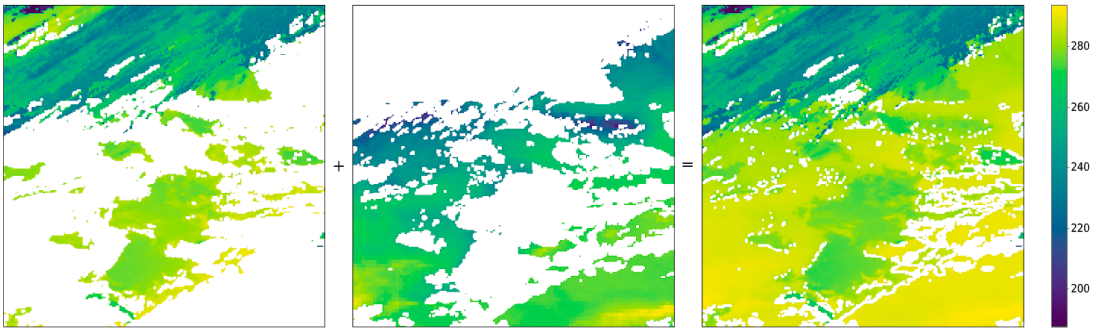

Figure 4.3: Unification of Surface and Top Cloud Temperature Variables. This figure illustrates the process of consolidating two separate temperature readings into a singular measure. The left panel represents the top cloud temperature variable, which offers temperature measurements at the top of the clouds and missing values for cloud-free pixels. The middle panel demonstrates the surface temperature variable, which provides measurements when there are no occluding clouds and returns missing values when cloud cover is present. The right panel showcases the unified temperature variable, derived from combining the surface and top cloud temperature readings, providing a comprehensive measure of temperature irrespective of cloud presence. All values are expressed in Kelvin (K).

1. **The Core Challenge**: This primary task necessitates the prediction of four selected weather parameters, namely, temperature, convective rainfall rate, probability of occurrence of tropopause folding, and the cloud mask. The geographical focus is centered on the core dataset regions, identified as R1, R2, R3, R7, and R8, where comprehensive training data is provided. These regions are visualized in Fig. 4.1 for better comprehension.

2. **The Spatial Transfer Learning Challenge**: This challenge expands upon the boundaries of the Core Challenge, requiring the models to extrapolate their predictions to six new, unseen regions (R4, R5, R6, R9, R10, and R11). Significantly, no training or evaluation data is provided for these regions, thereby challenging the models to extend their learned pattern recognition and prediction abilities to novel locations.

Under both tasks, the models are required to forecast the next 32 weather frames (equating to 8 hours of weather, sampled at 15-minute intervals) based on the preceding 4 images (corresponding to the last 1 hour of weather). This translates into predicting the four target weather characteristics over the entire test dataset regions. The format of the prediction necessitates an array of size (32, 4, 256, 256)

for each of the 36 test days provided, as demonstrated in Fig. 4.4. In the Spatial Transfer Learning Challenge, models can only use 1 hour of weather data for the 36 test days.

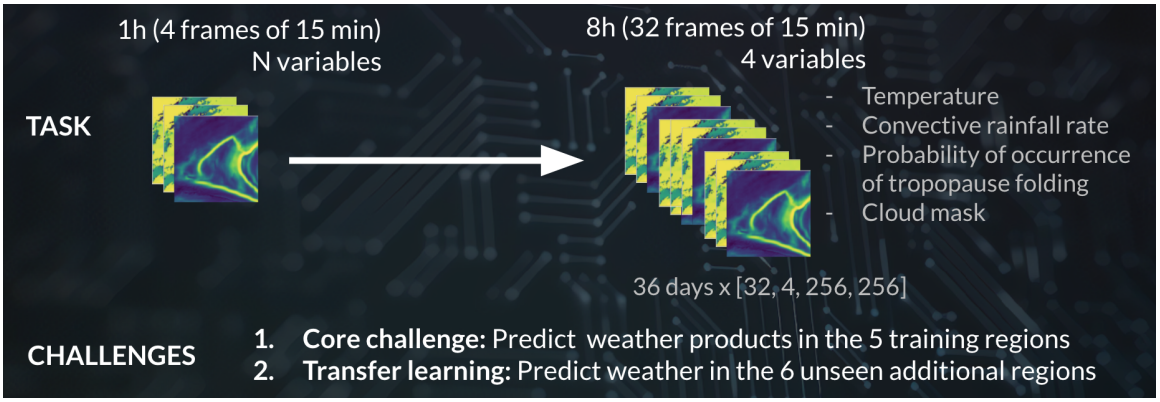

Figure 4.4: The Weather4cast competition posed a unique task and set of challenges to the participants. They were required to generate predictions spanning 8 hours for four distinct variables. The core challenge required the generation of predictions for locations for which training data was made available. On the other hand, the spatial transfer learning challenge demanded predictions for regions where no training data was provided.

4.3.2.1 Target Variables Distributions

The benchmark's proposed target variables are inherently heterogeneous, embodying a matrix of intricate correlations. A clear visualization of this complexity is depicted in Figure 4.5, which presents the distribution of each variable within region 3 as a representative example.

The distribution of the temperature variable appears to mimic a mixture of Gaussians, suggesting the operation of two underlying processes. There is a higher mean temperature associated with the surface level measurements, and a lower mean corresponding to the temperatures at the tops of clouds. This bimodal distribution signifies the distinct temperature profiles associated with these two environmental strata.

Contrastingly, the rainfall rate variable exhibits a zero-inflated distribution, indicating that the majority of pixels across all regions typically register no rainfall, rendering most of the dataset dry. Despite this, the distribution of rainy days displays substantial variability across different regions, hinting at regional differences in precipitation patterns.

Tropopause folding occurrence is quantified as a probability and demonstrates a distribution that is skewed to the right. This skewed distribution reflects the relative rarity of this meteorological phenomenon.

Lastly, the cloud mask variable operates on a binary system, wherein the value of one signifies the presence of a cloud at a specific pixel, and a value of zero denotes cloud absence. This simple binary distinction translates into a distribution that distinctly marks the extent of cloud cover within a region.

In summary, the distributions of these target variables, as illustrated in Figure 4.5, underscore the complex, heterogenous nature of the dataset and highlight the varied statistical patterns each variable exhibits. This richness in data and the associated complexities warrant careful consideration in the subsequent modeling processes.

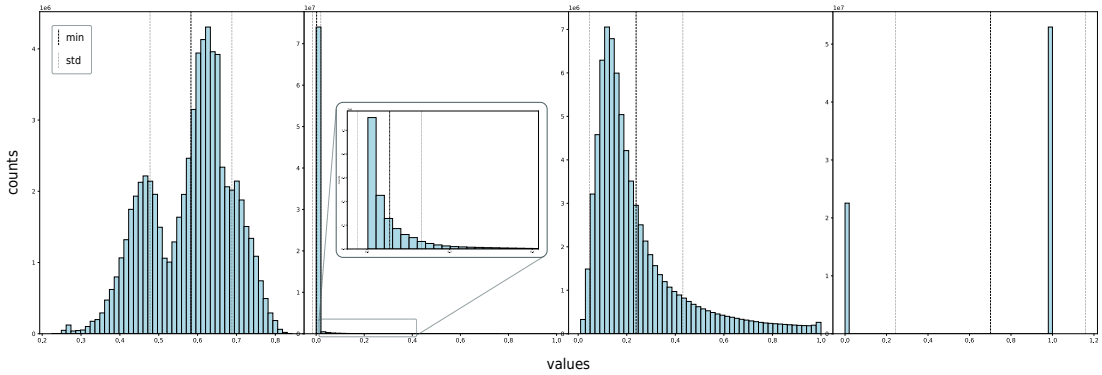

Figure 4.5: Distribution of Target Variables in Region 3. From left to right, this figure illustrates the distribution characteristics of the four target variables - surface and top cloud temperatures, rainfall rate, tropopause folding probability, and cloud mask. The temperature variable appears to follow a Gaussian mixture distribution, with higher mean values for surface temperatures and lower mean values for top cloud temperatures. The rainfall rate exhibits zero inflation, with most pixels across all regions recording no rainfall, although the distribution of rainy days varies by region. The tropopause folding probability presents a right-skewed distribution. The cloud mask, a binary variable, indicates the presence (value 1) or absence (value 0) of cloud cover at each pixel.

This novel structure of the Weather4cast benchmark offers a comprehensive platform for testing and comparing different weather forecasting models. It allows for the exploration of innovative techniques and methodologies in machine learning, weather forecasting, and spatio-temporal transfer learning, with the unique feature of dealing with both known and unknown regions. This, in turn, con-

tributes to the broader discourse in the field of meteorology and machine learning, pushing the boundaries of what these models can achieve.

4.3.3 Metrics

Our proposed evaluation metric extends the standard average mean squared error, with modifications to account for the idiosyncrasies of the data. This involves three key steps: (i) we stabilized the score variance by employing a subsampling approach on the measured data points to account for varying levels of missing values in the target data derived from satellite images, (ii) we ensured all error variables are equally weighted, and (iii) we transformed variables representing probabilities into real number representations.

In the test dataset, the evaluation encapsulates a total of $D = 36$ days, with $T = 32$ time intervals and $T_{\text{pixels}} = 256 \times 256$ total pixels. This spans across regions R1, R2, R3, R7, and R8 in the Core Challenge (a total of $N_{\text{Core}} = 5$ regions), and regions R4, R5, R6, R9, R10, and R11 in the Transfer Learning Challenge ($N_{\text{Transfer}} = 6$). The models are tasked with predicting $V = 4$ target variables.

Here's a more detailed explanation of how the error computation works:

- For the temperature variable, a fixed number of non-missing values ($T_{\text{temperature}} = T_{\text{pixels}} - M$) are sub-sampled to stabilize the error variance, a necessary measure due to the substantial variance in the number of missing values across different frames (up to $M = 53936$ missing values in an image). For the remaining variables (v), all pixels are utilized ($T_{\text{v}} = T_{\text{pixels}}$).

- The variable representing the tropopause folding probability is mapped to a logit scale (like in a logistic regression), ensuring that the computation of a mean squared error is meaningful. $\pm\infty$ are mapped to ± 5.806, a value distinct from the other finite values, through a data-driven approach. The total range is subsequently normalized to $[0, 1]$. Listing 4.1 shows the actual implementation in Python.

- To ensure an equitable combination of errors from different variables, we normalize the errors of each variable by dividing it by the error of a simple baseline model (the persistence model), $w(v)$. This scoring scheme assigns a score greater than 1 to predictions that perform worse than the simple persistence baseline, and a score less than one to those that perform better.

Then, the formula to compute the score for each task $C \in \{core, transfer\}$ is defined in Equation 4.1.

$$\text{Score}_C = \frac{1}{DTN_CV} \sum_{d=1}^{D} \sum_{t=1}^{T} \sum_{r \in R_C} \sum_{v \in \text{Variables}} \frac{1}{T_v \times w(v)} \sum_{p=1}^{T_v} (y_p - \hat{y}_p)^2 \tag{4.1}$$

where y and $\hat{y}$ are the real and prediction values at pixel p.

Listing 4.1: Python code for logit function

```
def logit(x):
    return np.log(x/(1-x))

def norm_logit(x, M = 0.997, m = 0.003):
    """ log(M) = - log(m) = 5.806 """

    # 1. clip pre-logit to avoid log(0) and log(1/0)
    x[x>=1] = M
    x[x<=0] = m

    # 2. apply logit
    x = logit(x)

    # 3. normalize logit
    M = -logit(m)
    x += M
    x /= 2*M

    return x
```

4.4 Associated Competitions and Provided Software

The Weather4cast datasets and benchmarks serve as a unifying force across diverse disciplines, bridging machine learning, time series analysis, geospatial science, meteorology, and Earth observation. These resources aim to foster collaborative research efforts and invite insights from a broad spectrum of academic communities. As a testament to this goal, we have successfully orchestrated two scientific competitions which were accepted at distinct conferences.

