Astronomical Image Time-series Classification Using Deep Learning

Anass Bairouk

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Informatique

École doctorale: Information, Structures, Systèmes

Unité de recherche: LIRMM

Astronomical Image Time-series Classification Using Deep Learning

Présentée par Anass BAIROUK
le 17-Octobre-2023

Sous la direction de Marc CHAUMONT, Dominique FOUCHEZ, Jerome PASQUET et Frederic COMBY

Devant le jury composé de

Alexandre BENOIT	PR	Université Savoie Mont Blanc	Rapporteur
Thierry ARTIERES	PR	Ecole Centrale Marseille	Rapporteur
Emille ISHIDA	IR	Université Clermont-Auvergne	Examinatrice
Sandra BRINGAY	PR	Université Paul-Valéry Montpellier 3	Examinatrice
Marc CHAUMONT	MCF-HDR	Université de Nîmes	Directeur de thèse
Dominique FOUCHEZ	PR	Université d'Aix-Marseille	Codirecteur de thèse
Jerome PASQUET	MCF	Université Paul-Valéry Montpellier 3	Encadrant
Frederic COMBY	MCF	Université de Montpellier	Encadrant

UNIVERSITÉ DE MONTPELLIER

DEDICATION

This scholarly opus is tenderly consecrated as a humble act of Sadaqah, in loving remembrance of my revered grandmother, Ymana ZAIKOUK. May her soul forever rest in the serenity of peace.

ACKNOWLEDGMENTS

"Urti nna γ ur tendert, ur rad gi-s temgert."

Amazigh proverb

I feel deeply privileged and honored to express my sincere gratitude to my supervisors, Doctor Marc CHAUMONT, Professor Dominique FOUCHEZ, Doctor Jerome PASQUET, and Doctor Frederic COMBY. Their tireless guidance, invaluable insights, and continuous support have been the pillars upon which this thesis has been constructed. They have demonstrated unwavering patience in their review and feedback, providing constructive suggestions to improve the quality of my work. It has been an enriching experience working under their mentorship, and for this, I am extremely thankful.

To my examination committee, Professor Alexandre BENOIT, Professor Thierry ARTIERES, Professor Sandra BRINGAY, and Doctor Emille ISHIDA, I extend my heartfelt appreciation. The time, energy, and expertise you committed to this process have been a critical component of my academic journey.

In the quiet background of my academic endeavors, my family has been my beacon of hope and strength. To my mother and father, your unwavering faith in my abilities and constant encouragement has been the fuel that drives me forward. To my wife, your understanding and companionship have been my solace in times of stress and confusion. To my brother and my precious sisters, your support and love have kept me grounded and focused.

I am also grateful to the administrative staff of the LIRMM laboratory for their efficient and diligent support. Your dedication to maintaining a conducive research environment is noteworthy and much appreciated. To all the members of the ICAR team, your camaraderie and cooperation have made this journey an en-

joyable and enriching experience. Thank you for fostering a community of learning, growth, and shared knowledge.

Thank you all for being part of my journey towards this significant milestone. Your support and belief have been instrumental in shaping my research path and personal growth.

ABSTRACT

The vast and complex universe presents a formidable challenge to the astronomical community, especially with the advent of powerful telescopes capturing unprecedented amounts of data. This dissertation focuses on the application of deep learning techniques for astronomical image time-series classification, a task essential in contemporary astronomy for identifying and studying time-dependent phenomena.

The primary contribution of this research is the development of a deep learning-based model that uses sequences of images of celestial bodies at different times, rather than relying on light curves. This approach allows for the extraction of richer contextual information, leading to improved classification performance despite the intrinsic noise and high dynamic range of astrophysical data.

We introduced the ConvEntion model, a novel approach for classifying different types of space objects directly using images. Leveraging the power of convolutions and transformers, the ConvEntion model has shown impressive performance in mitigating the problem of missing observations, a common issue in astronomical data.

Additionally, we presented the Semiconformer framework, a semi-supervised learning framework for image time series representation. This framework leverages self-supervised approaches to reduce the effect of class imbalance with data augmentation, minimize intra-class variance, and improve results on small datasets without the need for extra labeled data.

These contributions represent significant advancements in the field of astronomical image time series classification, paving the way for future research in this area.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
AI	Artificial Intelligence
LSST	Vera Rubin Observatory Legacy Survey for Space and Time
AGNs	Active Galactic Nuclei
ConvLSTM ..	Convolutional Long Short-Term Memory
FC-LSTM ...	Fully Connected Long Short-Term Memory
AITS	Astronomical Image Time Series
3D CNN ...	3D Convolutional Neural Network
SDSS	Sloan Digital Sky Survey
ConvBERT ..	Convolutional BERT
t-SNE	t-Distributed Stochastic Neighbor Embedding
ROC	Receiver Operating Characteristic
PR	Precision-Recall
ViT	Vision Transformer
NLP	Natural Language Processing
T2T-ViT	Tokens-To-Token Vision Transformer
BYOL	Bootstrap Your Own Latent
DINO	Distillation of knowledge in the form of a student network
SemiConFormer	Semi-Convolutional transFormer
VideoBERT ..	Video Bidirectional Encoder Representations from Transformers

ViViT	Video Vision Transformer
SITS-BERT ..	Satellite Image Time Series BERT
TransVG ...	Transformer Video Generation

CHAPTER 1

Introduction

1.1 CONTEXT AND MOTIVATION

The quest to understand the universe and its evolution, including its expansion and the formation of large structures, is a central concern of modern cosmology. This journey of discovery relies heavily on the identification and study of various celestial bodies that serve as probes into the distant universe. Among these, Type Ia supernovae are of particular interest due to their standard explosion mechanism (producing almost identical flux energy for each supernova), which allows for the deduction of the luminous distance of the exploded supernova. However, the detection of these stars is a challenging task due to the similar behaviors exhibited by many other celestial bodies. This necessitates the implementation of robust machine learning methods.

The analysis of these celestial bodies is at first conducted using photometric information, which involves constructing a time series, known as a "light curve," representing the amplitude values of an object in a specific band over time. After this initial approach, classification could be done with spectral typing, which is very costly in terms of observation time and/or the need for a large telescope. Despite numerous studies on the light curves, the classification results remain insufficient for conducting astrophysical science. This is where the role of machine learning becomes invaluable. Machine learning techniques have been utilized in various aspects of astrophysics research, including detecting and classifying celestial objects, examining galaxy formation and evolution, and determining fundamental cosmological parameters.

The advent of big data in astronomy has transformed the field, with observatories worldwide generating terabytes of data every night. This presents unique opportunities to examine the cosmos with unparalleled precision. However, extracting significant information from these enormous volumes of data is a primary challenge. Conventional statistical methods are beginning to show their limits in managing the complexity and scale of contemporary astronomical datasets. As a result, researchers have turned to machine learning algorithms, which have shown exceptional success in handling large-scale and complex data.

The focus of this dissertation is to extend this work by using images of celestial bodies at different times directly, that is, sequences of images centered on the object of interest. The goal is to construct a deep learning-based model capable of extracting contextual information and thus improving performance compared to results obtained from the use of light curves. This is a complex task due to the intrinsic noise and high dynamic range of astrophysical data, as well as the problem of mismatch between the statistics of the training and test databases. The redshift phenomenon, which drastically alters the perceived image, further complicates this task.

In this context, the motivation for this research is to develop and apply advanced machine learning techniques to enhance the accuracy and efficiency of astronomical image time series classification, thereby fully unlocking the potential of the big data era in astronomy.

1.2 SCOPE AND CHALLENGES

The scope of this research is centered on the classification of astronomical image time series, a task that is essential in contemporary astronomy as it enables the

identification and study of time-dependent phenomena. This task is particularly challenging due to the vast and complex datasets generated by modern astronomical surveys. The primary challenges in this field can be broadly categorized into data-related challenges, methodological challenges, and computational challenges.

Data-related challenges stem from the inherent complexity and variability of astronomical phenomena. The data used in this field, which includes astronomical image time series, is intrinsically very noisy and has a high dynamic range. The database is subject to the mismatch problem, where the training data and real-world data have different statistical distributions. This is because the training data is primarily composed of observations of "nearby" astrophysical objects, which have a different spectrum than distant objects. The redshift phenomenon drastically alters the perceived image of distant objects, which can lead to misclassifications when the model is applied to real-world data.

Methodological challenges arise from the need for meticulous data preparation, the challenge of handling incomplete or irregular data, and the complexity and computational intensity of certain neural network models. The use of synthetic image sequences, as proposed by some researchers, is a novel approach to circumvent the issue of dataset scarcity. However, the synthetic data must be close enough to real-world data for effective training and testing, which may be challenging due to the inherent complexity and variability of astronomical phenomena.

Computational challenges are associated with the development of advanced algorithms and computational tools to effectively process, analyze, and interpret the information within these immense datasets. These algorithms have the poten-

tial to expose previously unknown relationships and patterns in data, ultimately leading to new discoveries and a more profound understanding of the universe. However, the computational demands of advanced neural networks and the need for managing the complexities of these networks present significant challenges.

Addressing these challenges requires continued research and development efforts and may also necessitate greater collaboration and resource sharing within the astronomical research community. The ultimate goal remains the development of robust, efficient, and accurate classification models that can handle the vast and complex datasets generated by modern astronomical surveys.

1.3 CONTRIBUTIONS

This dissertation makes several significant contributions to the field of astronomical image time series classification. The primary contribution is the development of a novel approach for classifying different types of space objects directly using images. This approach, named ConvEntion (CONVolutional attENTION), leverages the power of convolutions and transformers to process astronomical image time series. ConvEntion integrates spatiotemporal features and can be applied to various types of image datasets with any number of bands. The implementation of ConvEntion has led to substantial improvements in classification accuracy, with an increase of 13% compared to state-of-the-art approaches that use image time series and a 12% increase compared to approaches that use light curves.

A secondary contribution of this dissertation is the proposal of an end-to-end semi-supervised image time series representation learning framework. This framework leverages self-supervised approaches to reduce the effect of class imbalance with data augmentation, minimize intra-class variance, and improve results on

small datasets without the need for extra labeled data. The proposed framework, named Semiconformer, has demonstrated superior performance compared to a ConvEntion model, achieving an accuracy of 82.18% and an F1 Score of 75.33%.

These contributions represent significant advancements in the field of astronomical image time series classification. They not only improve the accuracy and efficiency of classification tasks but also pave the way for future research in this area. By demonstrating the potential of deep learning techniques in handling complex astronomical data, this dissertation provides a foundation for further exploration and development of advanced machine learning models for astronomical image analysis.

1.4 ORGANIZATION OF THE DISSERTATION

Chapter 2, titled "Machine Learning Approaches," provides a comprehensive review of the current machine learning techniques used in image time series classification. It discusses the strengths and weaknesses of these methods and identifies areas for improvement.

Chapter 3, "The Role of Big Data in Modern Astronomy," discusses the impact of big data on modern astronomy. It explores how the advent of big data has transformed the field and presents the challenges and opportunities it brings.

Chapter 4, "Astronomical Image Time Series Classification using CONVolutional attENTION (ConvEntion)," introduces the ConvEntion approach developed in this dissertation. It details the design and implementation of the ConvEntion model and presents the results of its application to astronomical image time series classification.

Chapter 5, "Semi-Supervised Image Time Series Representation Learning Us-

ing Convolutional Transformers," presents the Semiconformer framework, a semi-supervised learning framework for image time series representation. It discusses the design of the Semiconformer framework, its implementation, and the results of its application.

The dissertation concludes with Chapter 6, which provides a summary of the research findings, discusses the implications of the study, and suggests directions for future research. This chapter also reflects on the contributions of the dissertation to the field of astronomical image time series classification.

CHAPTER 2

Machine learning approaches

2.1 PROGRESSING FROM SHALLOW TO DEEP NEURAL NETWORKS

Deep learning, a specialized field of machine learning, involves the use of neural networks with three or more layers. These networks are designed to mimic the behavior of the human brain (Rosenblatt, 1958; McCulloch & Pitts, 1943; Hebb, 1949), work with significantly less capacity, and are capable of "learning" from vast amounts of data (Najafabadi et al., 2015). While a neural network with a single layer can still provide approximate predictions, adding hidden layers can greatly enhance its accuracy by enabling optimization and refinement. The tremendous potential of deep learning is driving its application in various areas of artificial intelligence (AI) where automation and analytical or physical tasks can be performed without human intervention. From digital assistants to voice-enabled TV remotes and credit card fraud detection, everyday products and services rely on the power of deep learning. In addition, emerging technologies such as self-driving cars are made possible by this cutting-edge technology (Badue et al., 2019). In this chapter, we will delve into different aspects of deep learning, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformers, to better understand their unique capabilities and applications.

2.1.1 Artificial Neural Networks

2.1.1.1 Perceptron

The perceptron, a fundamental building block of a neural network, serves as a binary linear classifier. It was initially introduced by Frank Rosenblatt in 1958, laying

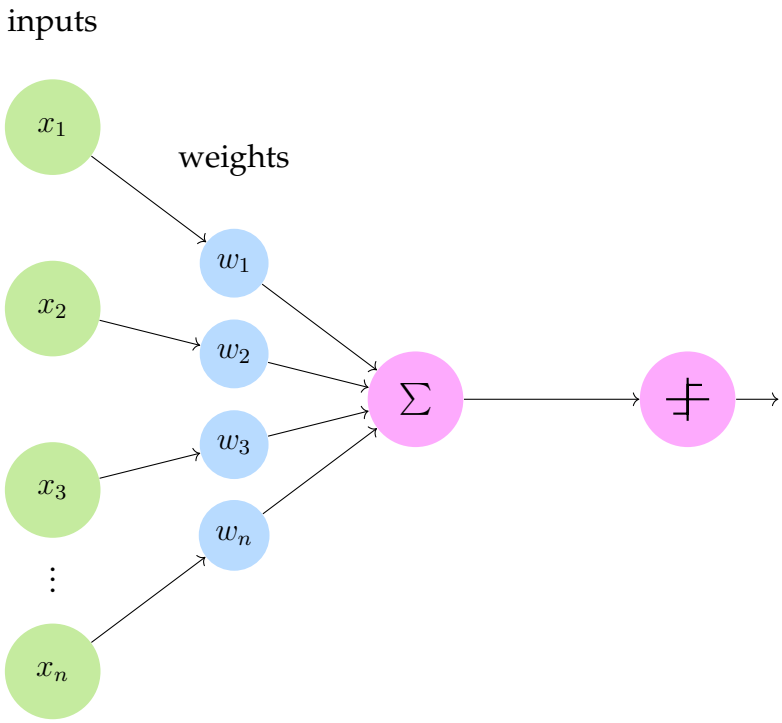

Figure 2.1: Illustration of a single perceptron. The perceptron takes input values $x_1, x_2, \ldots, x_n$ and computes a weighted sum $\sum_{i=1}^{n} w_i x_i + w_0$, where w_i is the weight associated with input x_i and w_0 is the bias term. The output of the perceptron is then passed through an activation function, typically the step function, to produce a binary output. The weights and bias term are adjusted during training using the perceptron learning rule to correctly classify input examples.

the foundation for modern machine learning (Rosenblatt, 1958). The functionality of a perceptron involves the accumulation of weighted inputs and the subsequent application of a basic activation function (refer to figure 2.1). The perceptron utilizes a bias term, which acts like an offset or shift of the activation function. If the activation function, considering this bias, results in a positive output, the perceptron fires, returning a 1. Conversely, if the output is negative, it returns a 0. Through this mechanism, the perceptron efficiently classifies its inputs into two distinct categories.

In terms of the mathematics, the output of a perceptron can be described by the

formula:

$$y = f\left(\sum_i w_i x_i - b\right) \tag{2.1}$$

where y is the output, w_i is the weight assigned to the ith input x_i, and b is the bias (McCulloch & Pitts, 1943). The function f is the activation function, often the Heaviside step function in the simplest form of a perceptron.

$$H(x) = \begin{cases} 0, & \text{if } x < 0 \\ 1, & \text{if } x \geq 0 \end{cases} \tag{2.2}$$

Each variable in the perceptron operation carries specific meaning. The w_i weights are parameters learned by the perceptron during the training process, and they determine how each input feature influences the final output. The bias term b provides a sort of offset, shifting the decision boundary away from the origin and therefore allowing for more flexible model fitting. The x_i are the inputs, or features, of the perceptron, and y is the output, which is the class prediction.

The process of updating the perceptron involves adjusting the weights and bias to improve the model's predictions over time. If the perceptron makes a correct prediction, the weights and bias are left unchanged. However, if the perceptron misclassifies an instance, the weights and bias are updated using the following rules:

$$w_i = w_i + \Delta w_i = w_i + \eta(t - y)x_i \tag{2.3}$$

$$b = b + \Delta b = b + \eta(t - y) \tag{2.4}$$

where t is the target output, η is the learning rate, and y is the perceptron's current

output (Novikoff, 1962).

The perceptron has a rich history. As mentioned earlier, it was first introduced by Rosenblatt in 1958, and it was a pioneering concept in the field of artificial neural networks. However, the initial perceptron model was greatly limited because it could only solve linearly separable problems. This drawback was highlighted by Marvin Minsky and Seymour Papert in their book "Perceptrons" in 1969, which led to a significant decrease in interest and research on neural networks (Minsky & Papert, 1969). However, the introduction of multi-layer perceptrons and backpropagation in the 1980s helped to overcome these limitations, reigniting interest in neural networks and leading to the explosion of research and applications we see in the current day.

2.1.1.2 Multi-Layer Perceptron (MLP)

A multilayer perceptron (MLP) neural network is a type of feedforward neural network that consists of multiple layers of perceptrons. It is a powerful and versatile model that can be used for a wide range of tasks, including classification, regression, and feature extraction.

The MLP neural network consists of an input layer, one or more hidden layers, and an output layer. The input layer receives the input data and passes it to the hidden layers (see figure 2.2). Each hidden layer applies a linear transformation followed by a non-linear activation function to the inputs, and then passes the output to the next layer. The final layer, the output layer, applies a linear transformation followed by a final activation function to produce the final output.

The linear transformation for each hidden layer is similar to that of a single perceptron and is represented by the equation:

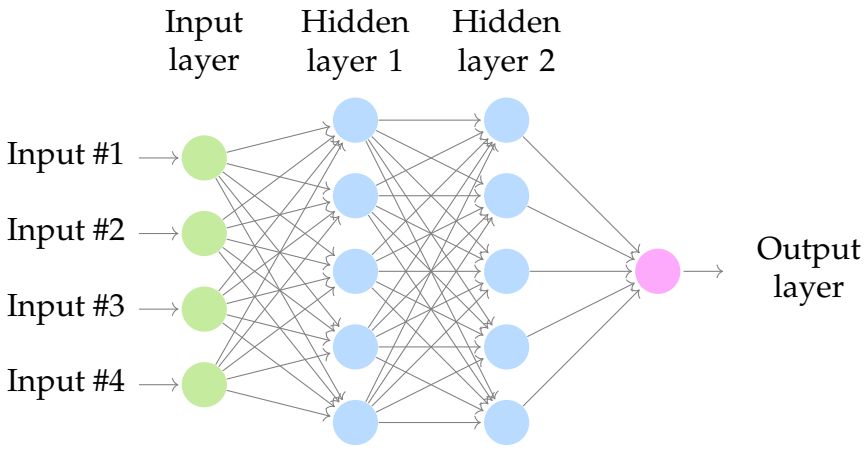

Figure 2.2: Neural network with two hidden layers.

$$z_j = \sum_{i=1}^{n} w_{ji} x_i + b_j \tag{2.5}$$

where z_j is the weighted sum for neuron j, w_{ji} is the weight for the connection between neuron i in the previous layer and neuron j in the current layer, b_j is the bias term for neuron j, and n is the number of neurons in the previous layer.

The activation function for each hidden layer can be any non-linear function, but the most commonly used functions are the sigmoid function, and the rectified linear unit (ReLU 2.3).

$$ReLU(x) = \begin{cases} x, & if\ x \geq 0 \\ 0, & else \end{cases} \tag{2.6}$$

The output layer also applies a linear transformation, but the final activation function depends on the task at hand. For binary classification problems, the sigmoid function is typically used, while for multi-class classification problems, the softmax function is used. For regression problems, the identity function is used (Goodfellow et al., 2016).

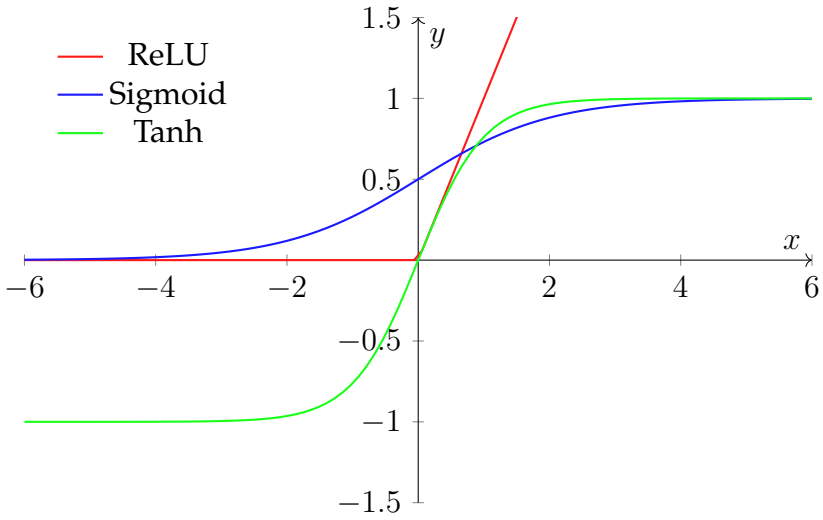

Figure 2.3: Illustration of activation functions: ReLU, Sigmoid, and Tanh.

To train the MLP, we need a way to measure how well it performs on a given task. This is done using a loss function, which compares the predicted output to the true output and outputs a scalar value representing the error. The most commonly used loss function for regression problems is the mean squared error (MSE), represented by the equation:

$$\mathcal{L} = \frac{1}{2m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 \tag{2.7}$$

where $\mathcal{L}$ is the loss function, m is the number of samples, y_i is the true label for sample i, and $\hat{y}_i$ is the predicted label for sample i.

For classification tasks, the cross-entropy loss is frequently used as the loss function. This can be represented by the following equation:

$$\mathcal{L} = -\sum_{j=1}^{C} y_j \log(\hat{y}_j) \tag{2.8}$$

In this equation, $\mathcal{L}$ represents the loss function, C is the number of classes, y_j is

the actual label for class j, and $\hat{y_j}$ is the predicted probability for class j. The value of y_j can either be 0 or 1, depending on whether the actual label belongs to class j or not. If the actual label is indeed class j, then $y_j = 1$, but if the actual label does not belong to class j, then $y_j = 0$.

To update the weights and biases of the MLP, we use an optimization algorithm called stochastic gradient descent (SGD). This algorithm updates the parameters in the direction that minimizes the loss function. The update equations for the weights and bias are:

$$\begin{aligned} w &= w - \eta \frac{\partial \mathcal{L}}{\partial w}, \\ b &= b - \eta \frac{\partial \mathcal{L}}{\partial b}. \end{aligned} \tag{2.9}$$

where w and b represent the weights and biases, $\mathcal{L}$ is the loss function, η is the learning rate, and $\frac{\partial \mathcal{L}}{\partial w}$ and $\frac{\partial \mathcal{L}}{\partial b}$ are the gradients of the loss with respect to the weights and biases, respectively.

2.1.2 Convolution Neural Network

Convolutional neural network (CNN) is a type of artificial neural network that can analyze visual data, such as images and videos. CNNs are composed of layers that perform different operations on the input data, such as convolution, pooling, and Multi-Layer Perceptron. CNNs can automatically learn features from the data, without requiring any manual feature engineering (O'Shea & Nash, 2015). CNNs have been widely used in computer vision tasks, such as image classification, object detection, and semantic segmentation (Gu et al., 2018).

The concept of Convolutional Neural Networks (CNNs) was materialized when

Fukushima (1980) proposed the neocognitron model in 1980. This model was the first neural network to utilize convolutional layers, marking a significant milestone in the field. The practical application of CNNs, however, became evident in 1989 with the development of LeNet-5 by LeCun et al. (1989) (see Figure 2.4), LeNet-5 demonstrated impressive performance on the task of handwritten digit recognition, proving the efficacy of CNNs in real-world tasks. Since then, CNNs have evolved significantly with various improvements on layer design, activation function, loss function, regularization, optimization and fast computation (Gu et al., 2018). Some of the most influential CNN architectures include AlexNet (Krizhevsky et al., 2017), VGGNet (Simonyan & Zisserman, 2014), ResNet (He et al., 2016), DenseNet (Huang et al., 2017).

CNNs are important in computer vision because they can effectively capture the spatial and hierarchical structure of visual data. They can also handle large-scale and high-dimensional data. Moreover, they can achieve state-of-the-art performance on various computer vision problems by leveraging large-scale datasets and powerful hardware (O'Shea & Nash, 2015).

2.1.2.1 The Convolution layer

Convolutional layers are the building blocks of convolutional neural networks (CNNs), which are widely used in computer vision tasks such as image classification, object detection, and segmentation. The convolutional layer applies a set of learnable filters to the input data, producing a set of feature maps that capture different aspects of the input.

Figure 2.5 illustrates the operation performed by a convolutional layer, called a convolution, which involves sliding a filter over the input and computing the dot

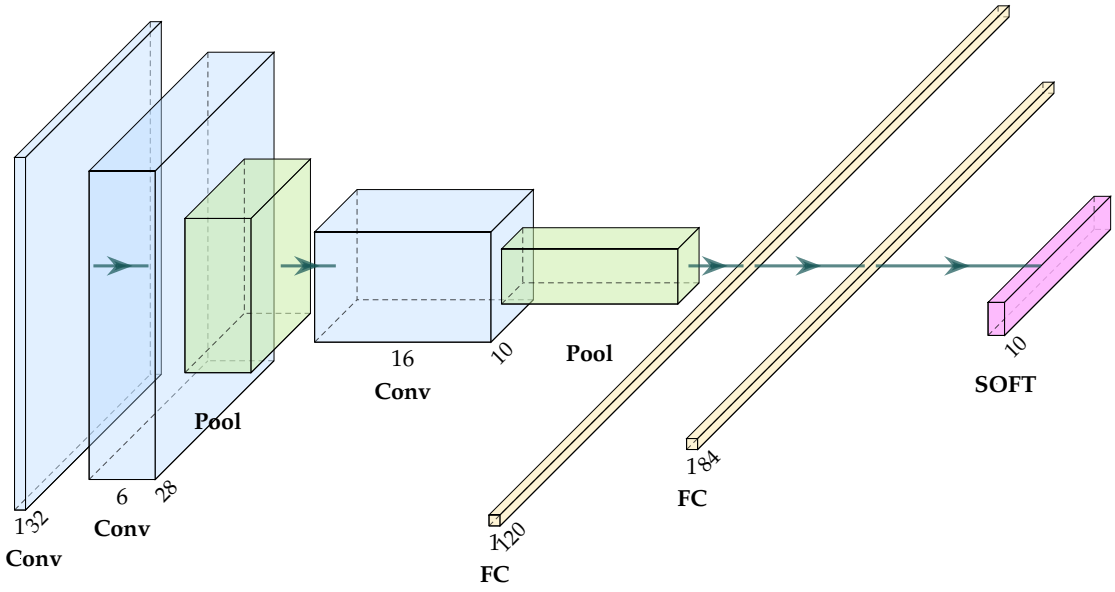

Figure 2.4: The original convolutional neural network architecture, first introduced by LeCun et al. (1989). The architecture alternates between convolutional layers (depicted in blue) with hyperbolic tangent non-linearities and subsampling layers (depicted in green). In this architecture, the convolutional layers already include non-linearities. The feature maps of the final subsampling layer are then fed into the classifier, which can consist of any number of fully connected layers (depicted in yellow). The output layer typically uses softmax activation functions (depicted in pink).

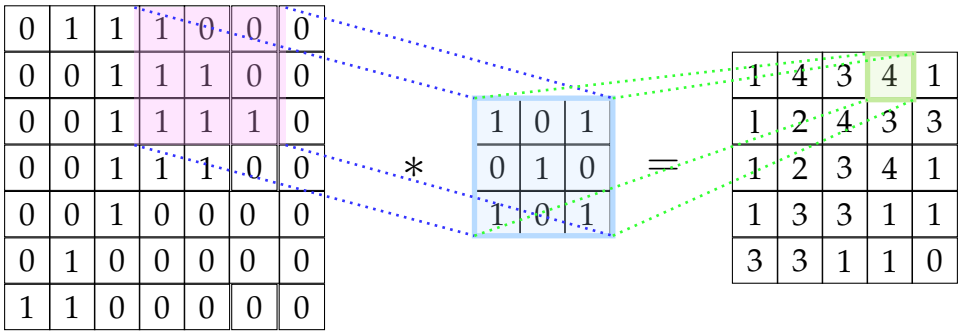

Figure 2.5: Illustration of a single convolutional layer. If layer l is a convolutional layer, the input image (if $l = 1$) or a feature map of the previous layer is convolved by one of the filters to yield the output feature map of layer l.

product between the filter and the input at each location. The resulting output is a feature map that highlights regions of the input that are correlated to the filter.

The convolution operation can be encapsulated mathematically as follows:

$$y_{i,j,k} = b_k + \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \sum_{c=0}^{C-1} x_{(i \cdot S+m-P),(j \cdot S+n-P),c} \cdot w_{m,n,c,k} \tag{2.10}$$

In this equation, $y_{i,j,k}$ signifies the output value at the position (i, j) in the k-th feature map. $w_{m,n,c,k}$ is the weight for the k-th feature map, while $x_{(i \cdot S+m-P),(j \cdot S+n-P),c}$ represents the input value at position $(i \cdot S + m - P), (j \cdot S + n - P)$ in the c-th input channel. b_k is the bias term associated with the k-th feature map. M and N denote the height and width of the filter, respectively, and c is the index for the number of input channels. S indicates the stride and P is the padding.

To reduce the size of the output a pooling operation is often applied after the convolution (See figure 2.6). The most commonly used pooling operation is the max pooling operation, which computes the maximum value within a local window of the input. The max pooling operation can be represented mathematically as follows:

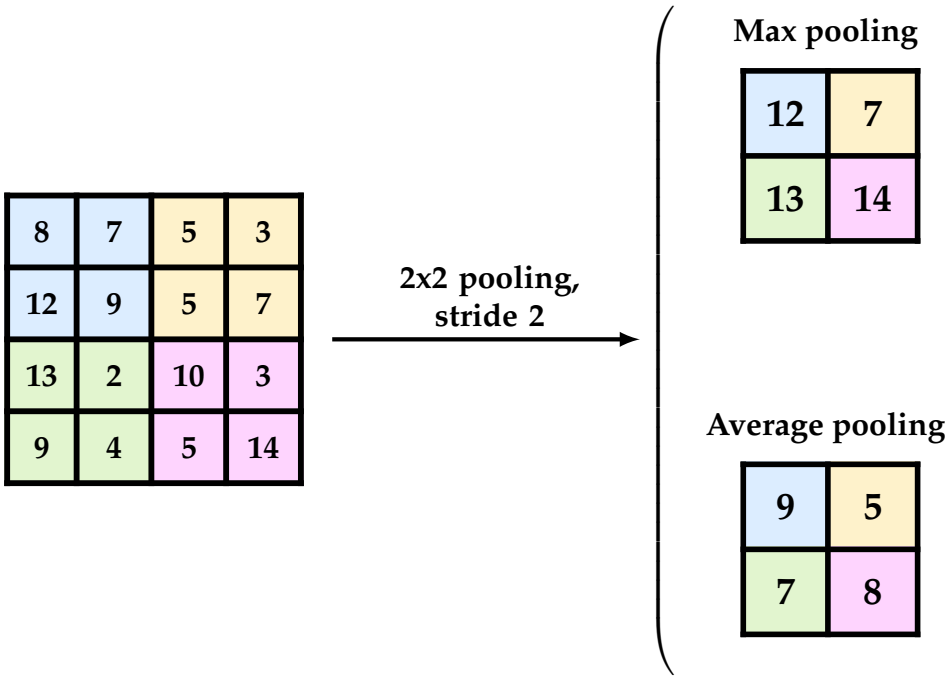

Figure 2.6: A pooling layer reduces the size and complexity of each feature map separately. For each unit in an output feature map, it takes the mean or the max of a fixed window in the corresponding input feature map.

$$y_{i,j,k} = \max_{m=0}^{H-1} \max_{n=0}^{W-1} x_{(i \cdot S+m),(j \cdot S+n),k} \tag{2.11}$$

where $y_{i,j,k}$ is the output value at position (i, j) of the k-th feature map, H and W are the height and width of the pooling window, respectively. m and n are the height and width indices of the pooling window, and S is the stride or the amount of shift between each pooling window.

In addition to the convolution and pooling operations, convolutional layers often include activation functions such as the ReLU function (See figure 2.3), which applies a non-linear transformation to the output of the layer.

The training process of a convolutional layer requires a proficient strategy to

gauge its performance on assigned tasks. This evaluation mechanism employs a loss function, which facilitates a comparison between the true output and the predicted output. The loss function generates a scalar value denoting the discrepancy, or "loss," hence guiding the training process to minimize this error.

The loss function serves as a guidepost during the process of training, indicating the direction and steps required for the model to improve. Commonly used loss functions include Mean Squared Error (MSE 2.7) for regression problems and Cross-Entropy Loss (2.8) for classification problems. The choice of the loss function depends largely on the type of problem at hand.

The training process also includes backpropagation (2.9), which uses the gradient of the loss function to update the weights and biases of the CNN. In essence, backpropagation operates by computing the gradient of the loss with respect to each weight and bias, and then adjusting these parameters in the direction that minimizes the loss. This iterative process is repeated over numerous epochs, or complete passes through the training dataset, until the model's performance ceases to significantly improve, or a preset number of epochs is reached.

2.1.2.2 3D Convolution Neural Network

3D Convolutional Neural Networks (3D CNNs) are a powerful extension of their 2D counterparts, specifically designed for processing volumetric data, such as 3D images or videos Ji et al. (2013). These networks are particularly useful in applications that involve the analysis of both spatial and temporal features, such as video classification and 3D object recognition. By applying 3D convolution operations, they are able to capture spatial correlations in three dimensions.

The primary mathematical formula involved in 3D CNNs is the 3D convolution

operation, which is an extension of the 2D convolution operation. This operation involves the use of 3D filters to process the input volume, capturing both spatial and temporal information. The 3D convolution operation can be represented as follows:

$$y_{i,j,k,p} = b_p + \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \sum_{l=0}^{L-1} \sum_{c=0}^{C-1} x_{(i \cdot S+m-P),(j \cdot S+n-P),(k \cdot S+l-P),c} \cdot w_{m,n,l,c,p} \tag{2.12}$$

Where $y_{i,j,k,p}$ is the output feature map at the spatial positions (i, j), temporal position k and for the p-th feature map. The filter weights are denoted by $w_{m,n,l,c,p}$, where indices m, n, and l iterate over the spatial and temporal dimensions of the 3D filter, and c is the index for the number of input channels. The input volume at each position and channel is represented by $x_{(i \cdot S+m-P),(j \cdot S+n-P),(k \cdot S+l-P),c}$. The bias term for the p-th feature map is b_p. S denotes the stride and P is the padding. This operation processes the input data in three dimensions, effectively capturing information across space, time, and multiple channels.

The history of 3D CNNs can be traced back to the work of Ji et al. (2013), who first introduced the concept of 3D convolutional networks for human action recognition in video analysis. Their pioneering work demonstrated the potential of 3D CNNs in capturing spatiotemporal features from videos, outperforming traditional methods that relied on handcrafted features and 2D CNNs.

Since the initial work of Ji et al., numerous research papers have explored and expanded upon the concept of 3D CNNs. One of the significant advancements in this field was the development of the C3D architecture by Tran et al. (2015). This architecture demonstrated the effectiveness of 3D CNNs in learning spatiotemporal features, leading to significant improvements in video classification performance

on standard benchmarks (Tran et al., 2015).

Another influential work in the field of 3D CNNs was the introduction of the I3D architecture by Carreira & Zisserman (2017). They proposed a novel approach that involved inflating 2D convolutional filters and pooling kernels into 3D, which allowed for the transfer of knowledge from pre-trained 2D CNNs to 3D CNNs. This led to significant improvements in action recognition performance and demonstrated the effectiveness of leveraging pre-trained networks for 3D CNNs.

In conclusion, 3D CNNs have emerged as a powerful tool for processing volumetric data and capturing spatiotemporal features. Their development has been driven by pioneering works like those of (Ji et al., 2013), (Tran et al., 2015), and (Carreira & Zisserman, 2017), which have paved the way for numerous applications in video analysis and 3D object recognition. The use of 3D convolution operations and the development of innovative architectures have enabled 3D CNNs to achieve impressive performance across various tasks.

2.1.3 Recurrent Neural Network

Recurrent Neural Networks (RNNs), an adaptive class of artificial neural networks, excel at modeling sequential data. RNNs boast a distinct architecture, enabling them to maintain hidden states, a memory form. This inherent ability to capture temporal dependencies makes RNNs highly suitable for a wide range of applications, from natural language processing and time-series prediction to speech recognition.

3D Convolutional Neural Networks (3D CNNs), while proficient in handling spatiotemporal data like video, lack the innate ability to capture long-term temporal dependencies (Tran et al., 2015), unlike RNNs. With convolutions across spatial

and temporal dimensions, 3D CNNs extract local features. Yet, their potential to model intricate temporal relationships is restricted. RNNs, on the other hand, outshine in managing sequences and extensive dependencies (Shi et al., 2015), making them the preferred choice for problems where time is a pivotal factor. Integrating memory into their design, RNNs gain a deeper understanding of sequential data, offering a context-rich and nuanced representation (Prakash et al., 2019).

As we venture into this section, we will unravel the complexities of Recurrent Neural Networks, shedding light on their architecture, functions, and applications. We will address critical aspects, such as RNN cell structure, and hidden state roles. Furthermore, we will explore RNN training challenges, including vanishing and exploding gradient issues, and delve into the solutions formulated to counteract these problems, like Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) cells.

2.1.3.1 Vanilla Recurrent Neural Network

Vanilla Recurrent Neural Networks (RNNs) are a class of artificial neural networks designed to process sequential data. These networks are particularly useful when dealing with time-series data, such as speech or text, as they can maintain an internal state that captures information about previous inputs. RNNs can be thought of as a chain-like structure (See figure 2.7), where each repeating module contains a hidden state and a weight matrix that updates the hidden state based on the input and the previous hidden state (Elman, 1990).