In the following section, we will detail the various stages of the competition, the resources made available to participants, and the support mechanisms provided. This includes a comprehensive starter kit equipped with code and baseline models, as well as data access provisions.

4.4.1 Weather4cast Venues 2021

We hosted 2 incremental competitions accepted in the following venues:

- Stage 1 (April 1 – June 30, 2021): The core challenge used the subset of regions R1, R2, and R3 while the transfer learning challenge used regions R4, R5, and R6. Selected papers were invited to present their work in the 1st Workshop on Complex Data Challenges in Earth Observation[3] at the 30th ACM International Conference on Information and Knowledge Management Gruca et al. (2021b).

- IEEE Big Data Cup (July 1 – October 27, 2021): This challenge used the complete dataset. The core challenge used the subset of regions R1, R2, R3, R7, and R8 while the transfer learning challenge used regions R4, R5, R6, R9, R10, and R11. Selected papers were invited to present their work in a special session devoted to the challenge at the IEEE Big Data conference [4] Herruzo et al. (2021a).

In any of the two competitions, participants could choose to take part in only the core challenge, the transfer, or both. Participants would have to upload their predictions for the test datasets to the leaderboard of the core and transfer learning competitions for quick feedback on model performance.

Everyone was allowed to submit up to five predictions but could delete older submissions to upload more – so an infinite number of submissions was possible but only in a rate-limited fashion. As is common to avoid over-fitting in the course of a competition, however, the final score was not based on the test dataset but on an additional held-out test dataset (*final dataset*), which was released during the last week of the competition. This *final dataset* contained data for the same regions in both core and transfer competitions but for the year following the period of the training data (temporal transfer learning).

For the *final dataset* scoring could only submit up to three predictions. Those submissions could not be deleted, and the best score of these three was used to determine the final place on the leaderboard for each participant.

[3] www.iarai.ac.at/CDCEO21

[4] `http://bigdataieee.org/BigData2021/BigDataCupChallenges.html`

4.4.2 Weather4cast Software

We developed a website to easily access Weather4cast: (http://weather4cast.ai). This provides the frontend and backend that enables (i) registration to the competition, (ii) downloading the data, (iii) submitting predictions, (iv) monitoring the leaderboards, and (v) interacting with other participants and organizers in the forum sections.

In addition, in the Weather4cast GitHub repository[5] we provided a starting kit featuring:

- Data links and descriptions
- Tutorials on how to read, visualize, and save the weather movies
- Sample data loaders to input data for model training
- Tutorials on how to train and run predictions for validation benchmarks
- Baseline models implemented in PyTorch, including:
  - A persistence model
  - A deep learning model based on the UNet architecture Ronneberger et al. (2015) which its variants did well in similar tasks before Choi (2019; 2020)
  - An ensemble model combining several deep neural networks inspired by Choi (2020)
- Instructions and examples of how to submit predictions to the online leaderboards.
- An example of an inference script to compute predictions with already trained models.

4.5 Baselines and Models

In this section, we first present our proposed model which will serve as a baseline. Then, we describe all the Weather4cast 2021 models used in the benchmark.

[5] `https://github.com/iarai/weather4cast`github.com/iarai/weather4cast

4.5.1 Conditional U-Net

The persistence model is a frequently used baseline in weather forecasting, recognized for its simplicity and surprisingly challenging performance to surpass Agrawal et al. (2019). This model functions as an identity mapping, using the present state as the forecast for all future states.

The U-Net architecture Ronneberger et al. (2015) has demonstrated remarkable success across various applications including image segmentation Ronneberger et al. (2015), precipitation forecasting Agrawal et al. (2019), and traffic prediction Kreil et al. (2020), Choi (2019; 2020), Herruzo and Larriba-Pey (2020). Thus, it offers a valuable point of comparison. Here, we extend and adapt this type of model proposing a few strong baselines for the Weather4cast 2021 benchmark:

Conditional U-Net, individual for regions R1, R2, R3, R7, and R8: Standard U-Nets have been utilized for one-hour ahead precipitation forecasting Agrawal et al. (2019), a configuration that permits complete hourly predictions by iteratively incorporating the preceding hour as input. In contrast, the MetNet approach Sønderby et al. (2020) introduced the future time to predict, $t + n$, as an additional input parameter. We found this adaptation more versatile as it permits querying the model for any trained lead time without requiring iterative input of subsequent predictions. Rather than adopting a one-hot encoding, as in MetNet, to represent the future time-bins for prediction, we introduce a scalar from 0 to 1 that indicates the intended prediction lead time. The scalar represents the ratio $n/32$ which is appended as an extra channel to the input sequence and fed into a U-Net (in 8 hours there are 32 intervals of 15 min). The time dimension T of the tensor is collapsed into the channels dimension C, expanding the depth of the channels: $[T, C, W, H] \mapsto [T \times C, W, H]$.

Ensemble of U-Nets: For the core challenge, individual conditional U-Nets are trained for each of the five regions. For the spatial transfer learning challenge, predictions for the additional regions are produced by averaging the outputs of all five U-Nets.

These baseline models were implemented using NumPy Paszke et al., PyTorch Paszke et al., and PyTorch Lightning Falcon et al. (2019), with a batch size of 32 and utilizing a Tesla V100 (see table **??**). We set the training to 10 epochs with early stopping implemented after 3 epochs without improvement, using the best-performing epoch for subsequent inference. The Adam optimizer Kingma and Ba (2017) was

employed with a learning rate of 10^{-3}. All associated code and model weights can be accessed on our GitHub repository at github.com/iarai/weather4cast.

Table 4.2: Machine and tools used for the preparation, training and evaluation of the models for the Weather4cast 2021.

GPU	8x Tesla V100-PCIE-32GB
CPU	2x Intel(R) Xeon(R) Gold 6146 CPU @ 3.20 GHz
Cores/Threads	12 cores / 24 threads (per CPU)
RAM	754Gi (Samsung DDR4-2600)
Tools	PyTorch, PyTorch Lightning, NumPy

4.5.2 Weather4cast 2021 Models

The top-scoring model in both the core and transfer learning competitions was a recurrent convolutional network employing residual units as opposed to traditional convolutions, as proposed by Antifugue Leinonen (2021a). A notable feature of this architecture was its capacity to maintain comparable performance with a 75% reduction in the number of parameters, achieved by implementing a shallower version of the network. The findings indicated that performance improvements were largely attributable to an increase in training data, particularly in the spatial transfer learning task, which saw an enhancement of over 5%. Alternative techniques also contributed to improvements, albeit marginally. These techniques included replacing the Adam optimizer Kingma and Ba (2017) with AdaBelief Zhuang et al. (2020), using two separate models conditioned on a rain rate threshold, and utilizing an ensemble of models, the latter being the second most significant factor for improvement.

The second-ranking model utilized an ensemble of U-Net models with densely connected blocks, similar to the architecture used in Stage-1 Choi (2021). In addition to the target variables, all other extra variables were also inputted to the model. Separate models were trained for each target variable, using mean square error for numerical variables and binary cross entropy for the binary variable cloud mask. Each model utilized all training regions as input.

The third-ranking model in both competitions was developed using a variant of the U-Net architecture, referred to as a Variational U-Net. This model introduces a probabilistic layer that samples from a normal distribution at the bottleneck, and

includes data augmentations such as rotations and flips, along with the application of model ensembles Kwok and Qi (2021a). Two ensemble strategies were employed: one utilized predictions from distinct models trained from scratch, while the other used predictions from the same model at different training epochs. The results underscored the importance of ensembles for enhancing performance in both competitions. Moreover, the findings showed that fine-tuning models for new training regions proved to be just as effective as initializing training from scratch with all regions, a crucial insight for the efficient scalability of models globally.

The model that secured the fourth place in both competitions was noteworthy, however, the specifics of its methodology remain unpublished as of now. The fifth-place model in both competitions, developed using a Visual Transformer, was based on a U-Net encoder-decoder scheme. Each block within this model was replaced with Swin-Transformers Bojesomo et al. (2021), which were adapted to 3D patches to encapsulate time. The model concurrently produced all target variables. One version of the model contained more parameters (achieved by enhancing the depth of U-Net) and incorporated extra variables as inputs. Conversely, the second version used 20% fewer parameters and solely utilized the target variables as inputs. Intriguingly, the first version performed superiorly in the core competition, while the second was more effective in the spatial transfer learning competition. The larger Visual Transformers were found to exhibit a higher tendency towards overfitting, particularly when applied to unseen regions. This indicates a balance that must be struck between model complexity and its capacity to generalize.

4.6 Results

The Weather4cast competition has emerged as a robust benchmark for the development and comparison of machine learning models in the realm of weather forecasting. Several models have proven their ability to meet the competition's main challenge of predicting high-resolution, multi-channel weather patterns for an eight-hour period in regions for which comprehensive training data is available. In addition, Weather4cast has introduced a pioneering benchmark framework for both temporal and spatial transfer learning, thereby enabling the creation of models with superior generalization. Such models are better poised to robustly capture the key patterns underpinning observed weather phenomena. This robustness is a critical factor for practical applications and the discovery of these essential patterns can provide insightful revelations about the physical dynamics of weather.

Model performance in this competition was consistent with results seen in related tasks, such as video encoding Kreil et al. (2020) and the Stage 1 competition of Weather4cast Gruca et al. (2021b). U-Net-like architectures demonstrated their utility in the core challenge of predicting eight hours of high-resolution, multi-channel weather. However, the model that clinched the top position employed a recurrent convolutional model incorporating residual units. The competition also witnessed the successful deployment of Variational Autoencoders and Visual Transformers, which yielded competitive results.

Table 4.3 provides a comparison of the top five models against baselines. Notably, the performance ranking was consistent between the core prediction challenge and the spatial transfer learning challenge, suggesting that the ability to **generalize to new spatial locations largely depends on the extent to which weather patterns are accurately learned in regions where ample training data is available**. The best model was able to reduce the error relative to the Persistence baseline by 53% and relative to the U-Net baseline by 30% in the core challenge. In the spatial transfer learning challenge, the error reduction relative to an ensemble of five U-Net models as baseline was 27%. Furthermore, the performance difference between the top two models and the next best models was significantly larger in the transfer challenge than in the core challenge, underlining the superior generalization capabilities of the top two models.

Stage 2 of the competition offered additional training data from new regions which, according to several participants, significantly improved model performance, especially in the spatial transfer learning challenge Leinonen (2021b), Kwok and Qi (2021b). An important question to explore is the extent to which this performance boost is attributed to the sheer increase in training set size versus the added diversity in the data. An in-depth exploration can be accomplished through experiments with training sets sub-sampled for size equal to the Stage-1 training sets. In addition, the incorporation of pre-trained models has proven to be an effective strategy to leverage additional data, as exemplified by the variational autoencoder deployed by Kwok and Qi (2021b).

Table 4.4 offers a comparison of the input data used by each model, the strategy for feeding different regions into models, and the application of any data augmentation. The winning model did not necessitate static data like latitude, longitude or variables other than the target variables for achieving impressive performance. However, other teams reported benefits from the inclusion of additional inputs. It was unanimously reported that substantial benefits were derived from training a

Table 4.3: Errors of the models' predictions on the leaderboards (the lower the better). The first column displays the prediction error score in the core competition and its improvement compared to a basic U-Net model. The last column shows the error scores for the spatial transfer learning competition and compares improvements achieved relative to (i) a basic U-Net trained only in one region and (ii) an ensemble of U-Nets using all five training regions. Interestingly, ranks from the core challenge align with ranks from the transfer challenge: Models performing better in the core challenge thus also generalized better to new locations. Observe that the top two models show greater performance in spatial transfer learning, showcasing their superior generalization capabilities.

Username	**Final Core (w.r.t. U-Net %)**	**Final Transfer (w.r.t. U-Net R1 — Ens R1,2,3,7,8 %)**
antifuge	0.4728 (30.3)	0.4323 (37.3 — 26.6)
sungbinchoi	0.4801 (29.2)	0.4376 (36.5 — 25.7)
Michael Fish Forecasting	0.4856 (28.4)	0.4603 (33.2 — 21.8)
moto	0.4935 (27.3)	0.4611 (33.1 — 21.7)
ai4ex	0.4985 (26.5)	0.4619 (33.0 — 21.5)
U-Net, individual, R1,2,3,7,8	0.6785 (0.0)	
Persistence	1 (-47.4)	1 (-45.1 — -69.9)
U-Net Ensemble R1,2,3,7,8		0.5887 (14.6 — 0.0)
U-Net Transfer from R1		0.6892 (0.0 — -17.1)

single model covering all locations, with the two top submissions training separate models for each prediction target variable.

All the top three submissions employed model ensembles, which consistently outperformed the predictions from the single model with the best validation score Kwok and Qi (2021b), Leinonen (2021b), Choi (2021). Notably, Kwok and Qi (2021b) demonstrated that weights from different epochs of training a single model can be used directly to generate an ensemble of models. Each individual model in this setup utilized the weights learned at a specific epoch. The highest performing individual models were then aggregated into an ensemble prediction, with the final result calculated by averaging. This innovative approach suggests that the advantages of ensemble models can be harnessed without incurring additional training costs. Finally, the team's ultimate submission comprised a combination of independently trained models and models derived from earlier training epochs.

Table 4.4: Combination of data features used by the participants (in order of the leaderboard ranks). The table illustrates the different strategies used by the models. It highlights a unique approach adopted by the two top-performing models: each of these employed a separate model for each variable, a distinct departure from the typical approach taken by other models, which encapsulated all variables into a single, unified model. Interestingly, the third and fifth-ranked models achieved additional performance improvements through the use of data augmentation techniques, a strategy that was not explored by the top two models. The term 'both' is used in the table to indicate models that compared results with and without a specific feature. Although not explicitly shown in the table, it was unanimously recognized that a crucial factor behind the individual improvements of each model was the strategy of training all regions within a single model. This approach marked a significant departure from the baseline, which utilized a distinct model for each region.

	Static Data	**Extra variables**	**Model per variable**	**Data augmentation**
antifuge	no	no	**yes**	no
sungbinchoi	no	**yes**	**yes**	no
Michael Fish Forecasting	**yes**	**yes**	no	**yes**
moto	n/a	n/a	n/a	n/a
ai4ex	**both**	**both**	no	**yes**
U-Net baselines	**yes**	no	**yes**	no

4.7 Discussion

This thesis chapter presented a novel benchmark for weather prediction, known as the Weather4cast, which treated weather forecasting as a video frame prediction task. It is composed of a high-resolution, multi-channel dataset collected from a constellation of geostationary satellites to present two main challenges: the Core Challenge and the Spatial Transfer Learning Challenge.

In the Core Challenge, models were tasked to predict four weather parameters—temperature, convective rainfall rate, probability of occurrence of tropopause folding, and the cloud mask—in selected regions with comprehensive training data. The Spatial Transfer Learning Challenge, on the other hand, pushed models to extrapolate their predictions to unseen regions without training or evaluation data. Both tasks required models to forecast the next 8 hours of weather based on the preceding 1 hour of weather data.

We successfully evaluated the value introduced by our benchmark through the Weather4cast competition, which served as a proving ground for various machine learning models, highlighting the effectiveness of U-Net-like architectures and re-

current convolutional models for this task. Notably, the winning model of the competition was a recurrent convolutional model incorporating residual units, demonstrating a high level of performance in predicting multi-channel weather patterns for an eight-hour period.

Interestingly, the competition's results suggested that the ability to generalize to new spatial locations largely depends on the extent to which weather patterns are accurately learned in regions with abundant training data. This indicates that capturing the key dynamics of weather phenomena in familiar territories is crucial to enabling robust forecasting in unfamiliar regions.

The competition also highlighted the potential benefits of data diversity over mere volume, as the inclusion of training data from new regions appeared to significantly improve model performance, especially in the Spatial Transfer Learning Challenge. The exploitation of pre-trained models was also recognized as an effective strategy for leveraging additional data.

The use of ensemble models was another significant trend among the top submissions. All the top three submissions employed model ensembles, which consistently outperformed single models. Moreover, it was demonstrated that ensemble models could be harnessed without incurring additional training costs by using weights from different epochs of training a single model.

In conclusion, the Weather4cast competition benchmark has proven instrumental in enhancing machine learning models to handle robust spatio-temporal predictions, thereby advancing the field of weather forecasting. This confluence not only contributes to the development of more efficient algorithms but also promotes innovative approaches to understanding and predicting complex weather patterns. Thus, this endeavor is dual-purpose, bolstering machine learning methodologies while also refining the accuracy and reliability of weather prediction - a synergy that promises to drive further advancements in both domains.

The related competition and its leaderboards continue to be available and open for new submissions on the Weather4cast website [6].

4.8 Author Contributions

The work encapsulated in this chapter reflects a collaborative effort involving various contributors whose inputs were invaluable in bringing the project to fruition.

[6] *Core Challenge leaderboard* www.iarai.ac.at/weather4cast/competitions/ieee-big-data-core-final *Spatial Transfer Learning Challenge Leaderboard* www.iarai.ac.at/weather4cast/competitions/ieee-big-data-transfer-learning-final

The novel Weather4cast benchmark, a cornerstone of this research, was the original concept of the author, P. Herruzo. Herruzo was also responsible for initiating and coordinating contact with data providers AEMet, and transforming raw data into the final distributed dataset.

Herruzo also invented, trained and validated all the deep learning baselines provided in the benchmark.

Moreover, Herruzo spearheaded the creation of the scientific competition content, crafting the proposal drafts for CIKM and IEEE Big Data competitions, and developing the software for the Weather4cast competition and for automatic leaderboard evaluation. Herruzo performed the initial analysis and provided insights into the models, and was responsible for drafting the manuscript of the published article.

In his analysis, Herruzo identified a one-to-one correlation between the performance of models in familiar and new areas. This observation underscores the crucial role of data quantity and quality in the successful generalization of models.

The engagement with IARAI team members A. Gruca, S. Hochreiter, M. Kopp, and D.P. Kreil was crucial in refining the project's focus and improving its execution. They contributed substantially to the selection of benchmark variables, provided insights on the competition's goals, had invaluable input on the metrics, and offered critical feedback and editing suggestions for the articles and competition proposals.

The collaboration with AEMet (L. Lliso, X. Calbet, P. Rípodas) was instrumental to this research. They supplied the raw data essential for the competition, shared their meteorological expertise to help define the benchmark's evaluation process, and contributed constructive criticism and refinements to the articles and competition proposals.

While the majority of the work was carried out by P. Herruzo, the collective intellectual and practical contributions of the mentioned individuals and institutions significantly enhanced the study's overall quality, depth, and impact. During this work's publication, P. Herruzo was a member of IARAI.

Chapter 5

Weather4cast at at NeurIPS 2022: Super-Resolution Rain Movie Prediction under Spatio-temporal Shifts

This chapter is based on materials from the following peer-reviewed paper:

[PENDING] Weather4cast at NeurIPS 2022: Super-Resolution Rain Movie Prediction under Spatio-temporal Shifts. In NeurIPS 2022 Competition and Demonstration Track. PMLR.

5.1 Introduction

In the preceding chapter, we introduced Weather4cast 2021, a new benchmark constituting a host of unique challenges, an exceptional dataset, and pioneering deep learning models. These models achieve state-of-the-art performance in multimodal weather forecasting with high spatio-temporal resolution. Additionally, we affirmed the scientific community's interest in this benchmark by successfully orchestrating scientific competitions and publishing the results in peer-reviewed papers.

The assembly of data, comprehensive study of related literature, model training, and delineation of significant challenges and metrics were made feasible through the joint effort of a diverse team. This collaborative group consisted of deep learning scientists, meteorologists, infrastructure experts, and data scientists. Renowned entities such as the Spanish Meteorology Agency (AEMet) and the Institute of Ad-

vanced Research in Artificial Intelligence (IARAI) were integral contributors to this committee.

Our meticulous investigation brought to light an unaddressed void within the scientific community that holds the potential to significantly impact both meteorology and machine learning. Presently, there are no efficient models capable of **predicting terrestrial precipitation using only satellite-observed spectral radiances**. Surface-level weather radars are sporadic, with complete non-existence in some developing nations. However, these regions are within the purview of satellites. Therefore, the development of precise models utilizing data gathered from space to predict surface-level precipitation could aid in preparing for and mitigating severe rainfall across the globe.

This area of research is of particular interest to both the European Space Agency (ESA) and the OPERA Network, which supervises and maintains the network of weather sensors across Europe. In this pivotal chapter, we align ourselves with these international agencies to address this grand scientific challenge, one with clear, immediate practical implications.

5.2 Related Work

In the realm of meteorological forecasting, Convolutional Neural Networks, as implemented in MetNet Sønderby et al. (2020) and MetNet-2 Espeholt et al. (2021), have been instrumental in enhancing physical models for both short-term (4 hour) and medium-term (12 hour) predictions. In fact, in a comparison, meteorological experts exhibited an 89% preference for the predictions generated by a deep generative model Ravuri et al. (2021a). Despite these advances, there remains a challenge concerning access to the high-resolution data essential for training such models.

Contemporary model architectures such as Graph Neural Networks Lam et al. (2022), Transformers Bi et al. (2022), and U-Nets Kaparakis and Mehrkanoon (2023), have primarily been constrained to publicly available resources. A key example includes the ECMWF ERA5 reanalysis archive Hersbach et al. (2020), which offers a range of variables at multiple vertical levels across the globe, albeit at lower resolutions (1,h, $\sim$ 30,km). These limitations align with those encountered in other common datasets in the domain Rasp et al. (2020), Schroeder de Witt et al. (2021).

The CloudCast dataset Nielsen et al. (2021) represents an advancement, providing 10 distinct cloud-related variables at 15 min intervals and $\sim$ 4 km resolution.

We have further enriched this resource by including 22 additional, more generalized multimodal variables in the Weather4cast 2021 dataset and benchmark Herruzo et al. (2021b).

The SEVIR dataset Veillette et al. (2020a), offering both satellite and high-resolution radar data, is unfortunately confined to the United States. While it has been employed in learning satellite-to-radar translation (at the same timestamp) and radar-to-radar prediction, its utility for satellite-to-radar prediction remains untapped. Recently, Generative Adversarial Networks have leveraged this dataset to refine U-Net predictions Hu et al. (2022).

As of our current understanding, the Weather4cast 2022 dataset represents a pioneering effort by integrating raw spectral bands from satellite sensors and ground-based high-resolution precipitation radar data across diverse geographical regions and varied time frames. This inclusion facilitates the establishment of **the first-ever satellite-to-radar forecasting benchmark**.

5.3 The Weather4cast 2022 Formulation

5.3.1 Datasets

The dataset is primarily assembled from two primary data sources: spectral bands derived from satellite observations and ground-based high-resolution precipitation radar. Both of these datasets are encoded as time-sequential image arrays, allowing for clear visualization and analysis of meteorological patterns over time.

To augment this data and provide additional context, we have included topographical information regarding the elevation of the terrain for each respective data point. This is crucial for understanding the impacts of geographical features on local weather patterns. Furthermore, we have provided geographical coordinates, specifically latitude and longitude, for each pixel. This granular localization information allows for precise mapping and identification of weather phenomena, thereby enhancing the applicability and relevance of our findings in practical meteorological forecasting.

5.3.1.1 Meteosat Second Generation SEVIRI Data

The Meteosat Second Generation (MSG) series of geostationary meteorological satellites are managed by the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT). Each satellite in this series is equipped with the

Table 5.1: Key characteristics of the SEVIRI instrument on board the Meteosat Second Generation (MSG) satellites from EUMETSAT. The table presents important details regarding the spectral channels, spatial resolution, and specific spectral characteristics associated with each channel.