The core of a vanilla RNN can be expressed mathematically by the following equations:

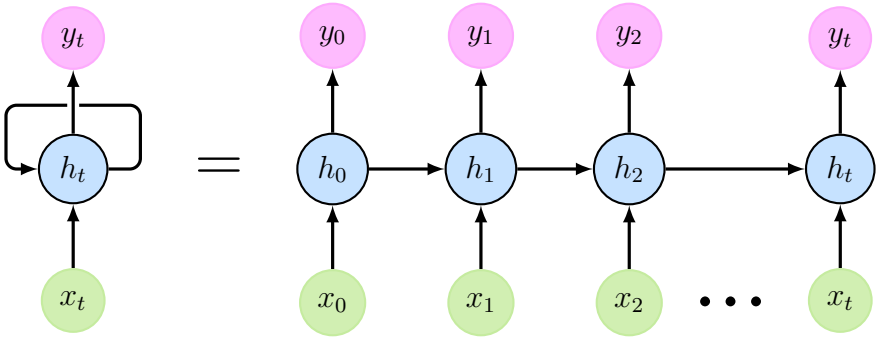

Figure 2.7: Illustration of a vanilla Recurrent Neural Network (RNN) depicting the hidden states $h_0, h_1, h_2, \ldots, h_t$, input sequence $x_0, x_1, x_2, \ldots, x_t$, and output sequence $y_0, y_1, y_2, \ldots, y_t$. The arrows represent the flow of information within the network."

$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t + b_h) \tag{2.13}$$

$$y_t = W_{hy}h_t + b_y \tag{2.14}$$

In these equations, h_t represents the hidden state at time t, x_t is the input at time t, and y_t is the output at time t. The weight matrices W_{hh}, W_{xh}, and W_{hy} connect the hidden state to itself, the input to the hidden state, and the hidden state to the output, respectively. The bias vectors b_h and b_y are added to the hidden state and the output, respectively. The $tanh$ function is used as the activation function for the hidden state, which squashes its output to the range of $(-1, 1)$ (Pascanu et al., 2013).

The history of RNNs can be traced back to the early 1980s when Hopfield networks were introduced (Hopfield, 1982). The concept of RNNs was later refined by Elman, who proposed the Simple Recurrent Network (SRN) in 1990 (Elman, 1990). RNNs have since been developed further, leading to the creation of more sophis-

ticated models such as Long Short-Term Memory (LSTM) networks (Hochreiter & Schmidhuber, 1997) and Gated Recurrent Units (GRUs) (Cho et al., 2014), which have been successful in addressing the vanishing gradient problem often encountered in vanilla RNNs.

RNNs have demonstrated remarkable success in various applications, including natural language processing, speech recognition, and time-series prediction (Prakash et al., 2019). As research in the field of deep learning continues, it is likely that new advancements and improvements in RNNs and other related models will be made, further enhancing their capabilities and expanding their range of applications.

2.1.3.2 Long Short-Term Memory Recurrent Neural Network

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network (RNN) architecture designed to address the vanishing gradient problem that occurs in standard RNNs (Hochreiter & Schmidhuber, 1997). LSTM networks are particularly effective at learning long-range dependencies in sequential data, such as text, speech, and time series. The key innovation of LSTM networks is the introduction of memory cells, which are capable of storing information for extended periods and selectively updating and retrieving it as needed.

The LSTM architecture consists of a memory cell c_t, an input gate i_t, a forget gate f_t, and an output gate o_t. The gates control the flow of information in and out of the memory cell (See figure 2.8). The LSTM cell equations are as follows:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{2.15}$$

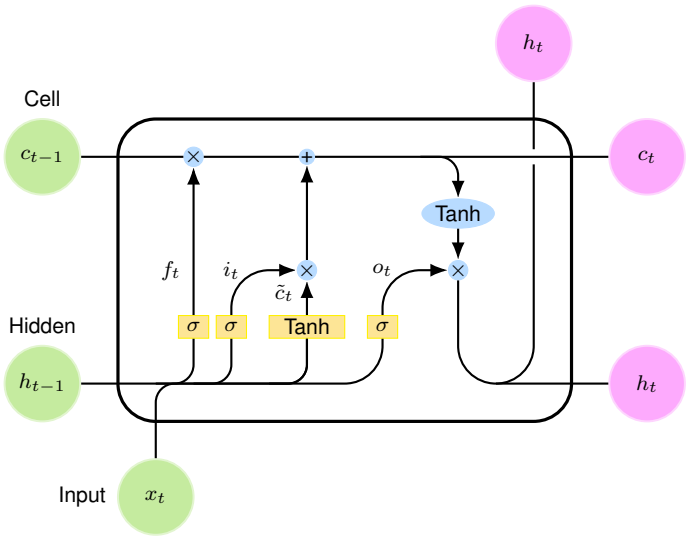

Figure 2.8: This diagram depicts an LSTM cell with input state x_t, cell state c_{t-1}, and hidden state h_{t-1}. The cell utilizes input, forget, and output gates (σ), as well as a Tanh activation function to update and generate new cell state c_t and hidden state h_t. The intricate structure of the LSTM cell allows for efficient handling of long-term dependencies in sequences.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2.16}$$

$$\tilde{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{2.17}$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}t \tag{2.18}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{2.19}$$

$$h_t = o_t * \tanh(c_t) \tag{2.20}$$

In these equations, σ represents the sigmoid function, $\tanh$ represents the hyperbolic tangent function, and $*$ represents element-wise multiplication. W_f, W_i, W_c, and W_o are weight matrices, and b_f, b_i, b_c, and b_o are bias vectors. The input gate i_t determines how much of the new information $\tilde{c}t$ should be added to the memory cell. The forget gate f_t decides which information to discard from the previous cell state c_{t-1}. The output gate o_t controls how much of the cell state c_t should be passed to the hidden state h_t.

LSTM networks were introduced by Hochreiter and Schmidhuber in 1997 to address the shortcomings of traditional RNNs in learning long-term dependencies (Hochreiter & Schmidhuber, 1997). Since then, LSTMs have been widely adopted in various applications (Gers & Schmidhuber, 2001; Sundermeyer et al., 2012).

Another development in LSTM research is the introduction of peephole connections (Gers & Schmidhuber, 2000). Peephole connections allow the gates to have direct access to the memory cell, which can improve the LSTM's ability to learn precise timing information.

Bidirectional LSTMs (BiLSTMs) are another modification to the standard LSTM architecture (Schuster & Paliwal, 1997). BiLSTMs process the input sequence in both forward and backward directions, enabling the model to capture information from both the past and the future context.

The attention mechanism, introduced by Bahdanau et al. (2014), has also been incorporated into LSTM architectures to improve their performance in tasks such as machine translation and text summarization. The attention mechanism allows the model to dynamically weigh different parts of the input sequence, enabling it to focus on the most relevant information for the current processing step.

LSTM networks have also been applied in combination with other deep learn-

ing techniques, such as convolutional neural networks (CNNs), for tasks like image captioning and video analysis (Donahue et al., 2015). By integrating LSTMs with CNNs, these hybrid models can process both spatial and temporal information, making them suitable for tasks that require an understanding of both visual and sequential data.

In recent years, the Transformer architecture (Vaswani et al., 2017) has gained significant attention as an alternative to RNN-based models, including LSTMs, for sequence-to-sequence tasks. Transformers rely solely on the attention mechanism and do not require recurrent connections, enabling them to process input sequences in parallel and scale more effectively. Despite the growing popularity of Transformer-based models, LSTMs continue to be a valuable tool for a wide range of applications.

In summary, LSTMs have served as a keystone in deep learning's evolution, empowering the proficient modeling of extensive dependencies within sequential data. Over time, numerous refinements and augmentations have emerged, bolstering LSTM architecture and adapting it to a variety of problem domains.

As we proceed to the subsequent subsection, we will explore Gated Recurrent Units (GRUs), an essential RNN variant. GRUs, with their more streamlined design compared to LSTMs, adeptly capture long-range dependencies. We will examine GRUs' distinct characteristics, benefits, and how they stack up against LSTMs in terms of efficacy and computational intricacy. By investigating GRUs, we aim to provide a comprehensive appreciation of the diverse strategies available for addressing sequential data and long-range dependencies within the realm of deep learning.

2.1.3.3 *Gated Recurrent Unit Neural Network*

Gated Recurrent Units (GRUs) are a type of recurrent neural network (RNN) architecture that was introduced by Cho et al. (2014) as a simpler alternative to Long Short-Term Memory (LSTM) networks (Cho et al., 2014). Like LSTMs, GRUs are designed to address the vanishing gradient problem and learn long-range dependencies in sequential data. GRUs have fewer parameters than LSTMs, making them computationally more efficient, while still delivering comparable performance on various tasks.

Figure 2.9 illustrate the GRU architecture which consists of a hidden state h_t and two gates: an update gate z_t and a reset gate r_t. The gates control the flow of information within the hidden state, determining how much of the previous state should be retained and how much of the new input should be incorporated. The GRU equations are as follows:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z) \tag{2.21}$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r) \tag{2.22}$$

$$\tilde{h}_t = \tanh(W_h \cdot [r_t * h_{t-1}, x_t] + b_h) \tag{2.23}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \tag{2.24}$$

In these equations, σ represents the sigmoid function, $\tanh$ represents the hyperbolic tangent function, and $*$ represents element-wise multiplication. $W_z, W_r,$

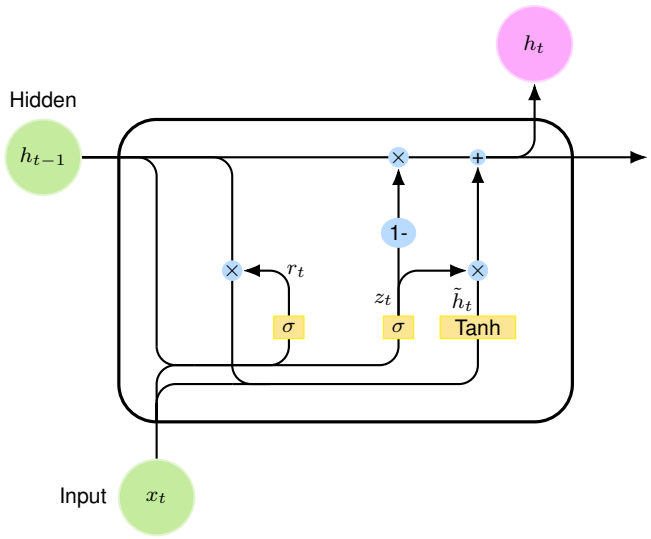

Figure 2.9: Gated Recurrent Unit (GRU) architecture illustrating the flow of information between input, hidden, and output states. The diagram displays the key components, including the update gate and reset gate, which control the flow of information between hidden states, and the Tanh activation function, which determines the candidate hidden state. The gates are regulated by sigmoid activation functions, while the hidden state is updated through a combination of element-wise multiplication and addition operations.

and W_h are weight matrices, and $b_z, b_r,$ and b_h are bias vectors. The update gate z_t determines how much of the previous hidden state h_{t-1} should be retained. The reset gate r_t decides how much of the previous hidden state should be considered when computing the new candidate hidden state $\tilde{h}_t$.

GRUs have been successfully applied in various applications (Chung et al., 2014). They have been shown to perform comparably to LSTMs on many tasks while being more computationally efficient due to their simpler architecture.

The GRU architecture has inspired several variations and extensions, such as the Bayesian GRU (Fortunato et al., 2017), which introduces Bayesian inference to the model for improved regularization and uncertainty estimation. Another example is the Minimal Gated Unit (MGU) (Zhou et al., 2016), which further simplifies the GRU architecture while maintaining its capability to capture long-range dependencies.

In conclusion, GRUs represent a significant advancement in RNN architectures, providing a more efficient alternative to LSTMs for modeling long-range dependencies in sequential data. They have been widely adopted in various applications and have inspired several variations and extensions. GRUs continue to be a valuable tool in the deep learning toolkit for processing sequential data.

2.1.4 Transformers Neural Network

2.1.4.1 Transformers Neural Network

Transformers Neural Networks have become a prominent and powerful architecture in the field of natural language processing and machine learning. Introduced by Vaswani et al. in their seminal paper "Attention Is All You Need" (Vaswani et al., 2017), this novel architecture was designed to address the limitations of recurrent

neural networks (RNNs) and convolutional neural networks (CNNs) in processing long-range dependencies in sequential data. The Transformer model relies on a mechanism called self-attention, which allows it to weigh the importance of different words or features in the input sequence, thereby enabling it to effectively capture the dependencies between words.

The core component of the Transformer architecture is the self-attention mechanism, which can be represented mathematically as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.25}$$

where Q, K, and V are the query, key, and value matrices, respectively, and d_k is the dimensionality of the key vectors. The softmax function ensures that the attention weights sum to one, and the scaling factor $\sqrt{d_k}$ is used to prevent the dot products from becoming too large in magnitude.

In the Transformer model, multi-head attention is used to allow the model to focus on different aspects of the input simultaneously. The multi-head attention mechanism can be expressed as:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h) \tag{2.26}$$

where each head_i is the result of the attention function applied to linearly projected versions of the input matrices:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{2.27}$$

and W_i^Q, W_i^K, and W_i^V are the learnable weight matrices for each head. The con-

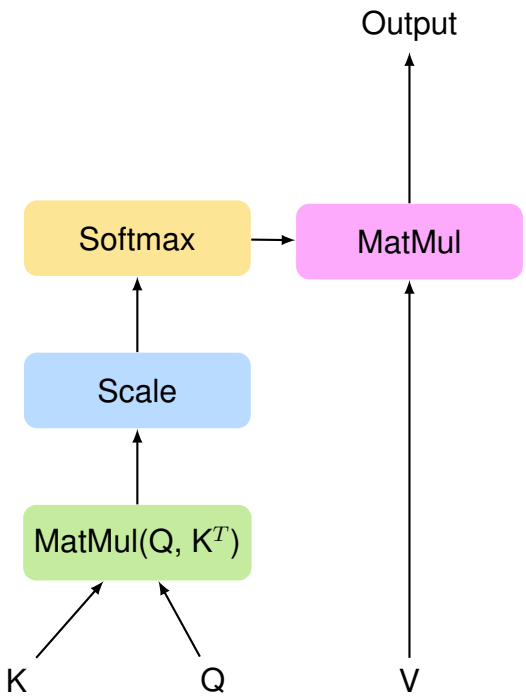

Figure 2.10: An in-depth illustration of the scaled dot-product attention method. The procedure commences with the input matrices Q (query), K (key), and V (value). The attention mechanism determines the compatibility of each query and key by executing matrix multiplication of Q and the transpose of K. Following that, each element is scaled by dividing it by the square root of d_k, which signifies the dimensions of the key, query, and value vectors. Afterward, the softmax function is employed on the scaled matrix to acquire attention weights. These weights are then utilized to carry out matrix multiplication with the V (value) matrix, ultimately generating the final result of the scaled dot-product attention technique.

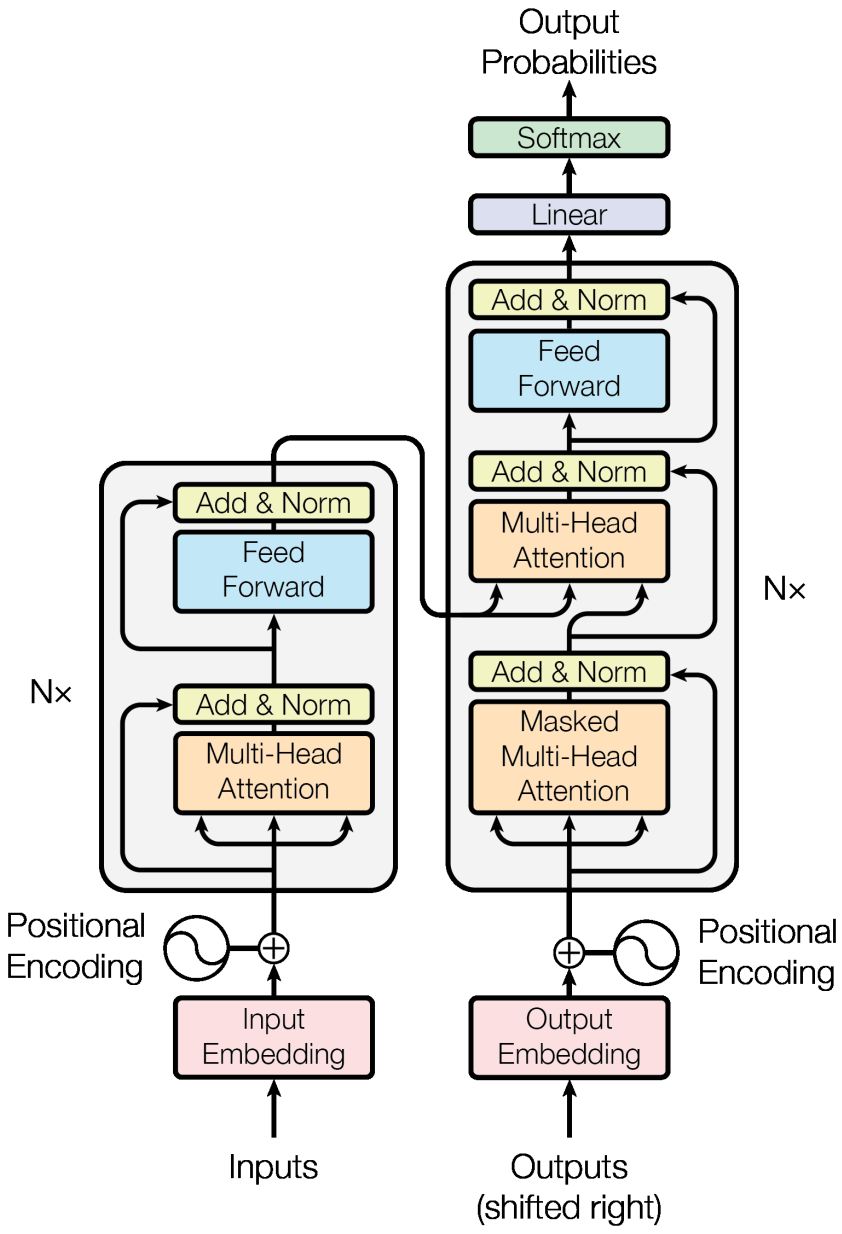

Figure 2.11: The Transformer Neural Network Architecture.

catenated heads (2.26) is then projected back to the original dimensionality using a linear transformation.

Figure 2.11 illustrate the architecture of transformers which consist of encoder and decoder blocks, each containing multiple layers. The encoder layers are composed of multi-head self-attention, followed by layer normalization, position-wise feed-forward networks, and another layer normalization. The decoder layers have an additional multi-head attention mechanism that attends to the output of the encoder.

The positional encoding is an essential component of the Transformer architecture, as it provides the model with information about the position of words in the input sequence. This is accomplished by adding sinusoidal functions to the input embeddings:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right) \tag{2.28}$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d}}}\right) \tag{2.29}$$

where pos is the position of the word in the sequence and i is the dimension within the encoding vector. The intuition behind this encoding is to help the model to easily learn relative positions of the words in the sequence as claimed by the authors of transformers (Vaswani et al., 2017). In order to understand these relative positions we follow a demonstration proposed by Timo Denk[1] where he explains the linear transformation between positions.