Channel Number	Central Wavelength (μm)	Spatial Resolution (km)	Spectral Zone Characteristic	Type of Channel
1	0.635	3	Solar Visible	Window (VIS)
2	0.81	3	Solar Visible	Window (VIS)
3	1.64	3	Solar Infrared	Window (VIS)
4	3.90	3	Solar/Thermal Infrared	Window (VIS/IR)
5	6.25	3	Thermal Infrared	H_2O Absorption (WV)
6	7.35	3	Thermal Infrared	H_2O Absorption (WV)
7	8.70	3	Thermal Infrared	Window (IR)
8	9.66	3	Thermal Infrared	O_3 Absorption (IR)
9	10.80	3	Thermal Infrared	Window (IR)
10	12.00	3	Thermal Infrared	Window (IR)
11	13.40	3	Thermal Infrated	CO_2 Absorption (IR)
12	Broad Band (0.4–1.1)	1	Visible/Infrared Solar	Window (VIS)

Spinning Enhanced Visible Infra-Red Imager (SEVIRI) instrument Shcmetz et al. (2002). This sophisticated device boasts twelve channels that continuously monitor the Earth across the visible and near-infrared (VIS), thermal infrared (IR), and water vapor absorption band (WV) spectrums.

Each of the eleven channels equipped with narrow spectral band filters offers a spatial resolution of approximately 3 km at nadir (the point on the ground that is directly below an observer or sensor in space). A detailed breakdown of the specific characteristics of each spectral channel can be found in Table 5.1. Given the geostationary orbit of the MSG satellites, they maintain a consistent viewpoint over the Earth's disk from a fixed location along the celestial equator. This positioning allows the SEVIRI instrument to generate images of 3712×3712 pixels for each of the eleven channels every 15 minutes, in its nominal mode. The data utilized in this research pertains to the MSG satellite positioned at zero degrees longitude, operating in nominal mode.

To visually demonstrate the capabilities of the SEVIRI instrument and the richness of the data it captures, Fig. 5.1 presents an 'Air mass' RGB composite image generated by this satellite. This image exemplifies the high-resolution, multispectral data collected by the MSG satellites and the potential insights this data offers for advancing our understanding of atmospheric dynamics and improving weather forecasting models.

In collaboration with EUMETSAT, we introduced a modest amount of noise to the original Meteosat sensor data. This modification enabled us to re-distribute the assembled dataset, with multiplicative noise deemed most appropriate given the

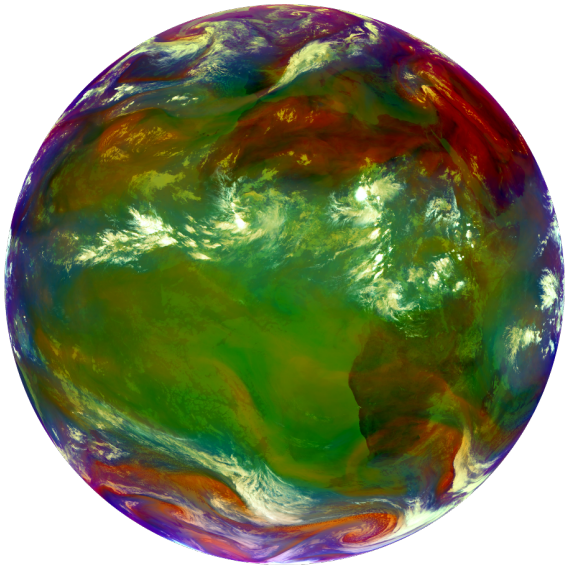

Figure 5.1: The 'Airmass' RGB composite image, crafted using a blend of data from four distinct channels (5, 6, 8, and 9) of the SEVIRI instrument, was sourced from the MSG satellite stationed at zero degrees longitude. This representative image was captured on August 20, 2019, at 10:00 Coordinated Universal Time (UTC).

.

range of data under review. Our sensitivity analyses confirmed that this addition had an inconsequential effect on the baseline predictions.

5.3.1.2 Weather Radar Data from the OPERA Project

Weather radars are a fundamental tool in the measurement of precipitation due to their ability to cover large areas, provide a three-dimensional structure of precipitation systems, and track these systems over time. When combined to form a network, these radars can monitor an even larger domain.

However, it's worth noting that radar precipitation measurements are subject to various sources of error. These include challenges associated with beam broadening and the increased distance to the Earth's surface as the distance from the radar site increases. Additional errors may arise from echoes originating from non-meteorological targets, beam blockage due to terrain (such as mountains), signal attenuation by rain (especially heavy rain), and the anomalous propagation of the beam under certain atmospheric conditions. A comprehensive discussion of the advantages and disadvantages of using weather radar for precipitation measurement, as compared to other sources of precipitation data, as well as an overview of the current state of the art in radar research, can be found in Sokol et al. (2021).

Despite these limitations, the distinctive characteristics of precipitation radar make a radar network an optimal choice for meteorological services conducting nowcasting and issuing weather warnings. Often, radar data is augmented with information from other sources like rain gauges and satellite data. In the context of this study, radar data serves as the reference point and is treated as the "ground truth" for the precipitation field.

The radar data used in this work are 2D composites sourced from the Operational Programme for the Exchange of Weather Radar Information (OPERA), a project run by EUMETNET (`www.eumetnet.eu`). OPERA generates 2D composites of instantaneous surface rain rates, maximum reflectivity, and one-hour rainfall accumulations. Specifically, for this study, we utilized the instantaneous rain rate composites provided every 15 minutes from February 2019 to 2021.

Radar reflectivity was converted into precipitation intensity in the 2D composites using the Marshall-Palmer Z–R relationship, employing coefficients $a = 200$ and $b = 1.6$ Marshall et al. (1947). Further details about the OPERA project can be found in Huuskonen et al. (2014) and Saltikoff et al. (2019a).

5.3.2 Data Compilation and Harmonization

The compilation of data for this study required careful consideration and manipulation due to the distinct geographical projections of our primary data sources: the OPERA radar network data and the Meteosat Second Generation (MSG) SEVIRI data.

OPERA radar data is structured in the Lambert Azimuthal Equal Area projection, with each pixel representing an area of 2000×2000 meters. This projection type is chosen for its property of preserving the area relative to the Earth's surface. Conversely, MSG data utilizes a geostationary projection, where the pixel size grows larger as the distance from the sub-satellite point increases. The area covered by an MSG pixel scales from 3000×3000 meters at the sub-satellite point to irregularly shaped pixels with a side size exceeding 24 km, as seen over locations like Iceland.

The differing formats posed a challenge in the integration of the datasets. For the convenience of model training, we performed a transformation of the OPERA data into a geostationary grid, allowing it to match the geographical layout of the MSG data. The restructured OPERA and MSG data, both rendered as 2D images, could then be seamlessly combined.

During the reprojection process, however, certain informational discrepancies may arise, primarily due to the variable and larger size of MSG pixels compared to those of OPERA. To address this issue, we elected a dense destination grid in which each MSG pixel was subdivided into 36 smaller units, with each side of the satellite pixel divided by a factor of 6.

Figure 5.2 exemplifies this reprojection process, showing a scene of approximately 30×30 km near Amiens, France. This figure illustrates the MSG grid, the reconfigured OPERA grid, and the result of reprojecting the OPERA pixels onto the final grid (highlighted in cyan).

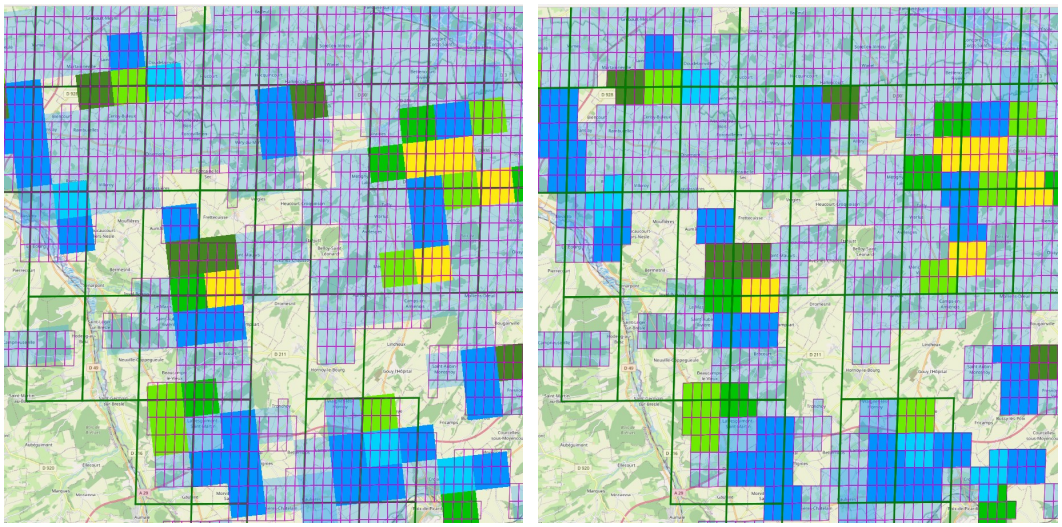

Figure 5.2: Diagram Illustrating the Reprojection Process: The illustration delineates the transformation of data from the OPERA radar network to a geostationary grid to align with the MSG SEVIRI data. The green lines represent the boundaries of the original MSG pixels, while the magenta lines mark out the smaller destination grid cells into which the OPERA data is reprojected. The colored squares represent the original OPERA pixels, each one holding specific precipitation data. The reprojection process ensures that both datasets are geographically matched, enabling them to be more effectively combined for analysis. The left image presents the original location of the OPERA data, while the right image displays the OPERA data post-reprojection, now situated in its newly adapted geostationary projection.

The dimension of the dense grid was selected based on an evaluation of the information loss occurring in a forward and backward reprojection of the OPERA data. The chosen 6×6 grid exhibited negligible information loss, even in the most unfavorable areas.

Notably, in the dataset provided for the competition, the MSG pixels are not further divided into 6×6 smaller units. Instead, each MSG pixel corresponds to

6×6 reprojected OPERA pixels. This distinction ensures consistency and usability within the competition dataset.

5.3.3 Geographical Context and Rainfall Variability: Criteria for Region Selection

Precipitation variability across Europe is largely influenced by seasonality, as highlighted by Zveryaev Zveryaev (2004). However, this variability is additionally modulated by an array of local and distant climate patterns as investigated by Karagiannidis Karagiannidis et al. (2008). Complicating the matter further, the increasing global warming, driven by human activities, is anticipated to intensify rainfall extremes across the continent, a point stressed by King King and Karoly (2017). Hence, the prediction of high-impact weather events becomes increasingly crucial.

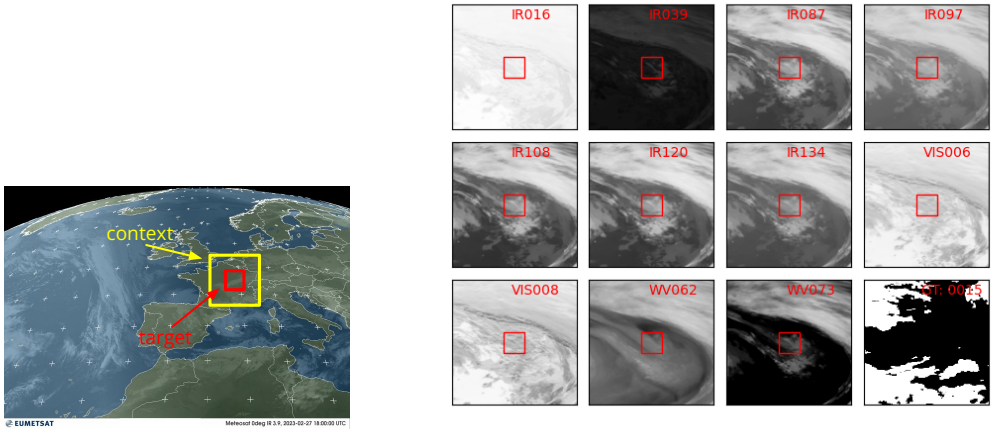

Figure 5.3: The diagram on the left illustrates the spatial context (highlighted in yellow), wherein satellite radiances are provided, and the target region (highlighted in red), which is the focus for rainfall predictions. The diagram on the right provides a snapshot of longitude-latitude maps for the eleven MSG band radiances for the context of patch 15, along with the OPERA binary mask ground truth (GT) using a threshold of 0.2 mm/hour (displayed at the bottom right). Within the MSG images, darker areas indicate lower values, while black represents rain in the OPERA image.