Let the function e induces a matrix $\boldsymbol{PE} \in \mathbb{R}^{n\times d}$, where each column vector $\boldsymbol{PE}_{pos,:}$ encodes the position $pos \in \{1, .., n\}$ in an input sequence of length n:

[1]https://timodenk.com/blog/linear-relationships-in-the-transformers-positional-encoding/

$$e(pos) = \boldsymbol{PE}_{pos,:} := \begin{bmatrix} \sin\left(\frac{pos}{f_1}\right) \\ \cos\left(\frac{pos}{f_1}\right) \\ \sin\left(\frac{pos}{f_2}\right) \\ \cos\left(\frac{pos}{f_2}\right) \\ \vdots \\ \sin\left(\frac{pos}{f_{\frac{d}{2}}}\right) \\ \cos\left(\frac{pos}{f_{\frac{d}{2}}}\right) \end{bmatrix} \tag{2.30}$$

where the frequencies are given by:

$$f_i = \frac{1}{\lambda_i} := 10000^{\frac{2i}{d}} \tag{2.31}$$

The authors of Vaswani et al. (2017) stated that a linear transformation $\boldsymbol{T}^{(k)} \in \mathbb{R}^{d \times d}$ exists, for which:

$$\boldsymbol{T}^{(k)} \boldsymbol{PE}_{pos,:} = \boldsymbol{PE}_{pos+k,:} \tag{2.32}$$

holds for any positional offset $k \in \{1, \ldots, n\}$ at any valid position $pos \in \{1, \ldots, n-k\}$ in the sequence. The matrix $\boldsymbol{T}^{(k)}$ can be defined as:

$$\boldsymbol{T}^{(k)} = \begin{bmatrix} \boldsymbol{\Phi}_1^{(k)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Phi}_2^{(k)} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Phi}_{\frac{d}{2}}^{(k)} \end{bmatrix} \tag{2.33}$$

where the $\frac{d}{2}$ transposed rotation matrices $\boldsymbol{\Phi}^{(k)}$ positioned on the main diagonal are defined by:

$$\mathbf{\Phi}_i^{(k)} = \begin{bmatrix} \cos(\lambda_i k) & \sin(\lambda_i k) \\ -\sin(\lambda_i k) & \cos(\lambda_i k) \end{bmatrix}^{\intercal} \tag{2.34}$$

Which proves the linear relationship between relative positions in the Transformer's positional.

The overall architecture of the Transformer can be summarized as follows: input embeddings with positional encoding are passed through a stack of identical encoder layers, followed by a stack of identical decoder layers, and finally, a linear layer and softmax function to produce the output probabilities.

One of the significant breakthroughs enabled by Transformers is the ability to pretrain large models on vast amounts of data, learning useful representations of language that can be fine-tuned for specific tasks. This idea was popularized by the BERT (Bidirectional Encoder Representations from Transformers) model Devlin et al. (2018), which demonstrated state-of-the-art performance on a wide range of NLP tasks.

Since the introduction of the Transformer architecture, numerous variants and improvements have been proposed. These include GPT (Generative Pre-trained Transformer) Radford et al. (2018), which focuses on unidirectional language modeling, and T5 (Text-to-Text Transfer Transformer) Raffel et al. (2020), which unifies various NLP tasks into a single text-to-text framework.

The impact of the Transformer architecture has been profound, leading to a revolution in natural language processing and the development of large-scale pretrained models that have set new state-of-the-art performance on a variety of NLP benchmarks. This has enabled significant progress in areas such as machine translation, text generation, question-answering, and sentiment analysis.

In conclusion, the Transformer architecture has revolutionized the field of natural language processing, enabling powerful and efficient models that can capture long-range dependencies in sequential data. The self-attention mechanism and multi-head attention allow the model to weigh the importance of different words or features in the input sequence, while the positional encoding provides information about the position of words. The ability to pretrain large models on vast amounts of data has led to significant breakthroughs in NLP, and the Transformer architecture has become the backbone of many state-of-the-art models.

2.1.4.2 Vision Transformers

Vision Transformers (ViT) have emerged as a powerful paradigm for computer vision tasks, initially inspired by the successes of Transformers in natural language processing Vaswani et al. (2017). The concept of self-attention, which was originally designed to process and understand text, has been effectively adapted for image understanding. ViT models have gained popularity due to their ability to scale up and generalize to a wide range of vision tasks, surpassing traditional convolutional neural networks (CNNs) in performance.

One of the first instances of Vision Transformers was introduced by Dosovitskiy et al. Dosovitskiy et al. (2021b), where they proposed a simple yet effective method for image classification. The core idea was to divide an image into fixed-size patches and linearly embed them as input tokens for a Transformer architecture. By doing so, the model could learn to attend to different parts of the image and capture both local and global contextual information.

The introduction of ViT sparked further research into the field, leading to advancements such as the Swin Transformer Liu et al. (2021a). This model employed

a hierarchical structure with shifted windows, which allowed it to maintain the advantages of Transformers while reducing computational complexity. The Swin Transformer achieved state-of-the-art results on various benchmarks, including image classification, object detection, and semantic segmentation tasks.

Another notable development in Vision Transformers is the DeiT (Data-efficient Image Transformer) Touvron et al. (2021). DeiT addressed one of the primary limitations of ViT: the reliance on vast amounts of labeled data for effective training. By incorporating knowledge distillation and regularization techniques, DeiT was able to achieve competitive results with significantly fewer training samples. This marked a crucial step towards making Vision Transformers more data-efficient and accessible for real-world applications.

Since the advent of Vision Transformers, research in this domain has continued to thrive. For instance, the CoaT (Co-Scale Conv-Attentional Image Transformers) model Xu et al. (2021) integrated convolutional layers into the Transformer architecture to further enhance its representation learning capabilities. This hybrid approach demonstrated improved performance and adaptability across various vision tasks. As the field progresses, we can expect to witness even more innovative architectures and approaches that leverage the strengths of both Transformers and convolutional networks for computer vision challenges.

2.1.5 Self-Supervised image classification

Self-supervised learning, a subset of unsupervised learning, has emerged as a potent paradigm for deep learning, particularly in the domain of image classification. It's an approach that leverages the inherent structure of the data to learn useful representations without the need for explicit labels (Jing & Tian, 2020). With the surge

of deep learning, the paradigm of self-supervised learning has undergone considerable evolution and refinement, gaining momentum over the past few years due to its potential for significant performance improvements.

The initial wave of self-supervised learning methods in deep learning was characterized by the development of pretext tasks, where the model learns to predict some aspect of the input data based on the rest (Doersch et al., 2015). For instance, an early method involved predicting the relative position of image patches, thereby encouraging the model to learn spatial hierarchies and object semantics. However, these approaches often required problem-specific tailoring and could struggle to transfer learned representations to downstream tasks.

Recent advancements in self-supervised learning have largely centered around contrastive learning, a method which trains models to identify similar and dissimilar instances (Chen et al., 2020a; He et al., 2020). This involves presenting the model with a pair of instances, such as two augmentations of the same image, and training it to recognize that they are similar while simultaneously distinguishing them from other instances. These contrastive methods have shown excellent results in various domains, with SimCLR Chen et al. (2020a) and MoCo He et al. (2020) among the most notable.

Bootstrapping Your Own Latent (BYOL) (Grill et al., 2020) and Distillation of self-training for Image Networks (DINO) (Caron et al., 2021) are two recent approaches that have further extended the boundaries of self-supervised learning. These methods deviate from the traditional contrastive learning paradigm, instead of using two views of the same image to train a student model and a teacher model. The two models interact and learn from each other, providing a rich, augmented view of the data.

In the BYOL framework (Grill et al., 2020), the two models are referred to as the "*online network*" and the "*target network*" as shown in figure 2.12. The online network, defined by a set of weights θ, consists of three stages: an encoder f_θ, a projector g_θ, and a predictor q_θ. The target network, on the other hand, uses a different set of weights ξ and provides the regression targets for training the online network. In other words, the online network is trained to make its output as close as possible to the output of the target network. The parameters ξ of the target network are updated as an exponential moving average of the online parameters θ. After each training step, this update is performed with a target decay rate $\tau \in [0, 1]$, following the equation $\xi \leftarrow \tau\xi + (1 - \tau)\theta$.

BYOL operates by producing two augmented views of an image, v and v'. The online network outputs a representation $y_\theta = f_\theta(v)$ and a projection $z_\theta = g_\theta(y_\theta)$ from the first augmented view v. The target network outputs $y'_\xi = f_\xi(v')$ and the target projection $z'_\xi = g_\xi(y'_\xi)$ from the second augmented view v'. The online network then generates a prediction $q_\theta(z_\theta)$ of z_θ. Both $q_\theta(z_\theta)$ and z'_ξ are ℓ_2-normalized. The loss function is defined as the mean squared error between the normalized predictions and target projections:

$$L_{\theta,\xi} = \left|\tilde{q}_\theta(z_\theta) - \tilde{z'}_\xi\right|_2^2 = 2 - 2 \cdot \frac{\langle q_\theta(z_\theta), z'_\xi \rangle}{||q_\theta(z_\theta)||_2 ||z'_\xi||_2} \tag{2.35}$$

The loss is symmetrized by separately feeding v' to the online network and v to the target network to compute $\tilde{L}_{\theta,\xi}$. At each training step, a stochastic optimization step is performed to minimize $L_{\text{BYOL}} = L_{\theta,\xi} + \tilde{L}_{\theta,\xi}$ with respect to θ only. This unique approach allows BYOL to learn rich image representations without the need for negative pairs, a common requirement in contrastive learning methods.

While BYOL and DINO both employ the dual model approach, they use dif-

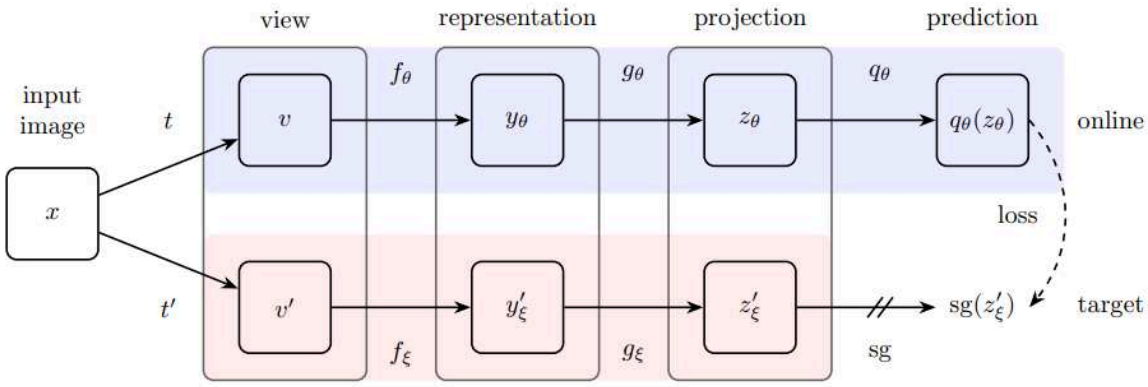

Figure 2.12: Bootstrapping Your Own Latent (BYOL)

ferent loss functions to train their models. In BYOL, a mean squared error loss is used to pull the student's representations towards a moving average of the teacher's representations (Grill et al., 2020). Conversely, DINO utilizes a distillation loss and a centering operation to ensure that the student's outputs stay close to the teacher's, while also preventing the outputs from collapsing to a single point (Caron et al., 2021).

An innovative addition to the self-supervised learning landscape is BEIT (Bao et al., 2022), a self-supervised vision representation model that uses a masked image modeling task to pre-train vision Transformers. BEIT encourages the model to predict masked-out portions of the input image, promoting the learning of rich, contextual representations that have been demonstrated to transfer well to downstream tasks.

EsViT (Li et al., 2021) and SelfSupSITSClassif Yuan & Lin (2021a) are other remarkable contributions to the field of self-supervised learning with vision Transformers. These works further illustrate the effectiveness and versatility of self-supervised learning methods, particularly when combined with the powerful, attention-based architecture of Transformers.

In summary, self-supervised learning has evolved significantly over the years, moving from the early days of pretext tasks to the current era of contrastive learning and beyond. The flexibility and performance potential of these methods make them an exciting area of research, with many opportunities for future exploration and advancement.

2.2 UTILIZING DEEP LEARNING FOR IMAGE SEQUENCE CLASSIFICATION

2.2.1 Image Sequence Classification through Convolutional Neural Networks

Image Sequence Classification via Convolutional Neural Networks (CNNs) plays a crucial role in understanding and interpreting temporal visual data, such as videos or a series of images. It involves the application of CNNs, renowned for their robustness in extracting spatial features from static images, to sequences of images, thereby capturing not only spatial but also temporal dynamics. This technique is extensively employed in numerous real-world applications ranging from action recognition in videos, surveillance monitoring, to patient monitoring in healthcare (Frizzi, 2016; Li, 2022; Vrskova et al., 2022), where recognizing patterns over a sequence of images is paramount.

In the work by Frizzi (2016), a method for fire and smoke detection was developed using a deep learning approach, particularly a convolutional neural network (CNN). The proposed method demonstrates high accuracy in fire identification within video sequences, outperforming traditional video fire detection methods. The CNN facilitates integrated feature extraction and classification, leading to an efficient, single-architecture solution.

Nevertheless, the current model exhibits some limitations, primarily due to its

processing of video input frame by frame, since conventional CNNs handle only 2D inputs. Future work aims to implement a 3D CNN, which can extract features from both spatial and temporal dimensions, thus encoding the motion information of fire and smoke. This change could further decrease time cost.

In their research, Li (2022) leveraged digital multimedia resources, primarily dance art videos, to automate scene classificationa fundamental aspect of scene semantic-based video content retrieval. They utilized a deep convolutional neural network (CNN), optimized using a differential evolution algorithm, for the classification task.

The proposed method begins by using the Canny operator in the YCbCr color space to detect human silhouettes in video keyframes. Subsequently, an AdaBoost algorithm with a cascade structure is employed to track and label human targets, along with the construction and updating of weak classifiers.

In the next stage, a differential evolution algorithm optimizes the structural parameters of the CNN, while an adaptive strategy enhances the accuracy of the optimization solution. The resultant optimized deep CNN is then used to train the labeled videos to achieve consistent scene classification results.

Experimental results confirmed that the proposed method, especially when the crossover rate of differential evolution and the CNN's convolutional kernel size are judiciously set, results in high scene classification performance. These results, characterized by high accuracy and low root-mean-square error, validate the approach's efficacy in dance art scene classification.

The study conducted by Vrskova et al. (2022) highlights the burgeoning interest in applying neural networks to diverse scientific, academic, and industrial fields, with video classification being a primary focus. Recognizing that object detection

from videos presents more challenges than single images due to the temporal continuity constraint, they propose a 3D Convolutional Neural Network (3DCNN) for detecting human activity from video data.

The team leveraged conventional neural networks like ConvLSTM and other 3DCNNs for object detection from video, but their proposed 3DCNN outperformed these, demonstrating better results for motion, static, and hybrid features. It achieved a remarkable recognition precision of 87.4%, significantly higher than the precisions attained by other neural network architectures, which were 65.4%, 63.1%, and 71.2%.

Furthermore, the proposed 3DCNN exhibited better performance than previous 3DCNN architectures on the UCF YouTube Action dataset. The same architecture, when applied to modified and full UCF101[2] datasets and the full UCF50[3] dataset, also showed strong performance, yielding overall precisions of 82.7%, 78.5%, and 80.6%, respectively. These results emphasize the potential and effectiveness of the proposed 3DCNN for human activity recognition.

The field of image sequence classification using convolutional neural networks (CNNs) continues to advance rapidly, as demonstrated by the three state-of-the-art approaches discussed in this subsection. Despite the distinctive advantages each method presents, they all converge on a singular theme: using deep learning and Convolutional Neural Networks to enhance image sequence classification. These models, whether focusing on fire and smoke detection (Frizzi, 2016), dance art video scene classification (Li, 2022), or human activity recognition (Vrskova et al., 2022), offer innovative solutions that outperform their predecessors. Future research directions should include the integration of 3D CNNs to take advantage of

[2]https://www.crcv.ucf.edu/data/UCF101.php
[3]https://www.crcv.ucf.edu/data/UCF50.php

both spatial and temporal information.

2.2.2 Implementing Recurrent Neural Networks for Image Sequence Classification

In the ever-evolving landscape of computer vision, a prominent and burgeoning research area is the classification of image sequences, where Recurrent Neural Networks (RNNs) have emerged as a powerful tool. As the very nature of image sequences involves temporal dependencies, RNNs, designed to process sequential data, are particularly well-suited for this task. Unlike their CNN counterparts that primarily handle spatial features, RNNs excel in capturing temporal dynamics and dependencies within sequential images, such as video footage or time-series medical scans. The following section aims to explore the state-of-the-art methodologies for image sequence classification using RNNs. We will delve into recent advancements and applications, highlighting how these networks have been optimized and applied in different contexts to achieve superior performance.

In the work of Ng et al. (2015), two video classification methods were proposed that aggregate frame-level CNN outputs into video-level predictions. The first approach employs feature pooling, which applies max-pooling to local information across time. The second utilizes Long Short-Term Memory (LSTM), whose hidden state evolves with each successive frame. Both methods are underpinned by the notion that integrating information from longer video sequences facilitates superior classification performance. In contrast to earlier approaches that were limited to training on mere seconds of video, these innovative techniques leverage up to two minutes, or 120 frames, to achieve superior outcomes. Furthermore, they demonstrated the necessity of motion information, particularly optical

flow, for achieving state-of-the-art results on the UCF-101 benchmark. However, they noted that optical flow might not always be beneficial, especially with uncontrolled, 'wild' videos, as in the Sports-1M dataset. In such cases, sophisticated sequence processing architectures like LSTM are necessary.

Yang et al. (2017) proposed an innovative approach to circumvent a significant challenge faced by Recurrent Neural Networks (RNNs) handling very high-dimensional inputs. While RNNs have shown considerable promise in sequence modeling tasks, such as Natural Language Processing, they are often impractical and difficult to train when confronted with high-dimensional inputs, as they necessitate a large input-to-hidden weight matrix. This complication has possibly hindered RNNs' wide-scale application in tasks like video modeling, where current solutions involve reducing the input dimensions via various feature extractors. The authors, however, have proposed a more general and efficient method, which factorizes the input-to-hidden weight matrix using Tensor-Train decomposition, trained concurrently with the weights. Despite being orders of magnitude less complex, their model achieves competitive performances with state-of-the-art models on classification tasks using multiple real-world video datasets. This approach, therefore, presents a significant contrast to the methods proposed by Ng et al. (2015), wherein RNNs and LSTMs are employed without addressing the issue of high-dimensionality. The approach of Yang et al. (2017) offers a new building block for modeling high-dimensional sequential data with RNNs and paves the way for transferring sophisticated architectures from other domains to high-dimensional sequential data modeling.

Zhu (2019) proposed a novel recurrent architecture, FASTER, for video-level classification that leverages the correlation between adjacent video clips, thereby

significantly enhancing computational cost efficiency, particularly during the aggregation stage. Traditional video classification methods often treat highly-correlated short clips as independent units, neglecting the temporal structure and necessitating high computational costs as each clip must be processed individually. The proposed FASTER architecture combines high-quality, computationally expensive representations of clips, which capture actions in detail, with lightweight representations that detect scene changes in the video, thereby avoiding redundant computations. Moreover, they introduce a novel processing unit, FAST-GRU (based on Gated Recurrent Unit), to learn the integration of clip-level representations and their temporal structure. This innovative approach greatly improves the FLOPs vs. accuracy trade-off during inference, achieving a ten-fold reduction in FLOPs while maintaining similar accuracy on popular datasets like Kinetics, UCF101, and HMDB51, when compared to existing methods.

In conclusion, recurrent neural networks (RNNs) demonstrate significant potential in the realm of image sequence classification, with advancements in dealing with high-dimensional inputs, temporal sequence information, and computational cost efficiency at the aggregation stage. Various approaches have demonstrated the benefits of incorporating temporal sequence data in the processing architecture, such as Ng et al. (2015)'s use of Long Short-Term Memory networks (LSTM), Yang et al. (2017)'s Tensor-Train decomposition for handling high-dimensional inputs, and Zhu (2019)'s FASTER architecture leveraging clip correlations for computational efficiency. These evolving methods attest to the dynamism and versatility of RNNs in image sequence classification tasks.

2.2.3 Image sequence classification Using Transformers

Image time series classification using Transformers has seen significant progress in recent years. Chen et al. (2019) introduced VideoBERT, a joint model for video and language representation learning using Transformer-based architecture . VideoBERT outperforms previous methods by learning representations from video clips and their associated captions, allowing it to excel in several downstream tasks.

In addition to VideoBERT, Zhou et al. (2021) proposed ViViT, a Video Vision Transformer specifically designed for video processing . ViViT leverages a spatiotemporal attention mechanism to capture both spatial and temporal information in videos, achieving state-of-the-art performance on multiple benchmark datasets.

Another approach is TransVG by Chen et al. (2021), an end-to-end framework for visual grounding using Transformers to jointly encode image features and natural language queries. TransVG demonstrates its effectiveness by outperforming previous state-of-the-art methods on several benchmark datasets.

Furthermore, Yuan & Lin (2021b) proposed a novel self-supervised pretraining scheme for Transformers in satellite image time series (SITS) classification to address overfitting when labeled data are scarce. By leveraging the inherent temporal structure of satellite time series, the pretrained network can be adapted to various SITS classification tasks with improved generalization performance.

Moreover, Lin et al. (2019) proposed the Temporal Shift Module (TSM) for efficient video understanding, which uses a modified version of the Transformer architecture to process temporal information in videos. The TSM achieves state-of-the-art performance on various action recognition benchmarks with fewer computations than previous methods. This collection of research papers demonstrates the growing interest and success in using Transformers for image time series clas-

sification and video understanding.

In conclusion, recent advancements in image time series classification using Transformers have led to innovative methods such as VideoBERT, ViViT, SITS-BERT, and TransVG. These approaches have effectively captured spatial and temporal information in videos, achieving state-of-the-art performance on multiple benchmark datasets. As the field continues to progress, the integration of Transformers with other architectures and the development of new techniques will likely further enhance video understanding and related applications. However, it is worth noting that Transformers have not yet been applied to astronomical image time series, which presents a potential avenue for future research. Building on these advancements, in our current work, we have chosen to adopt these potent and contemporary methodologies, specifically harnessing the capabilities of Transformers, for the classification of astronomical image time series. Through their ability to process sequential data, Transformers have potential to understand and interpret complex temporal changes in these astronomical images, which often contain subtle variations. The insights gained from this application could be instrumental in broadening our understanding of the cosmos, thereby opening up exciting new frontiers in astronomy.

CHAPTER 3

The Role of Big Data in Modern Astronomy

3.1 ASTRONOMY

Astronomy, the study of celestial objects and phenomena beyond Earth's atmosphere, has been an essential field of human inquiry since ancient times. Delving into the vastness of the cosmos, astronomy seeks to understand the intricate patterns and behaviors of stars, planets, galaxies, and other celestial bodies. This discipline not only inspires wonder and curiosity but also provides us with valuable insights into the origins and evolution of the universe. Moreover, it fosters technological advancements and drives innovation in various sectors, from satellite communications to navigation systems. By pushing the boundaries of our knowledge, astronomy plays a critical role in shaping our perspective of the cosmos and our place within it, ultimately enriching our understanding of the universe and our own existence. In the following chapter, we will explore the exciting intersection of astronomy and machine learning, discussing the potential applications and benefits of leveraging these cutting-edge technologies in the study of the cosmos.

3.1.1 Observational Cosmology

Observational cosmology studies astrophysical objects or phenomena, often referred to as probes, which grant us a window into the universe's earliest moments. By observing these probes, we can determine the universe's constituents and trace its evolution over time. The key to understanding this evolution, or dynamics, is by accessing measures of distance and time.

Time is discerned through redshift, which provides insights into the universe's

scale. This scale is directly associated with the time elapsed since the Big Bang, otherwise known as cosmic time.

As for distance, it is gauged through the use of standard rulers, found within the vast structures of the cosmic web, or standard candles.

In this manuscript, our attention is particularly drawn towards standard candles, especially those derived from stellar explosions that mark the end of their lifecycle, commonly known as supernovas.

However, it is essential to note that only a specific class of supernova exhibits the standard candle property. This makes it crucial to differentiate them from other objects that could potentially cause confusion.

This leads us to the process of observing these astrophysical objects or cosmic probes. Large surveys are conducted, meaning large areas of the sky are observed using increasingly precise telescopes. The observation is carried out through spectroscopic or photometric techniques (using filters), and we can incorporate the temporal dimension by arranging observation sequences that produce time series.

3.1.2 Cosmology and Astronomy in The Era of Big Data

The field of cosmology and astronomy has been transformed by the emergence of big data. The remarkable growth in both the quantity and complexity of data has drastically enhanced our comprehension of the universe. Observatories worldwide, such as the Vera C Rubin Legacy Survey of Space and Time (Rubin/LSST) (Ivezić et al., 2019) and the Square Kilometre Array (SKA) (Lazio, 2009), will start operating in the next few years, and it will generate terabytes of data every night, presenting unique opportunities to examine the cosmos with unparalleled precision. To effectively process, analyze, and interpret the information within these

immense datasets, the development of advanced algorithms and computational tools is essential (Zhang & Zhao, 2015).

A primary challenge in this field is extracting significant information from the vast amounts of data being collected. While conventional statistical methods are still widely used, they are beginning to show their limits in managing the complexity and scale of contemporary astronomical datasets. As a result, researchers have turned to machine learning algorithms, which have shown exceptional success in handling large-scale and complex data. These algorithms have the potential to expose previously unknown relationships and patterns in data, ultimately leading to new discoveries and a more profound understanding of the universe.

As data volumes continue to grow, astronomers are placing increasing emphasis on the use of machine learning techniques for time series analysis (Pasquet et al., 2019; Qu et al., 2021; Möller & de Boissière, 2020). This type of data, which includes astronomical image time series (Carrasco-Davis et al., 2019; Gómez et al., 2020), offers insights into the temporal development of astronomical objects and phenomena. Such data is vital for understanding transient events, variable stars, and other time-dependent processes. Machine learning has been applied to various aspects of time series analysis in astronomy, encompassing feature extraction, classification, and anomaly detection.

A key objective of our work is to contribute to the integration of machine learning in astronomy by developing a solution for astronomical image time series classification. This task is essential in contemporary astronomy, as it enables the identification and study of time-dependent phenomena, ultimately fostering a more comprehensive understanding of the universe. By creating and applying advanced machine learning techniques, we aim to enhance the accuracy and efficiency of as-

tronomical image time series classification, fully unlocking the potential of the big data era in astronomy.

3.1.3 Basics of astronomy

Astronomy seeks to understand the universe's origins, evolution, and properties by studying celestial objects and phenomena. One essential concept is **apparent magnitude**, which gauges the brightness of celestial objects as seen from Earth. The equation for apparent magnitude (m) is:

$$m = -2.5 \log_{10} \frac{F}{F_0} \tag{3.1}$$

Where F represents the observed flux, and F_0 is a reference flux. A difference of 5 magnitudes corresponds to a 100-fold change in brightness.

Absolute magnitude (M) is another crucial concept, measuring an object's intrinsic brightness. Unlike apparent magnitude, absolute magnitude calculates the object's brightness if it were 10 parsecs[1] away from the observer. The equation relating apparent magnitude, absolute magnitude (Hughes, 2006), and distance (d in parsecs) is:

$$m - M = 5 \log_{10}(d) - 5 \tag{3.2}$$

Knowing an object's apparent and absolute magnitudes allows distance estimation.

Redshift (Hubble & Tolman, 1935), where an astronomical object's observed light wavelength shifts toward the red (longer) end of the electromagnetic spec-

[1] 1pc = 3.2616 light-years
[2] https://ssb.stsci.edu/cdbs/calspec_ascii_review/p330e_stisnic_003.ascii

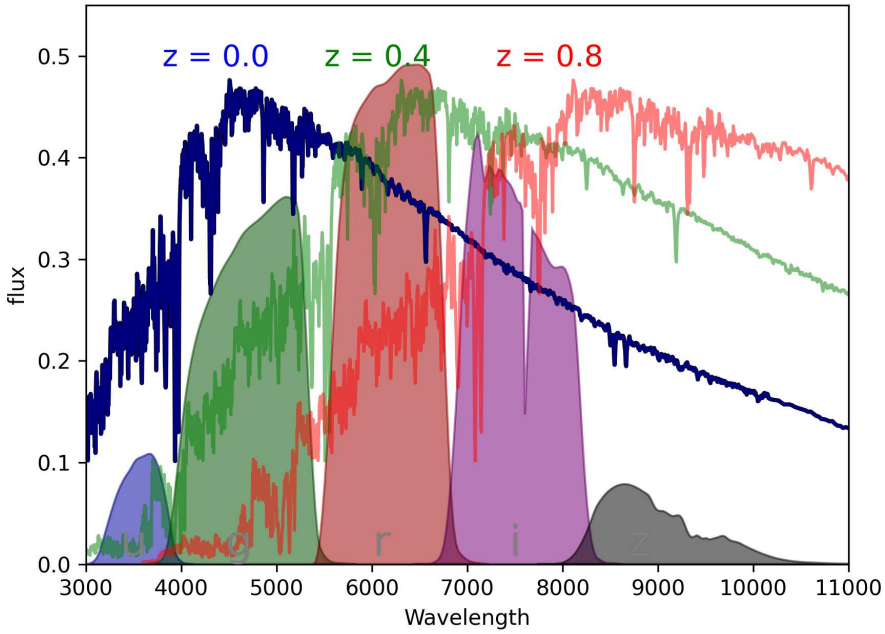

Figure 3.1: This figure presents the Sloan Digital Sky Survey (SDSS) broadband filters u, g, r, i, and z superimposed on the reference spectrum of the star P330E[2]. The filters are shown in their respective colors, while the star's spectrum is redshifted at three different values (z = 0.0, 0.4, and 0.8) to demonstrate the effect of redshift on the observed spectrum. The redshifted spectra are displayed with varying transparency, and the filters help highlight the spectral shifts. The plot emphasizes the importance of accounting for redshift when studying astronomical objects at different distances.

trum (See figure 3.1), occurs due to the Doppler effect as the emitting object moves away[3]. Redshift (z) is defined as:

$$z = \frac{\Delta\lambda}{\lambda_e} \tag{3.3}$$

With $\Delta\lambda$ being the change in wavelength and λ_e being the emitted wavelength.

[3]In fact this is a simplified description, the true physical effects comes from both velocity in a comoving frame and from the universe expansion but as a first approximation and at low redshift value it can be approximate by a Doppler effect.

Redshift helps on measuring the universe's expansion and distant galaxy distances.

Hubble's Law (Hubble, 1929), a foundational cosmology observation, relates galaxy recession velocity (v) to their distance (d) through:

$$v = H_0 d \tag{3.4}$$

Here, H_0 represents the Hubble constant, approximately 70 km/s/Mpc Riess et al. (1998); Perlmutter et al. (1999). Hubble's Law estimates the observable universe's age and size while providing expansion evidence.

Redshift and Hubbles Law connections are essential for determining distances to faraway galaxies. By measuring a galaxys redshift, we can infer its recession velocity, and using Hubbles Law, we can then estimate the galaxys distance. This relationship between redshift and distance aids astronomers in studying the universes large-scale structure and evolution. However, at larger cosmological distances, or in terms of cosmic time, the straightforward Hubble Law is replaced by a more complex model, the ΛCDM model.

The Lambda-Cold Dark Matter (ΛCDM) model (Einstein, 1916; Guth, 1981), the current cosmology standard, describes the universe's large-scale structure based on the assumption that the universe comprises cold dark matter (CDM) and dark energy, represented by the cosmological constant Λ. The ΛCDM model, in the context of an expanding universe described by Hubble's Law, offers a framework for understanding galaxy formation, matter distribution, and expansion history up to the cosmic time just after the big bang.

In the ΛCDM model, the universe's expansion results from the competition between matter's gravitational pull, including dark matter, and dark energy's repulsive force. As the universe expands, matter density decreases, while dark energy

density remains roughly constant, accelerating the universe's expansion in recent (1 G year)[4] time, an observation derived from studying distant supernovae (Riess et al., 1998; Perlmutter et al., 1999; Kirshner, 1999). The accelerated expansion, as described by the ΛCDM model, has significant implications for understanding the universe's large-scale structure and evolution.

The ΛCDM model's primary success lies in its ability to predict the observed galaxy distribution in the universe (Blanchard et al., 2023). The model posits that galaxies form within dark matter halos, with dark matter's gravitational pull driving galaxy formation and growth. Observations of galaxy clustering[5] and cosmic microwave background radiation(CMB)[6] (Penzias & Wilson, 1965) have strongly supported the ΛCDM model's predictions.

Using redshift, magnitude measurements, and considering the ΛCDM model concepts enables astronomers to investigate the universe's history and evolution. By studying distant galaxies and their redshifts, astronomers can trace the universe's expansion history, which in turn informs our understanding of the underlying physical processes driving cosmic evolution. This has led to the establishment of a coherent picture of the universe, in which dark matter and dark energy play crucial roles.

The study of distant supernovae has played a critical role in the development of the ΛCDM model. By measuring the apparent magnitudes of Type Ia supernovae, whose absolute magnitudes are known to be constant (or quasi constant, they are

[4] 1 Gyr, or 1 gigayear, is a unit of time equal to one billion years. It's often used in astronomy and cosmology to describe vast time scales, such as the age of the universe or the time spans involved in the evolution of galaxies

[5] Refers to the observation that galaxies are not randomly distributed throughout the universe but tend to group together, forming structures such as clusters and superclusters.

[6] The CMB is the remnant radiation from the early universe, dating back to approximately 380,000 years after the Big Bang when the universe became transparent to light.

called standardizable candle), astronomers can determine their relative distances. The redshifts of these supernovae can then be used to determine determine the cosmic time of their explosion. The relationship between the distance and the redshift (so cosmic time) provided the first observational evidence for the accelerated expansion of the universe (Riess et al., 1998; Perlmutter et al., 1999; Kirshner, 1999).

In conclusion, the ΛCDM model is integral to our comprehension of the large-scale structure, evolution, and ultimate destiny of the universe. The need to create an effective method for detecting supernovae from astronomical image time series is imperative. This would greatly enhance our ability to gauge the distances to far-off galaxies and track the universe's expansion history. Through precise detection and classification of supernovae in astronomical image time series, we can improve our insight into the fundamental physical processes that drive cosmic evolution, as well as the roles of dark matter and dark energy. This will enable astronomers to decode the enigmas of the cosmos, further enriching our collective understanding in the field of astronomy.

3.1.4 Astronomical objects and cosmological probes

3.1.4.1 Cosmological probes

Cosmology, the study of the origin and evolution of the universe, relies on the observation and analysis of various astronomical probes to develop an understanding of the cosmos. One of the most crucial astronomical probes in cosmology is the Cosmic Microwave Background (CMB) radiation (Penzias & Wilson, 1965). The CMB is a nearly uniform, isotropic radiation that fills the universe, believed to be the remnant of the Big Bang. Observations of the CMB's anisotropies[7] and temper-

[7] Anisotropies, in the context of the CMB, refer to the small variations in temperature and density across the otherwise uniform and isotropic radiation field. While the CMB is incredibly ho-

ature fluctuations have offered invaluable insights into the early universe's composition, density, and geometry.

Another essential astronomical probe for cosmology is the large-scale distribution of galaxies. By mapping the positions and velocities of galaxies across the universe, researchers can study the underlying structure of the cosmos and the impact of dark matter and dark energy on its evolution (Springel et al., 2005). Observations of galaxy clusters and superclusters have led to the discovery of the cosmic web, a filamentary structure composed of dark matter and galaxies, which has furthered our understanding of the universe's formation and large-scale dynamics.

Supernovae, particularly Type Ia supernovae, are also vital tools in cosmology due to their role as "standard candles." Since Type Ia supernovae have a consistent peak luminosity, they can be used to measure cosmic distances accurately (Riess et al., 1998; Perlmutter et al., 1999). Observations of distant Type Ia supernovae led to the groundbreaking discovery of the universe's accelerated expansion, which in turn led to the introduction of the concept of dark energy, a mysterious force responsible for driving the expansion.

Lastly, gravitational waves, ripples in spacetime caused by the acceleration of massive objects, have emerged as a new window into the cosmos (Weinstein, 2016). The detection of gravitational waves from merging black holes and neutron stars[8] has provided a wealth of information about the nature of these extreme events and their implications for cosmology. As gravitational wave observatories become

mogeneous, with a nearly constant temperature of about $2.725K$, there are tiny fluctuations in temperature on the order of $10^{-5}K$.

[8]Neutron stars are the incredibly dense remnants of massive stars that have undergone a supernova explosion at the end of their life cycles. They are called neutron stars because they are primarily composed of neutrons

more sensitive, they will allow researchers to probe even deeper into the universe's history and potentially unveil new insights into its structure and evolution.

In conclusion, cosmological probes such as the Cosmic Microwave Background, galaxy distributions, Type Ia supernovae, and gravitational waves play a crucial role in advancing our understanding of the universe's origins, evolution, and underlying structure. Astronomical image time series are particularly important for the discovery and analysis of Type Ia supernovae. By monitoring changes in the brightness of celestial objects over time, researchers can identify and study Type Ia supernovae. The identification of these events has been instrumental in uncovering the accelerating expansion of the universe and the enigmatic presence of dark energy. As our observational capabilities continue to improve, we can anticipate further breakthroughs in cosmology, driven by the ongoing study of these astronomical objects.

3.1.4.2 Transient objects

Transient objects in astronomy refer to celestial phenomena that exhibit significant changes in their brightness or position over relatively short timescales, ranging from seconds to months. These events often result from highly energetic processes or the interaction of celestial bodies. The study of transient objects has contributed to our understanding of the dynamic and violent nature of the universe and has led to numerous breakthroughs in astrophysics.

One well-known example of a transient object is a supernova, the explosive death of a massive star, or the thermonuclear explosion of a white dwarf[9] in a bi-

[9] A white dwarf is a compact stellar remnant that forms after a low-to-intermediate-mass star (roughly 0.5 to 8 times the mass of the Sun) exhausts its nuclear fuel and evolves through its life cycle.

nary system[10]. Supernovae are some of the most luminous events in the universe, briefly outshining entire galaxies before fading away over weeks or months (Filippenko, 1997). The observation and study of supernovae have been crucial in determining the expansion rate of the universe, the production of heavy elements, and the life cycles of stars.

Another class of transient objects is gamma-ray bursts (GRBs), which are short-lived, highly energetic bursts of gamma-ray photons. These bursts can last anywhere from milliseconds to several minutes and are thought to result from the core-collapse of massive stars or the merger of neutron stars (Gehrels & Razzaque, 2013). GRBs are among the most energetic events in the universe and can provide insights into the behavior of matter and energy under extreme conditions, as well as the distribution of galaxies in the early universe.