In our investigation, for a defined area of interest (with the size 252×252 as per the OPERA data reprojected to geostationary pixels), we offer a wider spatial context. This context area encloses a square region six times larger than the area of interest, rendered in the same geostationary projection, as illustrated in Figure 5.3

(left). This relative size of the context is determined to encapsulate nearby weather systems which could potentially affect the target patch with rainfall in subsequent hours. Figure 5.3 (right) offers a comparative visual representation of satellite radiances and the binary rainfall mask for OPERA at a given time point. This figure underscores how the different bands provide diverse perspectives of the same scene and how the binary mask is not directly correlated with the radiance patterns.

Before selecting the target patches for the competition, we aimed to categorize them based on the monthly frequencies of rainfall events. Our objective in determining the target patches was specifically to include those regions where rain events are not only significant but also typically infrequent. The aim of this selection criteria was to ensure our dataset would adequately represent less frequent rain events and offer a broad spectrum of precipitation patterns for examination.

5.3.3.1 Selection and Characterization of Target Patches

Our selection process for the target patches was guided by a dual desire: to encapsulate a wide range of rain events across the entire spectrum of rain rates, and to ensure that even the relatively infrequent, yet intense rain events were adequately represented in our study.

To categorize the rain rates, we adhered to the standard meteorological classification as specified by the (WMO) World Meteorological Organization (2018). This classification divides rain rates into four distinct categories: no rain (0-0.1 mm/hour), low rain (0.1-2.5 mm/hour), moderate rain (2.5-7 mm/hour), and heavy rain (greater than 7 mm/hour). This approach is congruent with the methodologies followed by other research in the field, such as that by Ravuri et al. (2021b).

Armed with this categorical breakdown, we proceeded to calculate the monthly frequencies of rain events for each category, within every potential region of interest. We considered data spanning from February 2019 until December 2021 for this analysis. We meticulously evaluated these frequencies, with special attention to regions that reported a higher frequency of rain events. Interestingly, we observed that the moderate and heavy rain events were considerably rarer and their monthly frequencies exhibited high volatility, reflecting the inherent seasonal patterns of precipitation.

To determine the final selection of target patches for the competition, we sought a balance. We wanted to include regions prone to frequent but less intense (low) rainfall, alongside regions that experienced less frequent but more intense rainfall.

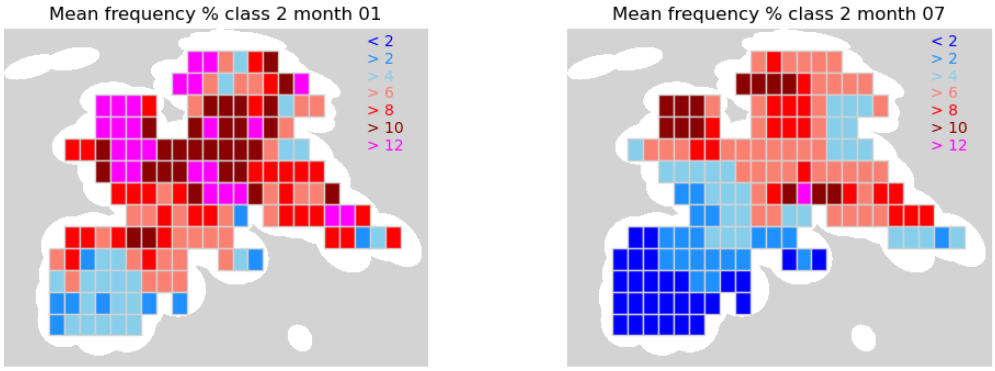

Figure 5.4: These charts illustrate the probability maps (%) for low rain rates, as recorded by the OPERA network between 2019 and 2021, for the months of January (left) and July (right). The values are displayed for square areas of the same size as the prediction outputs. Areas outside the OPERA coverage are shaded in grey.

This balance was key to ensuring a representative dataset that offered a comprehensive overview of different rainfall scenarios across the European domain.

Figure 5.4 underscores the geographical and temporal variability of low rain rate occurrences across the OPERA domain. It highlights the differences in rainfall patterns between the boreal winter and summer seasons.

During winter, a higher probability of rain (>12%) is evident across the British Isles, Northern Europe, the Alps, and Eastern Europe. Conversely, the summer months reveal a dominance of dry conditions across southern Europe, barring the Alpine region. Rain events also decrease in frequency across higher latitudes during this period. This contrast reaffirms the significant seasonal variability within the European domain Karagiannidis et al. (2008). It also shows the imbalance between rain and no-rain events, especially in southern regions.

This extensive analysis culminated in the selection of several target patches for both the core and the transfer learning challenges, which formed the data provided to competition participants. The locations of these selected regions are shown in Figure 5.5.

Through this rigorous process of characterization and selection, we ensured that our dataset would be able to capture the broad spectrum of rain events across Europe, while providing a challenging benchmark for the development of robust and accurate rainfall prediction models.

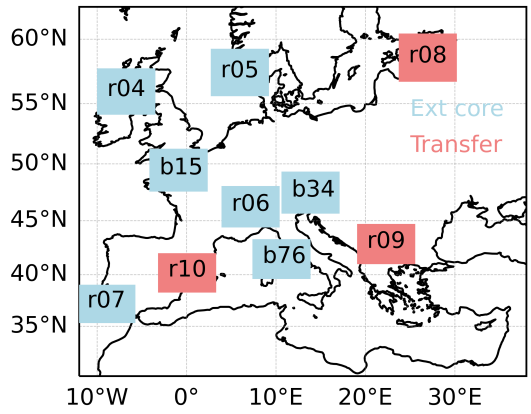

Figure 5.5: This map displays the locations of the competition regions across Europe. The Core regions used in Stages 1 and 2 are denoted by a 'b' and shown in blue. The Extended Core regions, used only in Stage 2, are denoted by an 'r'. The Transfer learning regions are indicated in red.

5.3.4 Tasks

The benchmark task focuses on the prediction of rainfall events over a future period of 8 hours, which is broken down into 32 individual time slots. Each time slot corresponds to a 15-minute increment. The predictive model is to be fed with data from the preceding hour, consisting of four 15-minute time slots, each represented by 11-band spectral satellite images. These spectral bands encompass the visible light (VIS), water vapor (WV), and infrared (IR) ranges.

The region of interest for the predictive model is confined to an area of 42×42 pixels within the satellite imagery. This specific area is enveloped by a larger spatial context represented by a square area of 252×252 pixels. This setup implies a dual-task for the predictive model. Firstly, the model needs to accurately forecast the weather conditions, specifically rainfall events, over the ensuing 8 hours. Secondly, it has to perform a "super-resolution" operation, in which the lower-resolution satellite data (one pixel in the satellite imagery corresponds to six pixels in the output) must be transformed into a higher-resolution prediction, equivalent to ground-radar reflectivity.

In addition to the satellite imagery, the model also utilizes static geographical information about the area of interest. These static variables include the latitude,

longitude, and elevation. To emulate the inherent stochasticity and unpredictability of meteorological phenomena, a slight random noise is incorporated into the 11 spectral bands of the satellite images.

The task essentially sets a benchmark for machine learning models in the domain of meteorological forecasting, particularly in the aspects of time-series prediction, super-resolution image processing, and working with both dynamic and static input data.

5.3.5 Evaluation Metric

In the context of the benchmark designed for predicting precipitation events, the evaluation metric needs to account for the core challenge of predicting rain occurrences, rather than the precise amount of rainfall. This specificity is based on the nature of the dataset and the fact that rainfall events are relatively rare. To this end, we have chosen to use the Intersection over Union (IoU), a metric frequently used in the field of computer vision for assessing the accuracy of models involved in object detection and segmentation tasks.

IoU is calculated by taking the ratio of the area of overlap between the predicted and actual rainfall areas to the area of union of these two. This metric is specifically tailored to target the accurate prediction of rainfall events by focusing on the correct prediction of rain-affected pixels.

To generate a single evaluative value for each model's prediction, we calculated IoU values for each region separately and then computed the average of these values across all regions. This process provides a comprehensive overview of the model's performance over a diverse range of regions.

The calculation of the final metric involves the elimination of certain data points from each region. Specifically, pixels outside of the OPERA coverage or those with missing data, represented by –9,999,000 and –8,888,000 values respectively, were removed prior to computation. The former represents instances where the region of interest extends over a sea area, thus exceeding the coverage of ground radar. The latter is associated with errors that may occur during the collection of OPERA radiance data.

Additionally, in order to avoid the misidentification of cluttering echoes and artefacts, and to account for the satellite's limitations in detecting precipitation, a threshold of 0.2 mm/h was established for rainfall detection. This amendment further unbalances the dataset, decreasing the frequency of rain events and thereby amplifying the challenge of the prediction task.

5.4 Associated Competitions and Provided Software

The weather forecasting benchmark described earlier offers a significant stepping stone for developing advanced machine learning models capable of accurately predicting meteorological events. It sets the stage for a unique blend of tasks: time-series prediction, super-resolution image processing, and the fusion of dynamic and static data. However, the true capabilities of a benchmark are realized only when it is put to the test in a real-world context. For this reason, the benchmark was transformed into a competition format. Organizing a competition enables the evaluation of the models' performances against each other and allows for the exploration of novel, innovative approaches that may not have been previously considered. The associated competition was accepted for inclusion in the highly competitive NeurIPS 2022 conference, reflecting the importance and relevance of this domain.

5.4.1 Weather4cast 2022 at NeurIPS

Building upon the theoretical foundation of the weather forecasting benchmark, the Weather4cast 2022 competition was designed to encourage practical advancements in this field. The competition was conducted in two stages.

During the first stage, which ran from August 1 to November 18, 2022, participants were given access to 2019 data for three regions. The purpose of this stage was to kickstart the model development process and to test the performance of the models in a relatively controlled setting. To provide participants with a sense of their models' standing, a public leaderboard was updated regularly with the performance of each model against the competition baseline and other submissions.

The second stage of the competition commenced on October 14, 2022, and concluded on November 18, 2022. This stage introduced two additional challenges, which further stretched the capabilities of the participating models. The Extended Core Challenge expanded the geographical scope by providing 2020 data for the initial three regions and four new regions. The Transfer Learning Challenge tested the adaptability and generalizability of the models by providing 2021 data for all seven regions and introducing three additional unknown regions. Public leaderboards were maintained throughout this stage to continue providing participants with comparative feedback.

The final evaluation of the models was performed on undisclosed data. This approach ensured an objective assessment of each model's predictive capabilities and its potential for application in real-world scenarios.

5.4.2 Weather2cast 2022 Software

To assist participants in the competition and to provide a foundation upon which they could build their models, the Weather2cast 2022 software was provided [1]. This starter kit contained essential code for data loading and exploration, as well as a baseline model for initial training.

The baseline model was a modified version of a 3D variant of the U-Net [2], a renowned architecture for image segmentation tasks. This served as a point of departure for the participants, who were encouraged to modify and improve upon this baseline in their submissions. See table 5.2 for details about the hardware and tools that we used.

Table 5.2: Machine and tools used for the Weather4cast 2022 benchmark.

GPU	8x Tesla V100-PCIE-32GB
CPU	2x Intel(R) Xeon(R) Gold 6146 CPU @ 3.20 GHz
Cores/Threads	12 cores / 24 threads (per CPU)
RAM	754Gi (Samsung DDR4-2600)
Tools	PyTorch, PyTorch Lightning, NumPy

By furnishing participants with these resources, the competition aimed to lower the barrier to entry, particularly for those who might be new to the field or who might not have had previous experience with similar tasks. It also helped ensure that all participants had a fair and equal start, fostering a spirit of innovation and advancement.