Fast Radio Bursts (FRBs) are another example of transient astronomical phenomena. These events are characterized by millisecond-duration pulses of radio emission and are of extragalactic origin. The exact nature of FRBs is still largely unknown, but they are believed to be associated with highly energetic processes involving neutron stars or other compact objects (Lorimer et al., 2007). The study of FRBs can potentially shed light on the properties of the intergalactic medium[11], as well as the distribution and evolution of matter throughout the universe.

In addition to supernovae, gamma-ray bursts, and fast radio bursts, there are numerous other transient objects and phenomena that enrich our understanding of the universe. These include novae, which are thermonuclear explosions on the

[10] A binary system, in the context of astronomy, refers to a system of two celestial objects, typically stars, that are gravitationally bound to each other and orbit around their common center of mass.

[11] The intergalactic medium (IGM) is the diffuse gas that exists between galaxies in the vast expanses of the universe. It is primarily composed of hydrogen and helium, along with trace amounts of heavier elements.

surface of white dwarfs in binary systems; X-ray transients, which are sources of highly variable X-ray emission associated with black holes or neutron stars accreting material from a companion star; and tidal disruption events, which occur when a star is disrupted by the tidal forces of a supermassive black hole. Other transient phenomena include flare stars, which are low-mass stars exhibiting sudden increases in brightness due to magnetic activity; and microlensing events, caused by the gravitational lensing of background stars by foreground objects, such as planets or brown dwarfs. The study of these diverse transient objects continues to unveil new insights into the dynamic processes shaping the cosmos, and advances in observational capabilities promise to reveal even more phenomena in the future.

3.1.4.3 Supernovae Types

Supernovae represent some of the most energetic and influential events in the universe, signifying the explosive death of certain types of stars or the result of thermonuclear reactions in binary systems. These cataclysmic explosions have fascinated astronomers for centuries (See the figure 3.2), with the earliest recorded supernova observations dating back to ancient China in 185 A.D (Stephenson & Green, 2003). The term "supernova" was introduced in the 1930s by Swiss astrophysicist Fritz Zwicky, who, in collaboration with Walter Baade (Osterbrock, 2001), pioneered the theoretical understanding of these phenomena.

There are two primary types of supernovae, categorized based on their distinct mechanisms and observed properties: Type I and Type II. Type I supernovae are further divided into Type Ia, Ib, and Ic, with Type Ia supernovae resulting from the thermonuclear explosion of a white dwarf in a binary system, as it accretes mass

[12]https://www.sdss3.org/science/gallery_sn_gallery24.php

Figure 3.2: A mosaic showing 36 of the 500+ Type Ia supernovae discovered by the Sloan Supernova Survey[12]. Each image is centered on the supernova, which usually stands out as a bright point near or within the galaxy that hosts it. The light of the supernova, powered by the thermonuclear explosion of a single white dwarf star, can outshine that of the tens of billions of stars in its host galaxy. Type Ia supernovae have a constant intrinsic luminosity (after a correction based on the time over which their light rises and falls), so their apparent brightness can be used to infer their distance. The primary goal of the Sloan Supernova Survey was to measure the expansion of the universe with high precision over the last four billion years of cosmic history, to help understand why that expansion is speeding up over time despite the decelerating gravitational effect of atoms and dark matter. Credit: B. Dilday and the Sloan Digital Sky Survey.

from a companion star. Type Ib and Ic supernovae arise from the core-collapse of massive stars, which have lost their outer hydrogen (and possibly helium) envelopes (Filippenko, 1997). Type II supernovae also result from the core-collapse of massive stars, but these stars retain their hydrogen envelopes, leading to different spectral features. This original classification has been made through spectroscopic measurements but this type of measurement is very demanding (large instruments, long time exposure) compared to photometric measurement.

Supernovae play a vital role in the universe, as they are responsible for the production and distribution of many heavy elements, which are essential for the formation of planets and life. The energy and particles released by supernovae also influence the dynamics of their host galaxies, triggering star formation and shaping the interstellar medium[13] (Woosley et al., 2002). Furthermore, Type Ia supernovae are essential tools in cosmology, serving as previously mentioned as "standard candles" for measuring cosmic distances due to their predictable peak brightness.

Observations of Type Ia supernovae have led to the groundbreaking discovery of the accelerated expansion of the universe, driven by dark energy (Riess et al., 1998; Perlmutter et al., 1999). This finding has profound implications for our understanding of the universe's composition, fate, and the nature of the enigmatic dark energy. Studies of supernovae also contribute to our knowledge of stellar evolution, as they showcase the final stages in the lives of stars with varying masses and initial conditions.

The investigation of supernovae continues to yield important insights into the life cycles of stars, the chemical evolution of galaxies, and the expansion history of the universe. With advances in observational capabilities and theoretical understanding, astronomers are uncovering new details about these explosive events and their substantial impact on the cosmos. As we briefly outlined earlier, the transient family is incredibly diverse. The initial stage of many specialized studies, such as distance measurements with type Ia supernova, involves the classification of different types of transients to select pertinent objects.

[13]The interstellar medium (ISM) is the matter and energy that exists between star systems within a galaxy. It consists of gas, dust, and cosmic rays, and is primarily composed of hydrogen and helium, along with trace amounts of heavier elements.

3.1.5 Astronomy surveys and Data acquisition

3.1.5.1 Astronomical Time domain

Time-domain astronomy focuses on the study of celestial objects that change over time, including transient events and periodic variations. This exciting field has seen significant growth in recent years, thanks to advancements in observational capabilities such as the Vera C Rubin Legacy Survey (Ivezić et al., 2019) and the Zwicky Transient Facility (ZTF) (Bellm et al., 2019). Time-domain astronomy encompasses various astronomical phenomena, including supernovae, gamma-ray bursts, variable stars, and exoplanet transits, providing valuable insights into their underlying physics and evolutionary processes.

Transient events, such as supernovae and gamma-ray bursts, offer a unique perspective on the final stages of stellar evolution and the explosive mechanisms driving these events. (Riess et al., 1998; Perlmutter et al., 1999). Gamma-ray bursts, the most energetic events in the universe, are associated with the formation of black holes and the collapse of massive stars (Mészáros, 2006).

Periodic variations, like those observed in variable stars, provide essential information on stellar structure, evolution, and pulsation mechanisms[14]. Cepheid variables, a well-known class of variable stars, have been instrumental in determining cosmic distances and establishing the cosmic distance ladder[15] (Freedman et al., 2001). Time-domain astronomy has also made significant contributions to exoplanet research, with transit photometry being a key method for detecting ex-

[14]Pulsation mechanisms refer to the physical processes that cause certain stars to undergo regular, periodic changes in their size, temperature, and brightness. These changes, or pulsations, are primarily driven by the interplay between gravity and pressure within the star, which causes the star to oscillate in a way that is similar to how sound waves propagate through a medium.

[15]The cosmic distance ladder, also known as the extragalactic distance scale, is a series of methods astronomers use to measure distances to celestial objects.

oplanets and characterizing their properties (Seager & Deming, 2010).

As we continue to enhance our observational capabilities, time-domain astronomy will assume an even more critical role in deepening our understanding of the dynamic and evolving universe. This field offers immense potential for unveiling new celestial phenomena and enriching our comprehension of familiar objects. With ongoing investments in cutting-edge technology and observational infrastructure, we can anticipate even more remarkable discoveries in time-domain astronomy. In the following subsections, we will explore various observational infrastructures, such as SDSS, LSST, and ZTF, that have significantly contributed to these advancements and discoveries.

3.1.5.2 Sloan Digital Sky Survey

The Sloan Digital Sky Survey (SDSS) is a pioneering astronomical project that has been gathering data since the early 21st century (Fukugita et al., 1996; Frieman et al., 2007; Holtzman et al., 2008; Sako et al., 2014). Its main objective is to develop a comprehensive map of the cosmos, documenting the positions, luminosity, and hues of celestial bodies such as stars, galaxies, and quasars. Positioned at Apache Point Observatory in New Mexico (Figure 3.3 shows the scale of the telescope), a dedicated 2.5-meter telescope furnished with an advanced digital camera and spectrographs is utilized. Consequently, the SDSS has generated an abundance of data, contributing significantly to our knowledge of the Universe's composition and evolution.

One major accomplishment of the SDSS is the assembly of the one the most extensive three-dimensional map of the cosmos (See figure 3.4), encompassing more

[16]https://www.sdss3.org/science/gallery_telescope.php
[17]https://www.sdss3.org/science/gallery_sdss_pie2.php

Figure 3.3: The 2.5-meter telescope at Apache Point Observatory, named the Sloan Foundation Telescope in recognition of the pivotal and generous support of the Alfred P. Sloan Foundation through all phases of the SDSS. All SDSS imaging and spectroscopy were carried out with the Sloan Telescope, equipped with a large format digital camera and fiber-fed spectrographs that measured spectra of 640 objects at a time. These facilities, together with some new instruments, are also being used for SDSS-III. [16].

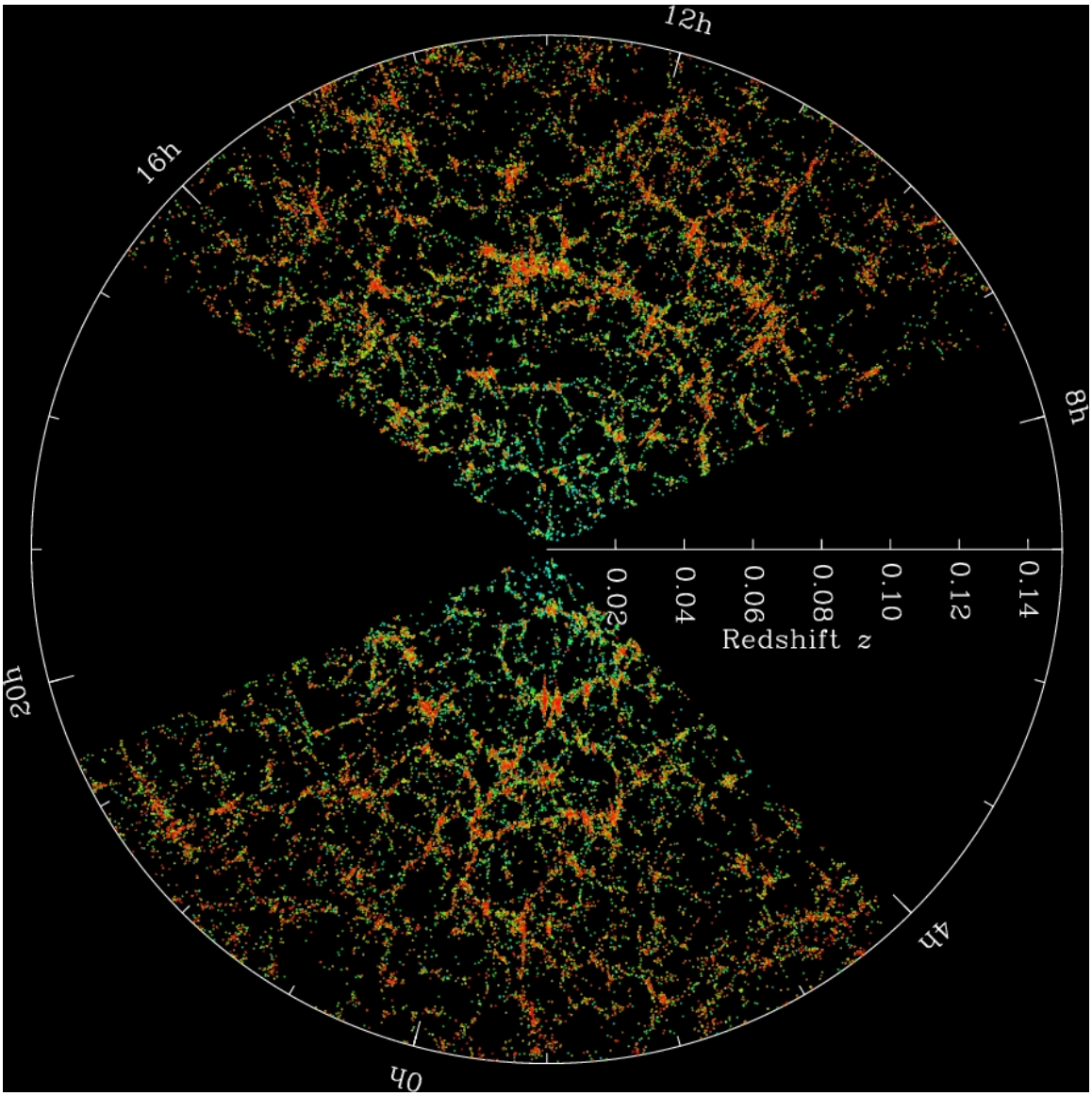

Figure 3.4: Slices through the SDSS 3-dimensional map of the distribution of galaxies. Earth is at the center, and each point represents a galaxy, typically containing about 100 billion stars. Galaxies are colored according to the ages of their stars, with the redder, more strongly clustered points showing galaxies that are made of older stars. The outer circle is at a distance of two billion light years. The region between the wedges was not mapped by the SDSS because dust in our own Galaxy obscures the view of the distant universe in these directions. Both slices contain all galaxies within -1.25 and 1.25 degrees declination. Credit: M. Blanton and the Sloan Digital Sky Survey[17].

than a million galaxies and quasars. This map has offered valuable perspectives on the distribution of matter and the impact of dark energy on the Universe's expansion. Furthermore, the data from SDSS has led to the discovery of numerous celestial entities, such as brown dwarfs, white dwarfs, and black holes, helping scientists to better comprehend their characteristics and the mechanisms governing their formation and development.

The SDSS has also profoundly influenced astrostatistics. The sheer volume of its dataset has necessitated the creation of refined statistical methods and computational instruments for processing and scrutinizing the information it contains . These developments have been crucial for deriving scientific knowledge from the SDSS data and have had wider implications for other areas of astronomy and astrophysics.

In conclusion, the Sloan Digital Sky Survey has been a cornerstone of modern astronomy, driving significant advancements in our understanding of the Universe. The project's success lies in its innovative approach to data collection, its ongoing commitment to providing high-quality data to the scientific community, and the collaborative spirit that has characterized its development and operation.

3.1.5.3 Zwicky Transient Facility

The Zwicky Transient Facility (ZTF) represents an impressive astronomical survey initiative, concentrating on detecting and scrutinizing transient and variable celestial objects. Housed at California's Palomar Observatory, the ZTF has been operational since 2018, employing the Samuel Oschin 48-inch Schmidt Telescope (See figure 3.5), which is outfitted with an advanced 576-megapixel camera (Bellm et al., 2019). This camera boasts an extensive field of view, enabling the ZTF to sur-

Figure 3.5: The Zwicky Transient Facility scans the sky using a state-of-the-art wide-field camera mounted on the Samuel Oschin telescope at the Palomar Observatory in Southern California Image credit: Palomar Observatory/Caltech[18].

vey the entire visible northern sky each night. Identifying transient phenomena, such as supernovae, asteroids, and other ephemeral celestial events, is the primary focus of the ZTF, as they offer crucial information on the Universe's dynamic processes.

The ZTF's accomplishments include the discovery of a wide array of transient events, encompassing supernovae, tidal disruption events, and gamma-ray bursts (Graham et al., 2019; Keller et al., 2021). By closely observing these transient phenomena, astronomers can gather valuable insights into the objects' physical properties and evolutionary processes. For example, studying supernovae can illuminate the factors causing the explosion of massive stars, while analyzing gamma-ray bursts can uncover details about the energetic processes at play in the early

[18]https://www.ztf.caltech.edu/new/deep-learning-helps-ztf-astronomers-classify-supernovae.html

Universe.

In addition to transient events, the ZTF has played a pivotal role in detecting and characterizing variable stars and other periodically varying objects (Masci et al., 2019). These observations prove essential for understanding stars' internal structure and evolution, as well as the formation and properties of exoplanetary systems. Moreover, the data gathered from the ZTF has been employed to examine the distribution and characteristics of small solar system bodies, like asteroids and comets, offering insights into the origins and development of our solar system.

The impact of the Zwicky Transient Facility on the field of time-domain astronomy is undeniable, and its ongoing observations continue to unveil new information about the dynamic processes within the Universe. However, recent studies have highlighted some challenges faced by the ZTF. One such study analyzed the impact of SpaceX's Starlink satellites on ZTF survey observations, finding that twilight observations are particularly affected by satellite streaks (Mróz et al., 2022). Despite this, as the ZTF accumulates more data and researchers devise novel analysis techniques, we can anticipate even more thrilling discoveries and advancements in our comprehension of the cosmos. The ZTF exemplifies the power of large-scale astronomical surveys in demystifying the Universe's secrets, while also highlighting the need to consider the effects of human-made objects on astronomical observations.

3.1.5.4 Vera C. Rubin Observatory

The Vera C. Rubin Observatory, is an ambitious ground-based astronomical facility under construction on Chile's Cerro Pachón. Figure 3.6 shows a three-dimensional rendering of the giant telescope which is equipped with an 8.4-meter primary mir-

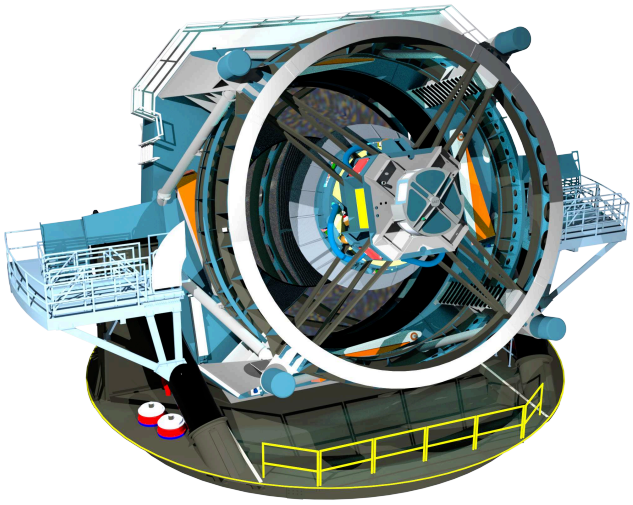

Figure 3.6: A three-dimensional rendering of the baseline design for the LSST with the telescope pointed at about 45 degrees of elevation. Credit: LSST Project Office[19].

ror and a staggering 3.2-gigapixel camera, it boasts one of the most sizable digital cameras ever assembled Ivezic et al. (2019). This remarkable pairing yields an incredibly broad field of view and high sensitivity, empowering the Rubin Observatory to scrutinize vast sections of the sky and identify faint cosmic objects. The observatory's primary objective is a 10-year survey of the southern sky, capturing high-quality imagery and collecting data to address a plethora of scientific inquiries in astronomy, astrophysics, and cosmology.

A central scientific goal of the Rubin Observatory is unraveling the mysteries of dark energy and dark matter, which jointly constitute around 95% of the Universe's total mass-energy content. To achieve this, the observatory will employ various observational probes, such as the study of supernovae, weak gravitational lensing, and large-scale structure. These observations will constrain the properties

[19]https://www.lsst.org/gallery/telescope-rendering-2013

of dark energy and dark matter, illuminating their roles in the Universe's expansion and evolution.

Another significant area of focus for the Rubin Observatory is the examination of transient and variable astronomical events. By persistently surveying the sky and monitoring changes, the observatory will detect and characterize an array of phenomena, encompassing supernovae, gamma-ray bursts, and active galactic nuclei. Furthermore, the Rubin Observatory is anticipated to discover and track numerous small solar system bodies, such as asteroids and comets, offering valuable insights into our solar system's formation and evolution.

In preparation for the flood of data from the Rubin Observatory, the Photometric LSST Astronomical Time-Series Classification Challenge (PLAsTiCC) was organized as a community-wide initiative to stimulate the development of algorithms for classifying astronomical transients (Ivezić et al., 2019). Given that the Rubin Observatory will uncover tens of thousands of transient phenomena every night, automated algorithms for classifying and sorting astronomical transients are crucial. Drawing inspiration from the highly successful Supernova Photometric Classification Challenge, PLAsTiCC consists of a set of realistic LSST simulations of various transient and variable phenomena.

The Vera C. Rubin Observatory's potent capabilities and comprehensive survey plans position it to revolutionize our understanding of the cosmos. Its extensive dataset will afford unprecedented opportunities for scientific discoveries and empower researchers to tackle some of the most pressing questions in astronomy and astrophysics Ivezic et al. (2019). The Rubin Observatory is currently in its final construction phase and the full setup starts its commissioning phase for a full survey start expected in early 2025. As the Rubin Observatory commences its decade-long

survey, the global scientific community eagerly awaits the transformative findings that will undoubtedly emerge from it.

3.1.6 Astronomical images and Light curves

3.1.6.1 Astronomical Images

Astronomical images provide captivating visual portrayals of celestial objects and phenomena, captured through a variety of observing instruments, from ground-based telescopes to space-borne observatories. Rich with information about the physical properties and processes within the observed subjects, these images hold the key to numerous insights. By employing advanced image processing techniques and analysis tools, astronomers can delve into various aspects, such as brightness, color, shape, and temporal changes, which contribute significantly to our understanding of the Universe.

Noise and artifacts, which might occur during image acquisition, are crucial elements to address when processing astronomical images. Such distortions can stem from the observing instrument, atmospheric conditions, and even cosmic rays. To tackle these challenges, astronomers implement techniques like dark frame subtraction (Levesque & Lelievre, 2010), flat field correction (Wang et al., 2022), and cosmic ray removal (Bai et al., 2017). These methods improve overall image quality and accuracy. Furthermore, advanced image processing algorithms, such as the Source Extractor software package Bertin & Arnouts (1996), are employed to detect and measure celestial objects' properties within the images.

Calibrating the brightness of astronomical images is essential for obtaining accurate measurements of objects' properties, such as luminosity and distance. This calibration process involves comparing the observed objects' brightness to

known reference stars with well-established brightness values. Consequently, astronomers can acquire reliable measurements of the intrinsic properties of the observed objects, enabling them to make meaningful comparisons and draw robust conclusions about the Universe's nature.

In conclusion, astronomical images, including time series data, serve as indispensable resources for astronomers, revealing a wealth of information about celestial objects and the underlying processes governing the Universe. Time series data, in particular, enables researchers to study transient events, variable objects, and temporal changes in the cosmos. By employing image processing techniques, noise reduction, and calibration, scientists can extract valuable data that furthers our understanding of the Universe. As advancements in technology continue to deliver increasingly detailed images and more powerful analysis tools, the role of astronomical images, including time series data, in shaping our knowledge of the Universe is poised to grow.

3.1.6.2 Light Curves

Light curve calculation is a fundamental component in the field of astronomy, which allows scientists to track and chart the brightness of celestial bodies over time. The term "light curve" refers to the graphical representation of the luminosity or intensity of light emitted or reflected by an astronomical object (See figure 3.7), such as a star, planet, or galaxy, as a function of time. These curves are of significant value to astronomers as they provide insights into the physical properties and behavior of these celestial bodies, aiding in understanding dynamic phenomena such as variable star brightness, eclipses, stellar rotation, and exoplanet transits.

The mathematical intricacies involved in the calculation of light curves are so-

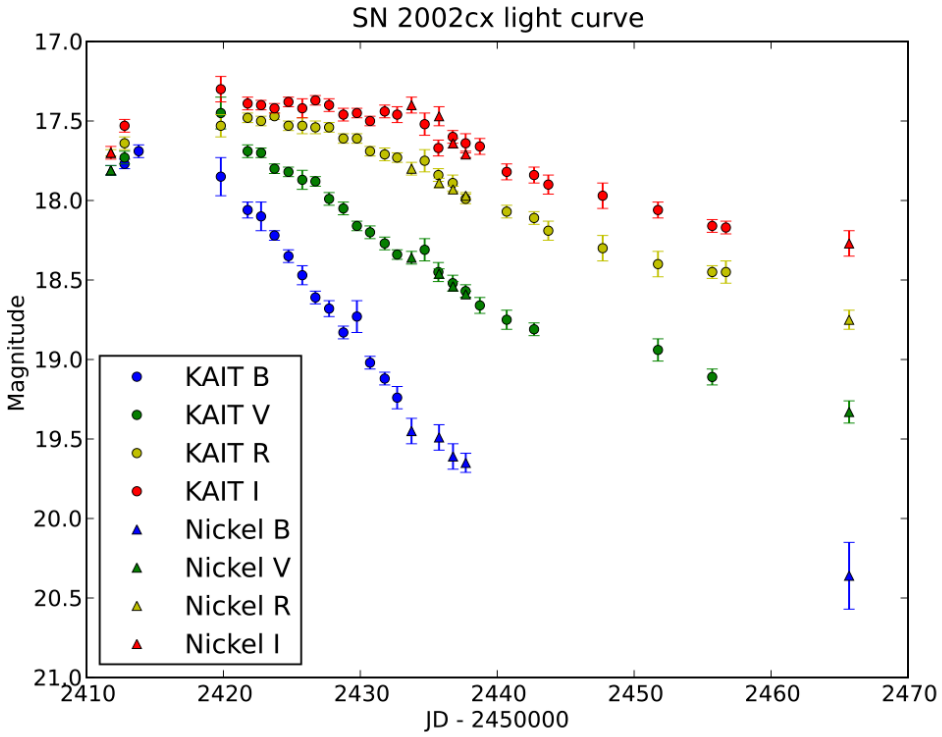

Figure 3.7: Supernova 2002cx light curve. It contains B-band (blue points), V-band (green points), R-band (yellow points), and I-band (red points) from two telescopes, KAIT (circles) and Nickel (Triangles). Credit: Li et al. (2003)

phisticated, involving various factors such as the average background brightness or sky level, variance for each pixel, and an estimate of the point spread function (PSF).

- Pre-processing

For each image, an initial round of analysis is performed to obtain its segmentation map, which is essentially a map of pixels assigned to each object present in the focal plane. Subsequently, for a complete series of images of the same field, the positions of different objects are compared, allowing for the establishment of a transfer function, T_j. This function enables the conversion of pixel coordinates in image j to their coordinates in a reference image chosen based on its quality.

Additionally, the Point Spread Function (PSF) of each image is evaluated using bright stars in the field. The PSF characterizes the response of the imaging system, which includes the telescope optics, the detector, and atmospheric blurring, to a point source of light. By analyzing the PSF, astronomers can gain insights into the instrument's performance and better understand the distribution of light from celestial objects. This information is crucial for optimizing the extraction of information from astronomical images and improving the accuracy of measurements such as photometry and astrometry.

- Resampled Simultaneous Photometry (RSP)

Drawing on the work presented in the research paper by Astier et al. (2013), we delve into the technique of Resampled Simultaneous Photometry (RSP) a method used in astronomy to enhance the precision of photometric measurements, specifically the brightness or flux of celestial bodies, such as stars or supernovae that appear with a host galaxy in the background. The goal of RSP is to account for

variations in instrumental and atmospheric conditions between different observations in a time series of astronomical images. The first step in RSP is to choose the best image from the series, which we refer to as "the reference". We then adjust all other images to match the pixel grid of this reference image. To do this, we adapt the geometric mappings to fit the catalogs, usually leaving residuals around 0.15 pixel, mainly influenced by shot noise. We also adjust the weight maps, especially to correctly account for pixels with zero weight. Then, to align the PSF of the reference with all the other images in the series, we fit discrete convolution kernels to the aligned image pairs. This fitting process is done on stamps centered on the objects in the frame with the highest peak flux that are not saturated.

Now, with our aligned images, we can create a model for the expected light in image i at pixel p:

$$M_{i,p} = \{[f_i \times \phi_{ref}\left(\boldsymbol{x}_p - \boldsymbol{x}_{SN}\right) + gal_{ref}] \otimes K_i\}_p + s_i \tag{3.5}$$

In this equation, f_i represents the supernova (SN) flux in image i, gal_{ref} is the galaxy pixel map in the reference image, K_i is the convolution kernel that helps us match the reference image PSF ϕ_{ref} and flux scale to the ones of image i, and s_i is the sky of image i. We then compare this model to our data using the least squares method:

$$\chi^2 = \sum_i \sum_p w_{i,p} \left(M_{i,p} - I_{i,p}\right)^2 \tag{3.6}$$

When performing a fit, we make certain assumptions. The most important one is that pixels are independent, which overlooks the correlations that resampling introduces between neighboring pixels. This simplification is not ideal, but it doesn't

usually cause a bias in linear least squares approximations. However, it might become an issue when the object's position doesn't enter linearly in the fit. But for the spatial sampling we usually work with, simplified simulations have shown that the effect of these correlations on the variance of the flux estimator is negligible. Despite this, the flux variance estimated from propagating the apparent sky variance is often underestimated because resampling mostly introduces positive correlations between neighboring pixels.

In conclusion, light curve calculation and the analysis of time-series data in astronomy are indispensable tools for studying various astronomical phenomena. By carefully processing and analyzing image data using techniques such as optimal extraction, time-series differential photometry, and resampled simultaneous photometry, astronomers can create accurate light curves that reveal crucial information about the behavior and properties of celestial objects. These light curves provide a wealth of knowledge, enabling researchers to better understand and model the underlying physical processes governing the universe.

3.2 MACHINE LEARNING IN ASTRONOMY

Machine learning will reshape our understanding of the cosmos. The expansive universe, teeming with celestial bodies and intricate phenomena, offers a colossal wealth of data, making it a prime candidate for machine learning applications. In this section, we will delve into the intersection of machine learning and astronomy. We will explore the intricate process of classifying light curves, an indispensable tool in studying the universe.

Light curves can be seen as cosmic fingerprints, they carry unique information about celestial objects, their behaviors, and interactions. We will look at vari-

ous approaches to light curve classification, exploring how these techniques have evolved over time, and the strengths and weaknesses that each holds. The goal is to create a comprehensive understanding of these methodologies.

However, with great power comes great challenges. The task of astronomical object classification using light curves is fraught with complexity. We will confront these challenges head-on, scrutinizing the obstacles that scientists often encounter. Alongside, we will journey through the annals of previous research on astronomical image sequence classification, identifying their contributions and their current limitations.

3.2.1 Light curves

In the early sections of this manuscript, we touched upon the identification of astronomical objects, which can be achieved through both photometric and spectroscopic measurements. Although both methods are vital, they differ significantly in their application and cost. Spectroscopic measurements, for instance, require substantial investment both in terms of the size of the necessary instruments and the observation time. This approach also necessitates a specific strategy for transient events, often involving follow-up after initial photometric detection. While important, spectroscopic measurements will not be the focus of this manuscript. Instead, we will concentrate exclusively on the study of light curves. However, it is essential to recognize the significant role that spectroscopic measurements play in achieving highly confident classifications. As will be detailed in subsequent chapters, these measurements are often used to label the training set for light curve classification, providing a critical layer of precision and reliability to the process. By honing in on light curves, we hope to shed light on a specific, yet vital aspect of

astronomical analysis, while acknowledging the broader context in which it operates.

3.2.1.1 Different light curve classification approaches

In the vast expanse of the universe, light curves serve as a critical tool, allowing us to discern the unique characteristics of celestial objects. As we delve into this subsection, we will unravel the techniques employed to classify light curves. From traditional methods to the cutting-edge machine learning algorithms, each approach has its own set of merits and challenges, and understanding them can lead to more accurate and efficient astronomical analyses.

Gabruseva et al. (2019) presented an automatic photometric classification model, a top contender in the PLAsTiCC astronomical challenge (PLAsTiCC-team et al., 2018). Their approach leverages the LightGBM model, feature extraction, and augmentation. The team tested a large amount of features, selecting the most pertinent ones through a permutation importance algorithm to avoid overfitting and underfitting. While the model achieved high accuracy rates of 88-100% for most astronomical objects, it struggled with supernova classification, suggesting the need for additional classifiers and template fittings for these types.

Alternatively Boone (2019) proposed a new photometric classification framework designed for transients and variables from surveys such as LSST. This approach leverages GP regression to augment the training data set, thereby extending the range of redshifts and observing conditions, without requiring a specific light curve model. The classifier excelled in the PLAsTiCC data set, outperforming any single algorithm to date with an AUC of 0.957 for SNe Ia classification. Furthermore, the classifier is designed to yield output probabilities independent of the

redshift distributions in the training sample, enhancing the understanding of the output probabilities.

Compared to the LightGBM-based model by Gabruseva et al. (2019), this new approach does not focus on feature selection but instead on augmenting the training data. While both models achieved high accuracies, Avocado approach showed remarkable performance in classifying SNe Ia, a category where the (Gabruseva et al., 2019) model struggled. Both methods highlight the importance of avoiding overfitting, but they tackle this issue differently (Gabruseva et al., 2019) through feature selection and (Boone, 2019) through data augmentation.

In another approach, a deeP architecturE for the LIght Curve ANalysis (PELICAN) Pasquet et al. (2019) is developed for the characterization and classification of light curves without requiring additional features. PELICAN effectively handles sparsity and irregular sampling, addressing the issue of non-representativeness between training and test databases due to spectroscopic follow-up limitations. The methodology was applied to different supernova light curve databases and demonstrated impressive results, including an accuracy of 0.811 on the Supernova Photometric Classification Challenge and 87.4% detection of SNe Ia with over 98% precision for LSST Deep Fields.

Möller & de Boissière (2020) introduced SuperNNova, an open-source supernova photometric classification framework that employs deep neural networks, specifically a recurrent neural network (RNN). The RNN is trained solely on photometric information, although the inclusion of additional data like host-galaxy redshift can enhance performance. The method demonstrates remarkable accuracy for type Ia vs non-Ia supernovae classification, exceeding 96.92% without redshift information and reaching up to 99.55% with redshift data. Furthermore, SuperN-

Nova shows unprecedented performance for incomplete light-curves, maintaining accuracies over 86.4% even two days before maximum light.

Contrasting with the previous approaches (Gabruseva et al., 2019; Boone, 2019; Pasquet et al., 2019), SuperNNova utilizes RNNs and does not require feature engineering, showing excellent scalability to large datasets. Moreover, the team addresses often overlooked machine learning pitfalls such as poor calibration and overconfidence on out-of-distribution samples, demonstrating better handling of uncertainties under a Bayesian light. The framework's effectiveness, scalability, and robust uncertainty handling position it as a promising tool for supernova photometric classification and subsequent cosmological studies.

While SuperNNova's approach has its advantages, it's worth noting that the Bayesian neural networks used in this method are still significantly more complex and computationally intensive than standard neural networks. This complexity could potentially pose challenges in terms of computational resources and processing time, especially when dealing with vast amounts of data typical in astronomical studies. Therefore, while the Bayesian approach enhances the model's ability to handle uncertainties, it does come with its trade-offs.

Another innovative method for Type Ia supernovae classification has been presented by the authors of SCONE (Qu et al., 2021), who utilize convolutional neural networks (CNNs), a type of neural network typically used for image recognition tasks. This model is trained solely on photometric data, bypassing the need for accurate redshift data. The unique feature of this approach lies in its preprocessing of photometric data into two-dimensional images via 2D Gaussian process regression. These images, dubbed "flux heatmaps," and corresponding "uncertainty heatmaps" represent the dataset for the model. This strategy smooths over irregu-

lar sampling rates between filters and allows SCONE to operate independently of the filter set used during training.

The performance of SCONE is noteworthy, achieving a test accuracy of 99.73 without redshift on the in-distribution SNIa classification problem and an accuracy of 98.18% when performing 6-way classification of supernovae by type. However, the model's out-of-distribution performance is not as strong as its in-distribution results, highlighting that the specific characteristics of the training sample relative to the test sample can significantly impact performance. This suggests the need for careful consideration when assembling training datasets for such models.

In conclusion, the classification of light curves is a complex and evolving field, with multiple methodologies being explored to enhance performance and accuracy. From traditional methods such as those employed by Gabruseva et al. (2019) and Boone (2019), to the application of deep learning techniques as seen in the PELICAN (Pasquet et al., 2019), SuperNNova (Möller & de Boissière, 2020), and SCONE (Qu et al., 2021) models, each approach has contributed significantly to our understanding and analysis of celestial objects. While these models have demonstrated impressive performance, they also underscore the importance of careful feature selection, data augmentation, and preprocessing, as well as the need for managing the complexities and computational demands of advanced neural networks.

However, it's crucial to acknowledge that these models, despite their innovative approaches and notable successes, also possess several limitations. These range from the need for meticulous data preparation, the challenge of handling incomplete or irregular data, the complexity and computational intensity of certain neural network models, to the differences in performance between in-distribution

and out-of-distribution data. These limitations highlight the room for further improvement and refinement in these models.

The ultimate goal remains the development of robust, efficient, and accurate classification models that can handle the vast and complex datasets generated by modern astronomical surveys. These ongoing efforts continue to enhance our knowledge of the universe, promising exciting discoveries in the future. In the following, we will delve deeper into these limitations, exploring their implications and potential strategies to overcome them.

3.2.1.2 Navigating the Challenges of Astronomical Object Classification using Light Curves

Classification of astronomical objects using light curves is a complex task that faces several significant challenges. A fundamental issue arises from the generation of light curves themselves. To create a light curve, two consecutive images need to be carefully aligned, and one of the images often has to be downgraded in quality to enable subtraction and thereby determine the flux. This procedure can lead to a loss of information, potentially obscuring important details and reducing the reliability of the resulting light curve.

Moreover, many astronomical objects are not isolated but are part of intricate scenes with multiple light sources, known as blended objects. This blending can distort the perceived light curve of an individual object, making its classification more challenging. While dedicated algorithms called scene modeling can help mitigate such issues, they are often computationally intensive, which can be prohibitive given the vast amounts of data typical in astronomical surveys. This computational demand makes it difficult to apply these algorithms on a large scale,

especially when resources are limited.

In addition to the computational limitations, most existing classification methods do not take into account the scene information, which is the background of the transient object. The background can contain valuable information that could potentially aid in the classification of the object. Ignoring this information could potentially lead to misclassification, particularly in cases where the background has distinct characteristics that could influence the observed light curve.

Furthermore, many of the novel methods for light curve classification, such as the deep learning-based approaches employed by PELICAN, SuperNNova, and SCONE, deal with challenges associated with the quality and completeness of the data. Irregular sampling rates, noisy data, and incomplete light curves can all hinder the performance of these models. While techniques such as Gaussian Process regression and data augmentation can help to some extent, these issues remain a significant challenge.

Lastly, the issue of non-representativeness between training and test databases also poses a significant challenge. Differences in redshift distributions, observational conditions, and inherent properties of the objects can lead to substantial differences between the training and test datasets. This mismatch can result in models that perform well on the training data but poorly on new, unseen data. As such, careful consideration must be given to the selection and preparation of the training data, and strategies must be developed to ensure that the models are robust to these differences. Despite these challenges, the field of light curve classification continues to evolve, with ongoing research aimed at developing more robust, efficient, and accurate methods.