5.5 Results

Our competition attracted significant global participation, with thirty teams from over a dozen countries submitting more than 1,600 entries. Impressively, over 900 of these submissions outperformed the benchmark set by a simplistic 3D U-Net model. The first pilot stage garnered the majority of these submissions, while the

[1] Weather4Cast starter toolkit: `https://github.com/iarai/weather4cast-2022`

[2] ELEKTRONN3 - Neural Network Toolkit: `https://elektronn3.readthedocs.io/en/latest/`

more extensive main competition brought in nearly half as many entries. The dedicated leaderboard for transfer learning received around 100 submissions, with approximately 60 models being assessed on the held-out test datasets.

Teams participating in the competition were encouraged to prepare short research papers that could be presented at the NeurIPS conference. More details can be found on the competition's website at `www.weather4cast.org`. Notable achievements and interesting models from the competition are described bellow.

WeatherFusionNet developed by team FIT-CTU, utilizes three distinct modules to estimate rainfall from satellite data. The key idea is to fuse the outputs of the modules using a U-Net architecture for final prediction Pihrt et al. (2022).

Model Ensemble for Probabilistic Rain Prediction by team meteoai, leverages ensemble method for predicting rain, incorporating preprocessing strategies and a focus on loss functions to optimize performance Li et al. (2022).

Team *team-name* developed a method using *Vision Transformers* with configurations to enhance various model performances and baseline-specific improvements Belousov et al. (2022).

The *SImple baseline for weather forecasting using spatiotemporal context Aggregation Network (SIANet)* proposed by team SI-Analytics, is an end-to-end model based solely on CNNs. SIANet achieved high standings in all stages of the competition through the implementation of innovative strategies Seo et al. (2022b).

RainUnet, developed by team KAIST-CILAB, is a hierarchical U-shaped network utilizing a Temporal-wise Separable block to capture interframe correlations. They further enhance the model's performance through preprocessing strategies and context-specific prediction Park et al. (2022).

Lastly, the *Region-Conditioned Orthogonal 3D U-Net* by team KAIST AI is a modified 3D U-Net architecture incorporating region information during propagation. Several training strategies were also employed to further enhance the performance Kim et al. (2022).

5.6 Discussion

In Table 5.3, we highlight the average Intersection over Union (IoU) scores for the top-performing teams and compare the characteristics of their winning models with a our baseline model.

Most high-performing models utilized dedicated preprocessing steps, leveraging either Earth observation domain knowledge or standard machine learning

Table 5.3: Top ranked teams and key features. **ex-aequo*

Rank	Team	avg IoU	Preprocess	Ensemble	Physics-based	Transformer
1	FIT-CTU	.316	✓	×	✓	×
2	meteoai	.307	✓	✓	×	✓
3*	SI Analytics	.305	✓	×	✓	×
3*	TEAM-NAME	.300	×	✓	×	✓
4	KAIST-CILAB	.287	✓	×	×	×
5	KAIST-AI	.274	✓	×	×	×
-	Baseline	.254	×	×	×	×

techniques. For instance, standard data augmentation was used to counter the significant imbalance between the rain and no-rain classes. Domain knowledge was also exploited for application-specific data augmentation, such as the estimation of physical properties like wind speed from cloud movement as additional model input.

Certain satellite bands from the Visible (VIS) or Water Vapor (WV) channels were discarded based on domain knowledge, thereby reducing the input data and improving computational efficiency.

Team *meteoai* Li et al. (2022) reported that incorporating static information, such as geographical coordinates and elevation, did not significantly enhance model performance. Other top-ranked teams did not report similar trials, suggesting that the role of static information in the dynamics of weather patterns remains an open research question.

Several top-ranking models prominently featured state-of-the-art machine learning techniques. Transformers were utilized for spatio-temporal modelling in the second and third best-ranking solutions. Additionally, ensemble models known for their robustness and efficiency were also effective. Meanwhile, incremental improvements to the baseline model were sufficient to elevate prediction performance and rank highly. For instance, team *KAIST AI* Kim et al. (2022) showcased the effectiveness of Feature-wise Linear Modulation (FiLM) layers Perez et al. (2018), which modify the output of neural networks by applying an affine transformation to intermediate features.

Interestingly, the winners of the core and transfer learning challenges differed. Most teams applied the same model to both challenges. However, team *SI-Analytics* took a distinct approach for the transfer learning challenge by applying geometric data augmentation and test-time ensemble models with a spatio-temporal smoother loss, earning the first position on the leaderboard and the special Transfer Learning Award by the Scientific Committee Seo et al. (2022a).

Models that incorporated physics knowledge delivered the best and third-best results in the core prediction challenge and the best performance in the transfer-learning challenge. Therefore, the development of models that combine the strengths of machine learning and application domain knowledge seems promising for future research.

In summary, this chapter introduces a significant new benchmark, established in the context of a successful competition held at NeurIPS. The competition fostered a diverse set of innovative approaches to tackle the challenges of weather prediction, demonstrating the feasibility and value of community-based efforts in advancing scientific understanding in this complex field. This newly introduced benchmark provides a robust reference point for future investigations, catalyzing further advances in weather forecasting. Looking ahead, we anticipate this resource will continue to stimulate new research, drive methodological innovations, and contribute substantially to our growing body of knowledge in the domain of meteorological prediction.

5.7 Author Contributions

This chapter heralds a remarkable collaborative achievement, realized through the joint efforts of various entities and individuals, including UPC, IARAI, AEMet, and the European Space Agency's Phi-Lab.

P. Herruzo, the author, played a crucial role in the creation of the innovative Weather4cast 2022 benchmark, which centers around satellite to radar data. Herruzo was instrumental in fostering partnerships with data providers, including AEMet, and led the project's multi-party coordination until his departure on May 23, 2022.

Throughout his involvement, Herruzo made substantial contributions to the project's pillars, including the transformation of raw data into an initial dataset and the construction of original deep learning baseline models. Moreover, he was at the helm of the associated scientific competition, drafting the NeurIPS 2022 competition proposal and laying the groundwork for the Weather4cast 2022 competition software.

Herruzo played a vital role in conducting initial analyses, offering insights into the baseline models. He also authored the literature review section of the published article. He provided extensive feedback and meticulous polishing of the final manuscript.

Despite leaving the project, Herruzo acknowledges the importance of the additional refinements made in his absence. These included defining the competition's regions, evaluating competitor models, and managing the competition. These post-departure adjustments were crucial to the project's successful completion. Nevertheless, the fundamental scaffolding of Herruzo's work continues to be integral to the project's eventual success.

Chapter 6

Research Findings and Conclusions

In this final chapter, we revisit the original research questions posed at the beginning of this thesis and provide a concise summary of their answers. We invite readers to refer to the specific chapters for more detailed insights and findings. Subsequently, we present the main conclusions derived from our research endeavors and discuss potential avenues for future work and research.

6.1 Research Question Analysis and Findings

In this section, we present a comprehensive analysis of the original research questions and offer a concise overview of the key findings discovered throughout this thesis.

Q1: How can traffic data be optimally encoded to exploit the capabilities of deep learning algorithms for forecasting?

In this thesis, our primary focus lies in exploring the realm of multimodality in traffic forecasting. We recognize that traffic patterns are influenced by various factors and, therefore, aim to leverage multiple data sources to enhance prediction accuracy. Specifically, we incorporate additional data such as weather information and seasonal patterns to capture the complexity of traffic dynamics. Our research demonstrates the encoding of traffic data as images and the integration of weather information with different spatiotemporal resolutions into deep neural networks. We found that combining these data sources at different stages of the network allows us to effectively capture the interplay between traffic and weather factors, even when the data sources have different resolutions. This approach enables us to uncover valuable insights and improve the accuracy of traffic forecasting models.

Q2: How can deep learning be leveraged to accurately predict traffic patterns and conditions?

We found that Convolutional neural networks (CNNs), originally developed for image processing, can be effectively applied to traffic data when it is encoded as images. By employing CNNs, we are able to capture spatial dependencies within the traffic data, allowing the model to learn meaningful representations of traffic patterns and extract relevant features for accurate prediction. To account for the temporal dynamics of traffic, we also utilize recurrent neural networks (RNNs) to capture temporal dependencies and capture the evolving nature of traffic over time. Furthermore, we highlight the importance of multimodal data integration by incorporating additional data sources, such as weather information and seasonal patterns, at different spatio-temporal resolutions. This approach enhances the prediction capabilities of the models by concatenating the multimodal information at various stages of the network. Our findings highlight the potential of leveraging deep learning techniques for traffic forecasting tasks, enabling more informed transportation management strategies.

Q3: What are the best practices for encoding weather data for deep learning applications, and which models yield superior performance in weather forecasting?

Our research shows that carefully selecting and preprocessing weather variables such as temperature, precipitation, and cloud cover, we can create input representations that capture the relevant spatio-temporal information into images. The encoding of weather data as sequences of images has allowed us to exploit the strengths of deep learning models in capturing complex weather patterns and making accurate predictions. Furthermore, we have extensively evaluated different deep learning architectures and model configurations to identify those that yield superior performance in weather forecasting. The winner models seem to be those that incorporate convolutional and recurrent neural networks, which are able to capture both spatial and temporal dependencies in the data. Interestingly, the same type of architecture has been found to be effective in traffic forecasting, suggesting that these models may be well-suited for spatio-temporal prediction tasks in general.

Through our investigation, we have established a set of recommended practices for encoding weather data and identified the deep learning models that demonstrate the highest predictive capabilities.

Q4: What are the fundamental components —including tasks, metrics, baselines, and evaluation methodologies— necessary to assess the generalization

and adaptability of deep learning models in the context of spatio-temporal forecasting?

In order to assess the generalization and adaptability of deep learning models in spatio-temporal forecasting, we have developed fundamental components that form the basis for comprehensive evaluations. These components include specific forecasting tasks, relevant evaluation metrics, viable baseline models, and rigorous evaluation methodologies.

We found that by defining well-defined tasks and associated metrics, we can objectively measure the performance of deep learning models in different forecasting scenarios. This involves defining datasets from locations and/or time periods that are not represented in the training data, thereby challenging the models to adapt to distribution shifts. Through this approach, we have gained insights into the generalization capabilities of the models and their ability to adapt to new environments. Our findings suggest that as models improve their performance on known locations or time periods, their ability to generalize to unfamiliar scenarios also tends to improve.

Additionally, the establishment of baseline models enables the comparison of novel approaches against established benchmarks. Our research has also contributed to the development of evaluation methodologies that account for the unique challenges of spatio-temporal forecasting, ensuring robust and reliable assessments of model performance. These fundamental components collectively provide a framework for evaluating the generalization and adaptability of deep learning models in spatio-temporal forecasting.

6.2 Conclusions and Future Work

This thesis embarked on an ambitious exploration of deep learning methodologies applied to spatio-temporal processes, centering on the intricate domains of traffic and weather forecasting.

We have delved into the challenges associated with encoding traffic data for deep learning, recognizing the transformative potential of these methodologies in shaping mobility patterns, streamlining city planning, and optimizing freight delivery services. We aimed to elucidate a clear and efficient pathway to harness GPS data, offering insights into how to employ deep learning models for accurate traffic prediction.

Similarly, we addressed the complex task of weather forecasting. Given the enormous impact of weather on human activities and the environment, we set out to establish best practices for encoding weather data and identify effective deep learning models to perform superior weather prediction.

Underpinning this investigation was the recognition of the critical need for a robust and comprehensive benchmark within the realm of spatio-temporal forecasting. We proposed an assemblage of benchmark components, including carefully selected tasks, relevant metrics, strong baselines, and thorough evaluation methodologies, to enable a systematic comparison of models and accelerate advancements in the field.

The apex of our exploration was the organization of the Weather4cast competition. This competition has served as an effective proving ground, establishing benchmarks for spatio-temporal prediction in both weather and traffic domains, and inciting a wave of innovative solutions in these intricate fields.

In conclusion, this thesis has made significant strides in bridging the gap between deep learning and spatio-temporal forecasting. By offering innovative solutions for traffic and weather forecasting, we have not only contributed to the development of machine learning methodologies but have also honed the accuracy and reliability of spatio-temporal predictions.

Looking ahead, we anticipate this work will serve as a springboard for future research in this domain. Future efforts should aim to improve the limitations of the work presented. First, improving the scalability and efficiency of the proposed models, explore the potential of other deep learning architectures, and continue refining the proposed benchmark to better reflect real-world complexities. For instance, the datasets presented in this thesis could potentially serve for other tasks such as anomaly detection, extreme events warnings, data imputation, and uncertainty estimation.

Moreover, we hope to extend the application of our findings to other spatio-temporal processes such as seismic activity prediction, urban planning, and environmental monitoring. The lessons learned from traffic and weather forecasting may provide valuable insights for these and other domains.

In essence, we envisage our research to be a catalyst, inspiring further advancements in the application of deep learning to complex spatio-temporal processes. We look forward to seeing how our findings will be leveraged and built upon in the future to further the impact of deep learning on our understanding of the world.

Bibliography

S. Agrawal, L. Barrington, C. Bromberg, J. Burge, C. Gazen, and J. Hickey. Machine Learning for Precipitation Nowcasting from Radar Images. *arXiv:1912.12132 [cs, stat]*, Dec. 2019. URL http://arxiv.org/abs/1912.12132. arXiv: 1912.12132.

P. R. Agudo, J. A. L. García, J. G. Pereda, M. Ángel Martínez Rubio, X. Calbet, L. L. Valverde, N. P. Galan, Óscar Alonso, J. A. G. Pastor, and P. Herruzo. Arcimis: Saf de nowcasting (nwc saf) liderado por aemet. *Congresos AET (Asociación Española de Teledetección)*, 2022. URL https://repositorio.aemet.es/handle/20.500.11765/14291.

D. H. Ballard. Modular learning in neural networks. In *Proceedings of the Sixth National Conference on Artificial Intelligence - Volume 1*, AAAI'87, page 279–284. AAAI Press, 1987. ISBN 0934613427.

P. Bauer, A. Thorpe, and G. Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55, Sept. 2015. ISSN 1476-4687. doi: 10.1038/nature14956. URL https://www.nature.com/articles/nature14956.

P. Bauer, P. D. Dueben, T. Hoefler, T. Quintino, T. C. Schulthess, and N. P. Wedi. The digital revolution of Earth-system science. *Nature Computational Science*, 1 (2):104–113, Feb. 2021. ISSN 2662-8457. doi: 10.1038/s43588-021-00023-0. URL https://www.nature.com/articles/s43588-021-00023-0. Number: 2 Publisher: Nature Publishing Group.

Y. Belousov, S. Polezhaev, and B. Pulfer. Solving the weather4cast challenge via visual transformers for 3d images, 2022. URL https://arxiv.org/abs/2212.02456.

Y. Bengio. *Practical Recommendations for Gradient-Based Training of Deep Architectures*, pages 437–478. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

ISBN 978-3-642-35289-8. doi: 10.1007/978-3-642-35289-8_26. URL https://doi.org/10.1007/978-3-642-35289-8_26.

L. Berthomier, B. Pradel, and L. Perez. Cloud Cover Nowcasting with Deep Learning. *2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6, Nov. 2020. doi: 10.1109/IPTA50016.2020.9286606. URL http://arxiv.org/abs/2009.11577. arXiv: 2009.11577.

K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian. Pangu-weather: A 3d high-resolution model for fast and accurate global weather forecast, 2022. URL https://arxiv.org/abs/2211.02556.

A. Bojesomo, H. Al-Marzouqi, and P. Liatsis. Spatiotemporal swin-transformer network for short time weather forecasting. In G. Cong and M. Ramanath, editors, *Proceedings of the CIKM 2021 Workshops co-located with 30th ACM International Conference on Information and Knowledge Management (CIKM 2021), Virtual Event, QLD, November 1-5, 2021*, CEUR Workshop Proceedings. CEUR-WS.org, 2021. *in press*.

C. Chen, K. Li, S. G. Teo, X. Zou, K. Wang, J. Wang, and Z. Zeng. Gated residual recurrent graph neural networks for traffic prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):485–492, Jul. 2019. doi: 10.1609/aaai.v33i01.3301485. URL https://ojs.aaai.org/index.php/AAAI/article/view/3821.

S. Choi. Traffic map prediction using UNet based deep convolutional neural network. *arXiv:1912.05288 [cs, stat]*, Nov. 2019. URL http://arxiv.org/abs/1912.05288. arXiv: 1912.05288.

S. Choi. Utilizing UNet for the future traffic map prediction task Traffic4cast challenge 2020. *arXiv:2012.00125 [cs, eess]*, Nov. 2020. URL http://arxiv.org/abs/2012.00125. arXiv: 2012.00125.

S. Choi. Utilizing unet for the future weather prediction: Weather4cast 2021'. In G. Cong and M. Ramanath, editors, *Proceedings of the CIKM 2021 Workshops co-located with 30th ACM International Conference on Information and Knowledge Management (CIKM 2021), Virtual Event, QLD, November 1-5, 2021*, CEUR Workshop Proceedings. CEUR-WS.org, 2021. *in press*.

P. G. C. College. Electromagnetic waves, 2020. URL https://phys.libretexts.org/Courses/Prince_Georges_Community_College/PHY_2040%3A_General_Physics_III/03%3A_Electromagnetic_Waves/3.1%3A_The_Electromagnetic_Spectrum#:~:text=or%20HD%20format.-,Microwaves,300%20MHz%20and%20300%20GHz. Accessed: 2023-05-08.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

S. Du, T. Li, X. Gong, and S.-J. Horng. A hybrid method for traffic flow forecasting using multimodal deep learning, 2019.

V. Dumoulin and F. Visin. A guide to convolution arithmetic for deep learning, 2016.

C. Eichenberger, M. Neun, H. Martin, P. Herruzo, M. Spanring, Y. Lu, S. Choi, V. Konyakhin, N. Lukashina, A. Shpilman, N. Wiedemann, M. Raubal, B. Wang, H. L. Vu, R. Mohajerpoor, I. Kim, L. Hermes, A. Melnik, R. Velioglu, M. Vieth, M. Schilling, A. Bojesomo, H. Al Marzouqi, P. Liatsis, J. Santokhi, D. Hillier, Y. Yang, J. Sarwar, A. Jordan, E. Hewage, D. Jonietz, F. Tang, A. Gruca, M. Kopp, D. Kreil, and S. Hochreiter. Traffic4cast at neurips 2021 – temporal and spatial few-shot transfer learning in gridded geo-spatial processes. In H. J. Escalante and K. Hofmann, editors, *Proceedings of the NeurIPS 2021 Competition Track*, volume forthcoming of *Proceedings of Machine Learning Research*. PMLR, 2022.

L. Espeholt, S. Agrawal, C. Sønderby, M. Kumar, J. Heek, C. Bromberg, C. Gazen, J. Hickey, A. Bell, and N. Kalchbrenner. Skillful twelve hour precipitation forecasts using large context neural networks. *arXiv preprint arXiv:2111.07470*, 2021. URL https://arxiv.org/abs/2111.07470v1.

W. Falcon et al. Pytorch lightning. *GitHub. Note: https://github.com/PyTorchLightning/pytorch-lightning*, 3:6, 2019.

P. Ghamisi, O. Ghorbanzadeh, Y. Xu, P. Herruzo, D. Kreil, M. Kopp, and S. Hochreiter. The landslide4sense competition 2022, 12 2022.

R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

A. Gruca, P. Herruzo, P. Rípodas, A. Kucik, C. Briese, M. K. Kopp, S. Hochreiter, P. Ghamisi, and D. P. Kreil. *CDCEO'21 - First Workshop on Complex Data Challenges in Earth Observation*, page 4878–4879. Association for Computing Machinery, New York, NY, USA, 2021a. ISBN 9781450384469. URL `https://doi.org/10.1145/3459637.3482044`.

A. Gruca, P. Herruzo, P. Rípodas, A. Kucik, C. Briese, M. K. Kopp, S. Hochreiter, P. Ghamisi, and D. P. Kreil. Cdceo'21 - first workshop on complex data challenges in earth observation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 4878–4879, New York, NY, USA, 2021b. Association for Computing Machinery. ISBN 9781450384469. doi: 10.1145/3459637.3482044. URL `https://doi.org/10.1145/3459637.3482044`.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015.

P. Herruzo and J. L. Larriba-Pey. Recurrent Autoencoder with Skip Connections and Exogenous Variables for Traffic Forecasting. In *NeurIPS 2019 Competition and Demonstration Track*, pages 47–55. PMLR, Aug. 2020. URL `http://proceedings.mlr.press/v123/herruzo20a.html`. ISSN: 2640-3498.

P. Herruzo, L. Portell Penadés, A. Soto, and B. Remeseiro. Towards objective description of eating, socializing and sedentary lifestyle patterns in egocentric images. In *BMVC Cardiff 2019*, 2019.

P. Herruzo, A. Gruca, L. Lliso, X. Calbet, P. Rípodas, S. Hochreiter, M. Kopp, and D. P. Kreil. High-resolution multi-channel weather forecasting – first insights on transfer learning from the weather4cast competitions 2021. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 5750–5757, 2021a. doi: 10.1109/BigData52589.2021.9672063.

P. Herruzo, A. Gruca, L. Lliso, X. Calbet, P. Rípodas, S. Hochreiter, M. Kopp, and D. P. Kreil. High-resolution multi-channel weather forecasting – First insights on transfer learning from the Weather4cast Competitions 2021. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 5750–5757, Dec. 2021b. doi: 10.1109/BigData52589.2021.9672063.

H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellan, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, G. De Chiara, P. Dahlgren, D. Dee, M. Diamantakis, R. Dragani, J. Flemming, R. Forbes, M. Fuentes, A. Geer, L. Haimberger, S. Healy, R. J. Hogan, E. Hólm, M. Janisková, S. Keeley, P. Laloyaux, P. Lopez, C. Lupu, G. Radnoti, P. de Rosnay, I. Rozum, F. Vamborg, S. Villaume, and J.-N. Thépaut. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: https://doi.org/10.1002/qj.3803. URL `https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803`.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8): 1735–1780, nov 1997a. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL `https://doi.org/10.1162/neco.1997.9.8.1735`.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8): 1735–1780, nov 1997b. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL `https://doi.org/10.1162/neco.1997.9.8.1735`.

Y. Hu, L. Chen, Z. Wang, X. Pan, and H. Li. Towards a more realistic and detailed deep-learning-based radar echo extrapolation method. *Remote Sensing*, 14(1), 2022. ISSN 2072-4292. doi: 10.3390/rs14010024. URL `https://www.mdpi.com/2072-4292/14/1/24`.

A. Huuskonen, E. Saltikoff, and I. Holleman. The operational weather radar network in europe. *Bulletin of the American Meteorological Society*, 95:897–907, 2014. doi: https://doi.org/10.1175/BAMS-D-12-00216.1.

W. Jiang and J. Luo. Graph neural network for traffic forecasting: A survey. *Expert Systems with Applications*, 207:117921, nov 2022. doi: 10.1016/j.eswa.2022.117921. URL https://doi.org/10.1016%2Fj.eswa.2022.117921.

W. Jiang and L. Zhang. Geospatial data to images: A deep-learning framework for traffic forecasting. *Tsinghua Science and Technology*, 24(1):52–64, 2019. doi: 10.26599/TST.2018.9010033.

D. Jonietz and M. Kopp. Towards modeling geographical processes with generative adversarial networks (gans) (short paper). In *COSIT*, 2019.

C. Kaparakis and S. Mehrkanoon. Wf-unet: Weather fusion unet for precipitation nowcasting, 2023. URL https://arxiv.org/abs/2302.04102.

A. Karagiannidis, A. Bloutsos, P. Maheras, and C. Sachsamanoglou. Some statistical characteristics of precipitation in Europe. *Theor Appl Climatol*, 91:193–204, 2008. doi: 10.1007/s00704-007-0303-7.

T. Kim, S. Kang, H. Shin, D. Yoon, S. Eom, K. Shin, and S.-Y. Yun. Region-conditioned orthogonal 3d u-net for weather4cast competition, 2022. URL https://arxiv.org/abs/2212.02059.

A. D. King and D. J. Karoly. Climate extremes in europe at 1.5 and 2 degrees of global warming. *Environmental Research Letters*, 12(11):114031, 2017. doi: 10.1088/1748-9326/aa8e2c.

D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, Jan. 2017. URL http://arxiv.org/abs/1412.6980. arXiv: 1412.6980.

M. Kopp, D. Kreil, M. Neun, D. Jonietz, H. Martin, P. Herruzo, A. Gruca, A. Soleymani, F. Wu, Y. Liu, J. Xu, J. Zhang, J. Santokhi, A. Bojesomo, H. A. Marzouqi, P. Liatsis, P. H. Kwok, Q. Qi, and S. Hochreiter. Traffic4cast at neurips 2020 - yet more on the unreasonable effectiveness of gridded geo-spatial processes. In H. J. Escalante and K. Hofmann, editors, *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *Proceedings of Machine Learning Research*, pages 325–343. PMLR, 06–12 Dec 2021. URL https://proceedings.mlr.press/v133/kopp21a.html.

D. P. Kreil, M. K. Kopp, D. Jonietz, M. Neun, A. Gruca, P. Herruzo, H. Martin, A. Soleymani, and S. Hochreiter. The surprising efficiency of framing geo-spatial time series forecasting as a video prediction task – Insights from the IARAI Traffic4cast Competition at NeurIPS 2019. In *NeurIPS 2019 Competition and Demonstration Track*, pages 232–241. PMLR, Aug. 2020. URL `http://proceedings.mlr.press/v123/kreil20a.html`. ISSN: 2640-3498.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, page 1097–1105, Red Hook, NY, USA, 2012. Curran Associates Inc.

P. H. Kwok and Q. Qi. A variational u-net for weather forecasting. In G. Cong and M. Ramanath, editors, *Proceedings of the CIKM 2021 Workshops co-located with 30th ACM International Conference on Information and Knowledge Management (CIKM 2021), Virtual Event, QLD, November 1-5, 2021*, CEUR Workshop Proceedings. CEUR-WS.org, 2021a. *in press*.

P. H. Kwok and Q. Qi. Enhanced variational U-Net for weather forecasting. In *Proceedings of the 2021 IEEE International Conference on Big Data*, 2021b. *in press.*

R. Lam, A. Sanchez-Gonzalez, M. Willson, P. Wirnsberger, M. Fortunato, A. Pritzel, S. Ravuri, T. Ewalds, F. Alet, Z. Eaton-Rosen, W. Hu, A. Merose, S. Hoyer, G. Holland, J. Stott, O. Vinyals, S. Mohamed, and P. Battaglia. Graphcast: Learning skillful medium-range global weather forecasting, 2022. URL `https://arxiv.org/abs/2212.12794`.

Y. LeCun, B. Boser, J. Denker, D. Henderson, W. Hubbard, and L. Jackel. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 2, 1989.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

J. Leinonen. Spatiotemporal weather data predictions with shortcut recurrent-convolutional networks: A solution for the weather4cast challenge. In G. Cong and M. Ramanath, editors, *Proceedings of the CIKM 2021 Workshops co-located with 30th ACM International Conference on Information and Knowledge Management*

(CIKM 2021), Virtual Event, QLD, November 1-5, 2021, CEUR Workshop Proceedings. CEUR-WS.org, 2021a. *in press*.

J. Leinonen. Improvements to short-term weather prediction with recurrent-convolutional networks. In *Proceedings of the 2021 IEEE International Conference on Big Data*, 2021b. *in press.*

Y. Li, H. Dong, Z. Fang, J. Weyn, and P. Luferenko. Super-resolution probabilistic rain prediction from satellite data using 3d u-nets and earthformers. *arXiv preprint arXiv:2212.02998*, 2022. URL `https://arxiv.org/abs/2212.02998`.

T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. Microsoft coco: Common objects in context, 2015.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

J. Marshall, R. Langille, and W. Palmer. Measurement of rainfall by radar. *Journal of Meteorology*, 4, 1947. doi: https://doi.org/10.1175/1520-0469(1947)004⟨0186:MORBR⟩2.0.CO;2.

A. H. Nielsen, A. Iosifidis, and H. Karstoft. Cloudcast: A satellite-based dataset and baseline for forecasting clouds. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pages 1–1, 2021. doi: 10.1109/JSTARS.2021.3062936.

J. Park, M. Son, S. Cho, I. Lee, and C. Kim. Rainunet for super-resolution rain movie prediction under spatio-temporal shifts. *arXiv preprint arXiv:2212.04005*, 2022.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library.

E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

J. Pihrt, R. Raevskiy, P. Šimánek, and M. Choma. Weatherfusionnet: Predicting precipitation from satellite data, 2022. URL https://arxiv.org/abs/2211.16824.

A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training. *openai*, 2018.

S. Rasp, P. D. Dueben, S. Scher, J. A. Weyn, S. Mouatadid, and N. Thuerey. WeatherBench: A Benchmark Data Set for Data-Driven Weather Forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11):e2020MS002203, 2020. ISSN 1942-2466. doi: https://doi.org/10.1029/2020MS002203. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020MS002203.

S. Ravuri, K. Lenc, M. Willson, D. Kangin, R. Lam, P. Mirowski, M. Fitzsimons, M. Athanassiadou, S. Kashem, S. Madge, and et al. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878): 672–677, Sep 2021a. ISSN 1476-4687. doi: 10.1038/s41586-021-03854-z. URL http://dx.doi.org/10.1038/s41586-021-03854-z.

S. Ravuri, K. Lenc, M. Willson, D. Kangin, R. Lam, P. Mirowski, M. Fitzsimons, M. Athanassiadou, S. Kashem, S. Madge, R. Prudden, A. Mandhane, A. Clark, A. Brock, K. Simonyan, R. Hadsell, N. Robinson, E. Clancy, A. Arribas, and S. Mohamed. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878):672–677, Sept. 2021b. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-021-03854-z. URL https://www.nature.com/articles/s41586-021-03854-z.

C. Requena-Mesa, V. Benson, M. Reichstein, J. Runge, and J. Denzler. EarthNet2021: A large-scale dataset and challenge for Earth surface forecasting as a guided video prediction task. *arXiv:2104.10066 [cs]*, Apr. 2021. URL http://arxiv.org/abs/2104.10066. arXiv: 2104.10066.

S. Respati, A. Bhaskar, and E. Chung. Traffic data characterisation: Review and challenges. *Transportation Research Procedia*, 34:131 – 138, 2018. ISSN 2352-1465. doi: https://doi.org/10.1016/j.trpro.2018.11.024. URL http://www.sciencedirect.com/science/article/pii/S2352146518303132. International Symposium of Transport Simulation (ISTS'18) and the International Workshop on Traffic Data Collection and its Standardization (IWTDCS'18)Emerging Transport Technologies for Next Generation Mobility.

O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv:1505.04597 [cs]*, May 2015. URL http://arxiv.org/abs/1505.04597. arXiv: 1505.04597.

S. Ryu, D. Kim, and J. Kim. Weather-aware long-range traffic forecast using multi-module deep neural network. *Applied Sciences*, 10:1938, 03 2020. doi: 10.3390/app10061938.

E. Saltikoff, G. Haase, L. Delobbe, N. Gaussiat, M. Martet, D. Idziorek, H. Leijnse, P. Novák, M. Lukach, and K. Stephan. Opera the radar project. *Atmosphere*, 10,320, 2019a. doi: https://doi.org/10.3390/atmos10060320.

E. Saltikoff, G. Haase, L. Delobbe, N. Gaussiat, M. Martet, D. Idziorek, H. Leijnse, P. Novák, M. Lukach, and K. Stephan. OPERA the Radar Project. *Atmosphere*, 10 (6):320, June 2019b. doi: 10.3390/atmos10060320. URL https://www.mdpi.com/2073-4433/10/6/320. Number: 6 Publisher: Multidisciplinary Digital Publishing Institute.

J. Schmetz, P. Pili, S. Tjemkes, D. Just, J. Kerkmann, S. Rota, and A. Ratier. AN INTRODUCTION TO METEOSAT SECOND GENERATION (MSG). *Bulletin of the American Meteorological Society*, 83(7):977–992, July 2002. ISSN 0003-0007, 1520-0477. doi: 10.1175/1520-0477(2002)083⟨0977:AITMSG⟩2.3.CO;2. URL https://journals.ametsoc.org/view/journals/bams/83/7/1520-0477_2002_083_0977_aitmsg_2_3_co_2.xml. Publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society.

C. Schroeder de Witt, C. Tong, V. Zantedeschi, D. De Martini, A. Kalaitzis, M. Chantry, D. Watson-Parris, and P. Bilinski. RainBench: Enabling Data-Driven Precipitation Forecasting on a Global Scale. other, pico, Mar. 2021. URL https://meetingorganizer.copernicus.org/EGU21/EGU21-1762.html.

M. Seo, D. Kim, S. Shin, E. Kim, S. Ahn, and Y. Choi. Domain generalization strategy to train classifiers robust to spatial-temporal shift. *arXiv preprint arXiv:2212.02968*, 2022a.

M. Seo, D. Kim, S. Shin, E. Kim, S. Ahn, and Y. Choi. Simple baseline for weather forecasting using spatiotemporal context aggregation network. *arXiv preprint arXiv:2212.02952*, 2022b.

J. Shcmetz, P. Pili, S. Tjemkes, J. Kerkmann, S. Rota, and A. Ratier. An introduction to meteosat second generation (msg). *Bulletin of the American Meteorological Society*, 83(7):977–992, 2002.

X. SHI, Z. Chen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. WOO. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 802–810. Curran Associates, Inc., 2015.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

Z. Sokol, J. Szturc, J. Orellana-Alvear, J. Popová, A. Jurczyk, and R. Célleri. The role of weather radar in rainfall estimation and its application in meteorological and hydrological modelling—a review. *Remote Sensing*, 13, 351, 2021. doi: https://doi.org/10.3390/rs13030351.

C. K. Sønderby, L. Espeholt, J. Heek, M. Dehghani, A. Oliver, T. Salimans, S. Agrawal, J. Hickey, and N. Kalchbrenner. MetNet: A Neural Weather Model for Precipitation Forecasting. *arXiv:2003.12140 [physics, stat]*, Mar. 2020. URL `http://arxiv.org/abs/2003.12140`. arXiv: 2003.12140.

M. Veillette, S. Samsi, and C. Mattioli. Sevir: A storm event imagery dataset for deep learning applications in radar and satellite meteorology. *Advances in Neural Information Processing Systems*, 33:22009–22019, 2020a.

M. S. Veillette, S. Samsi, and C. J. Mattioli. SEVIR : A Storm Event Imagery Dataset for Deep Learning Applications in Radar and Satellite Meteorology. *NeurIPS2020*, page 11, 2020b.

A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf`.

O. W. Weather. Weather data and api, 2019. URL https://www.worldweatheronline.com/developer/. Accessed: 2019-07-30.

N. Wichers, R. Villegas, D. Erhan, and H. Lee. Hierarchical long-term video prediction without supervision. In *ICML*, 2018.

Wikipedia contributors. Transportation forecasting — Wikipedia, the free encyclopedia, 2023. URL https://en.wikipedia.org/wiki/Transportation_forecasting. [Online; accessed 21-May-2023].

World Meteorological Organization. Guide to instruments and methods of observation. Technical Report WMO No. 8, 2018.

L. Zhao, O. Gkountouna, and D. Pfoser. Spatial auto-regressive dependency interpretable learning based on spatial topological constraints. *ACM Trans. Spatial Algorithms Syst.*, 5(3), aug 2019. ISSN 2374-0353. doi: 10.1145/3339823. URL https://doi.org/10.1145/3339823.

J. Zhuang, T. Tang, Y. Ding, S. Tatikonda, N. Dvornek, X. Papademetris, and J. S. Duncan. Adabelief optimizer: Adapting stepsizes by the belief in observed gradients, 2020.

I. I. Zveryaev. Seasonality in precipitation variability over europe. *Journal of Geophysical Research: Atmospheres*, 109(D5), 2004. doi: https://doi.org/10.1029/2003JD003668.