3.2.2 Astronomical image sequences

3.2.2.1 Previous Research on Astronomical Image Sequence Classification

Astronomical image sequences offer several advantages over light curves for the classification of celestial objects. Light curves, while highly useful, essentially offer a one-dimensional view of a celestial object's brightness over time. In contrast, image sequences provide multidimensional information, capturing not only the brightness variation of an object but also its spatial information, morphology, and any associated contextual information from surrounding celestial objects.

This spatial-temporal data richness enhances the ability to discern more complex and subtle patterns and relationships like disentangling a static background and varying source in changing observational conditions, potentially leading to more accurate and comprehensive classification of astronomical objects. In some specific cases, image sequences can capture the evolution of objects, their interaction with the surrounding environment, or changes in their shape or structure, such occurrences can be noted in strongly gravitationally lensed transients. Indeed, for supernovae in general, we do not anticipate any interaction with surroundings or alterations in shape or structure. All observable changes are attributed to variations in observational conditions.

Moreover, image sequences can be used directly without the need for preprocessing steps, such as those required for generating light curves, reducing the potential for information loss during these stages.

In their research, Carrasco-Davis et al. (2019) proposed a novel classification model for astronomical objects, leveraging a recurrent convolutional neural network (RCNN). This model utilizes sequences of images as inputs, negating the need for light curve computations or difference images, marking the first instance

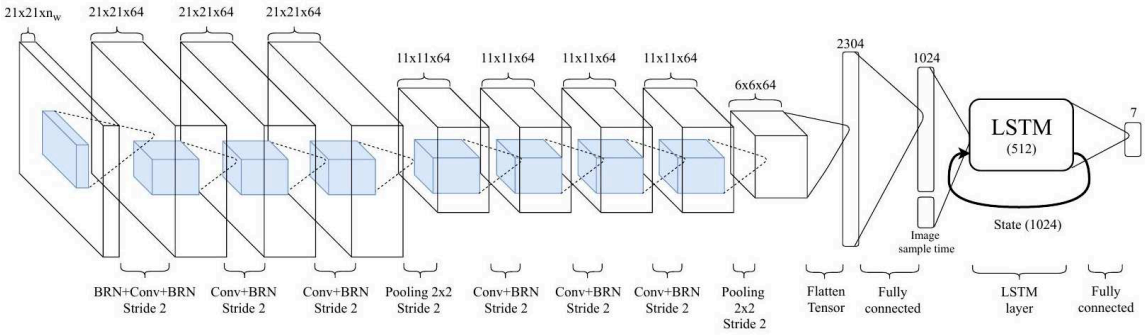

Figure 3.8: RCNN architecture proposed by Carrasco-Davis et al. (2019).

of direct image sequences being used for classifying variable objects in astronomy. The team also introduced an image simulation process to create synthetic image sequences, reflecting realistic, unevenly sampled, and variable noise sets of sequence for each astronomical object.

These simulated sequences offer multiple benefits, such as rapid dataset creation adaptable to various surveys and classification tasks. The aim is to develop a simulated dataset closely mirroring the real one, enabling fine-tuning to match distributions and address the domain adaptation issue. The researchers used real-world data from the High cadence Transient Survey (HiTS) to test the RCNN classifier trained with the synthetic dataset, achieving an average recall of 85%, improved to 94% after fine-tuning with 10 real samples per class.

The RCNN model's performance was compared to a light curve random forest classifier, showing similar performance levels on the HiTS dataset. The RCNN approach offers numerous advantages, such as reduced data pre-processing, faster online evaluation, and easier performance enhancement using a few real data samples.

Alternatively, Gómez et al. (2020) introduced a deep learning approach for su-

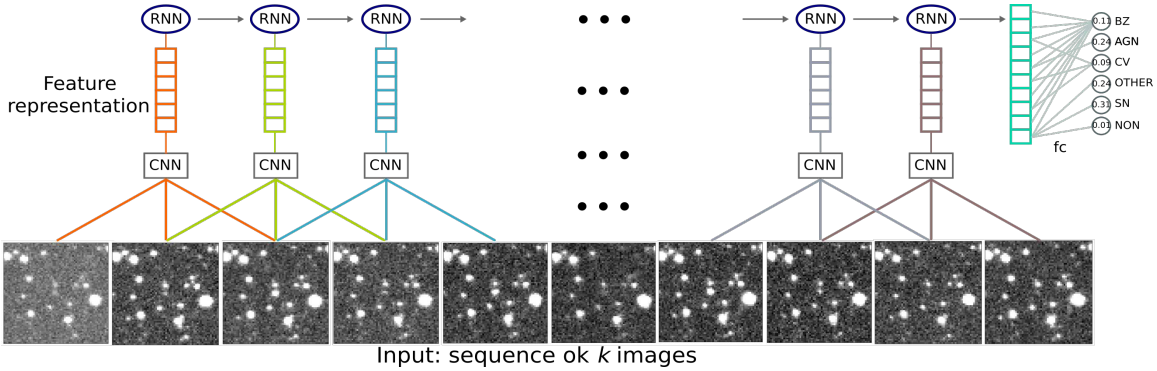

Figure 3.9: TAO-Net architecture proposed by Gómez et al. (2020).

pervised classification of temporal sequences of astronomical images into distinct transient astrophysical phenomena. Traditionally, this task required human experts' intervention to find heuristic features, process images through subtraction, or extract sparse information such as flux time series, or light curves. The proposed method, known as TAO-Net (Transient Astronomical Objects Network), employs Deep Convolutional Neural Networks and Gated Recurrent Units to model the spatio-temporal patterns explicitly.

The researchers trained these deep neural networks on 1.3 million real astronomical images from the Catalina Real-Time Transient Survey, classifying the sequences into five different types of astronomical transient classes. TAO-Net demonstrated a significant improvement over random forest classification on light curves, with a 10-percentage point increase in the F1 score for each class. The average F1 score across classes increased from 45% with random forest classification to 55% with TAO-Net.

This achievement with TAO-Net suggests the potential for developing new deep-learning architectures for early transient detection. The approach differs from the RCNN model proposed by Carrasco-Davis et al. (2019), as it focuses on learning directly from imaging data without requiring synthetic image sequences

The advancement in astronomical object classification through image sequences reflects a significant step forward, offering multi-dimensional information and reducing the need for preprocessing steps associated with light curve generation. Two approaches, namely the Recurrent Convolutional Neural Network (RCNN) model by Carrasco-Davis et al. (2019) and TAO-Net by Gómez et al. (2020), exemplify the potential of using image sequences for object classification in astronomy.

Despite their different approaches, both methods showcase the advantages of using image sequences over light curves, including capturing spatial and temporal patterns, reducing data pre-processing, and facilitating faster online evaluations.

3.2.2.2 Facing the Challenges: Current Limitations

While astronomical image sequences present a promising direction for the classification of celestial objects, they also bring forth a set of unique challenges that need to be addressed for their effective and broad application.

One significant challenge lies in the scarcity of available image sequence datasets. The construction of an image sequence dataset is a time and resource-intensive process requiring accurate time-stamped images of celestial objects taken at regular intervals. The collection of such data necessitates large-scale, ongoing sky surveys using powerful telescopes, which may not be feasible for all astronomical research institutions due to logistical and financial constraints. Additionally, the data must be carefully cleaned and curated to ensure its quality and usability, further adding to the complexity of dataset creation.

The limited number of research studies utilizing image sequences for astronomical object classification presents another challenge. As of now, we have discussed two prominent works, one by Carrasco-Davis et al. (2019) and the other by

Gómez et al. (2020). Both have shown promising results, but the field is still in its early stages, and more research is needed to further establish and validate these methods. A broader base of research would provide a more robust understanding of the potential advantages and limitations of image sequences and allow for the development of more refined and effective classification models.

Moreover, the existing works have focused on specific types of celestial objects and used certain types of neural network architectures, such as Recurrent Convolutional Neural Networks and Deep Convolutional Neural Networks with Gated Recurrent Units. However, the diversity of celestial objects and the complexity of their behaviors may require a broader array of modeling approaches. The development of these models may be hindered by the lack of diverse and high-quality image sequence datasets, creating a kind of 'chicken and egg' problem.

Further, the use of synthetic image sequences, as proposed by Carrasco-Davis et al. (2019), is a novel approach to circumvent the issue of dataset scarcity. However, the synthetic data must be close enough to real-world data for effective training and testing, which may be challenging due to the inherent complexity and variability of astronomical phenomena.

In conclusion, while astronomical image sequences offer exciting potential for improved object classification, they also present unique challenges related to data availability, the nascent state of research in the field, the diversity of celestial objects, and the use of synthetic data. Addressing these challenges will require continued research and development efforts and may also necessitate greater collaboration and resource sharing within the astronomical research community.

CHAPTER 4

Astronomical image time series classification using CONVolutional attENTION (ConvEntion)

4.1 INTRODUCTION

The exploration of the universe presents a significant challenge to the astronomical community. With the advent of increasingly powerful telescopes, the field is undergoing a transformation, capturing an unprecedented volume of data. The task at hand is to process and analyze these massive data sets, which require substantial computational resources and human effort. Future advancements are expected to detect astronomical phenomena at a rate that far exceeds current capabilities, discovering up to 10 million new objects each night (Some are artifacts (fake), some are old transients or variable objects). These include a variety of celestial bodies, such as active galactic nuclei (AGNs), variables, cepheids, RR Lyrae, and supernovae, the last of which are crucial for cosmological studies, particularly in understanding the accelerated expansion of the universe.

The classification of these diverse objects involves a complex pipeline. The first step is photometry, which measures the flux per band from a series of images, with the band number varying based on the survey. A brightness change time series, or light curve, is then generated and supplied to a machine learning classifier to identify the object type.

However, the use of light curves presents certain challenges. The generation of a light curve requires accurate alignment of consecutive images and a reduction in the quality of one image for flux subtraction, potentially leading to data loss. Dedicated algorithms known as scene modeling help address some of these

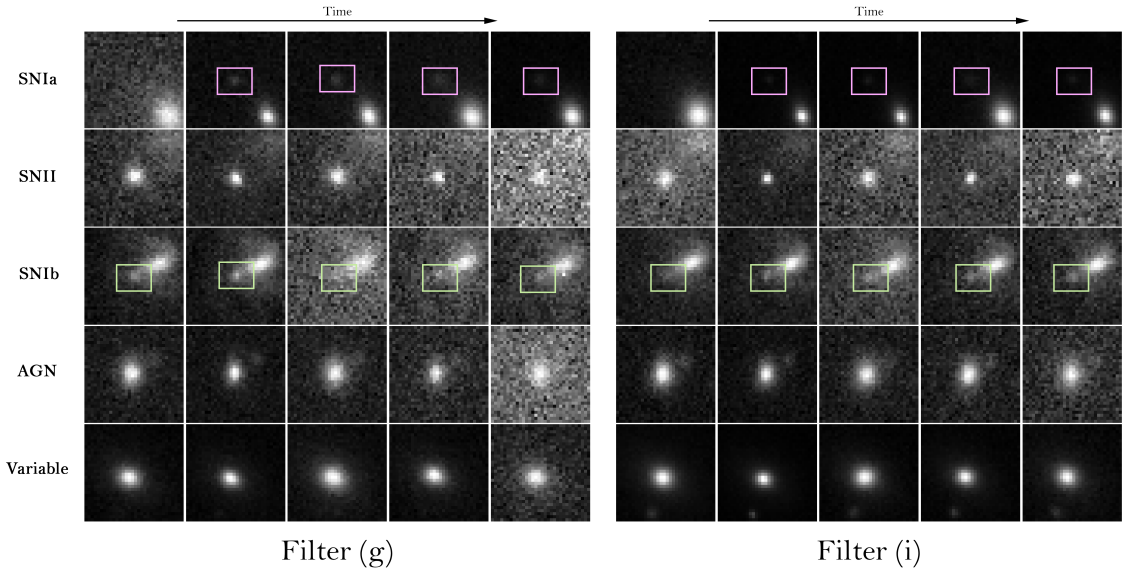

Figure 4.1: Sample of some objects present in our dataset. Each image in filter g/i corresponds to a different observation with the same filter.

challenges, but they come with significant computational resource requirements. Moreover, the scene information, the background against which the transient object is observed, is generally not incorporated into the classification process.

Recent studies, such as those by Carrasco-Davis et al. (2019) and Gómez et al. (2020), have proposed bypassing the feature extraction and light curve stages to classify objects directly using images. They employed a recurrent neural network (RNN) for classifying sequences after spatial features were extracted via a convolutional neural network (CNN). This strategy showed promising results, and we aim to enhance this classification approach and address existing challenges.

In our work, we present a novel deep learning transformer-based architecture for classifying astronomical image time series. Rather than separating spatial and temporal feature extraction, we consolidate these two processes, improving the network's object classification capabilities. Moreover, we offer a solution to the

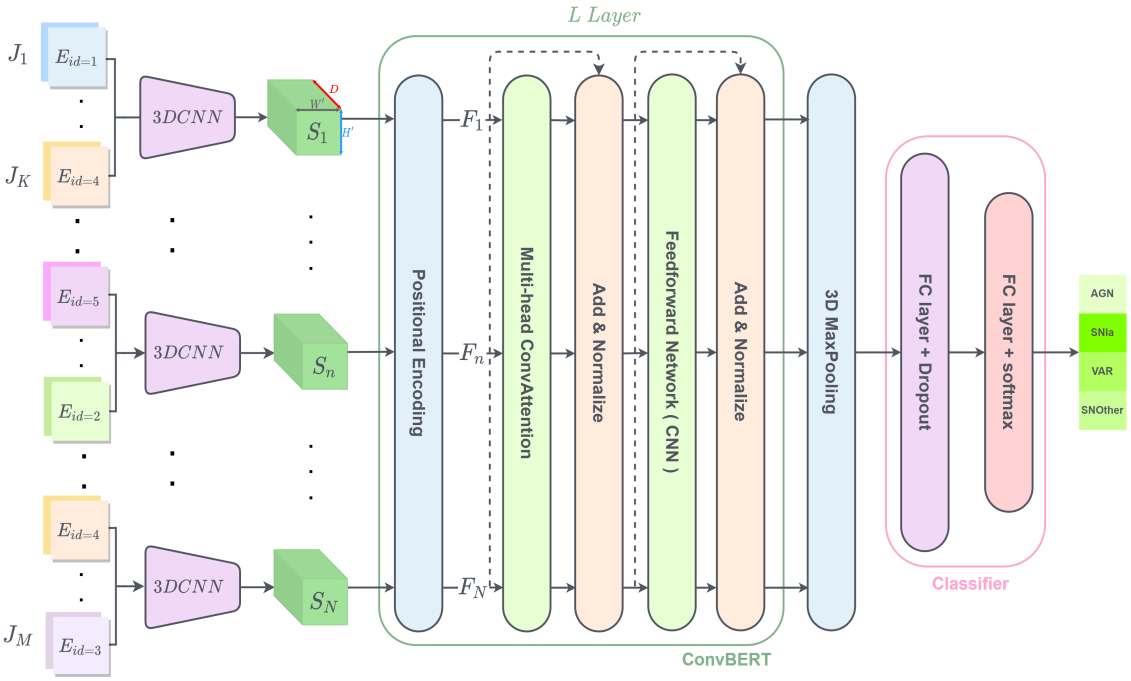

Figure 4.2: General architecture of the ConvEntion network. The image time series are first rearranged to embed the band information. Then each 3DCNN is fed with a sub-sequence of K inputs of the time series $J(\in \mathbb{R}^{M\times H\times W\times 2}$ for M elements of images of size HxW) to create the new downsized sequence $S(\in \mathbb{R}^{N\times H'\times W'\times D})$. S is fed to the positional encoder in order to add the information about the position, which outputs $F(\in \mathbb{R}^{N\times H'\times W'\times D})$. Then F is passed to ConvBERT which has L layers. The 3D max-pooling is used to downsize the output of ConvBERT for the classifier.

issue of missing observations, which has demonstrated a significant enhancement in the model's accuracy. We tested our model using actual data from a specific survey Holtzman et al. (2008); Sako et al. (2014); Frieman et al. (2007). We outline our dataset in Section 4.2, detail our architecture ConvEntion in Section 4.3, present our results and comparisons with other architectures in Section 4.4, and conclude with our findings and future perspectives in Section 4.5.

4.2 DATASET

4.2.1 Database description

The Sloan Digital Sky Survey (SDSS) Holtzman et al. (2008); Frieman et al. (2007) is a very ambitious and successful large-scale survey program using a dedicated 2.5-meter telescope at Apache Point Observatory, New Mexico, equipped with photometric and spectroscopic instruments that have released images, spectra, and catalog information for several hundred million celestial objects. The dataset used in this paper was collected during the SDSS Supernova Survey Sako et al. (2014), one of three components (along with the Legacy and SEGUE surveys) of SDSS-II, a three-year extension of the original SDSS that operated from July 2005 to July 2008. The Supernova Survey is a time-domain survey, involving repeat imaging of the same region of the sky every other night, weather permitting.

The images are obtained through five wide-band filters Fukugita et al. (1996) named u′, g′, r′, i′, and z′, simplified as u, g, r, i, and z in the following, which corresponds to an effective mid-point wavelength of u (365nm), g (475nm), r (658nm), i (806nm), and z (900nm). The survey region observed repeatedly over three years is a 2.5-degree-wide stripe centered on the celestial equator in the Southern Galactic Cap that has been imaged numerous times in the last twenty years, allowing for the construction of a big image database for the discovery of new celestial objects. Most of the sources have included galactic variable stars, active galactic nuclei (AGN), supernovae (SNe), and other astronomical transients, all of which have been processed to generate multi-band (ugriz) light curves. The imaging survey is reinforced by an extensive spectroscopic follow-up program that uses spectroscopic diagnostics to identify SNe and measure their redshifts. Light curves were

evaluated during the survey to provide an initial photometric type of the SNe and a selected sample of sources was targeted for spectroscopic observations.

In order to investigate the classification from images rather than light curves, we acquired the images from the public SDSS dataset through their platform. Our dataset contains many types of supernovas (see Table 4.1 and Sako et al. (2014)). The label of "unknown" mainly represents very sparse or poorly measured transient candidates, "variables" have signals spanning over two seasons, and "AGNs" have a spectral signature. The three other classes are supernovae of type Ia, Ib/c, and II. Among supernovae, the typing is performed from spectroscopy or from the light curve using different machine learning techniques (see Sako et al., 2014). We grouped the non-Ia supernovas because our focus in this study only on the Ia type for their interest in cosmology as standard candles and also because of the small number of non-Ia with spectral signatures. The very small class of three SLSN bright objects has been added to the non-Ia supernovae. Figure 4.1 shows an example of astronomical image time taken from the SDSS dataset.

Object name	Count
AGN	906
SNIa	499
SNOther	89
Unknown	2009
Variable	3225
SNOther_PT	2041
SNIa_PT	1448

Table 4.1: Number of objects per class in the SDSS dataset. PT: Photometrically typed, which means that the SNs are not spectroscopically verified.

4.2.2 Challenges

Most of the astronomical dataset suffers from a number of problems that should be dealt with before feeding it to the classification algorithm. Among difficulties contributing to the challenging nature of Astronomical Image Time Series (AITS), we can mention class imbalance (as shown in Table 4.1 of our dataset). In particular, we can clearly see that the classes we have are not balanced where the number of samples for variables is much bigger than SNIa. This imbalance significantly impacts machine learning models due to their higher prior probability, which means they tend to overclassify the larger class(es). As a result, instances belonging to the smaller class(es) are more likely to be misclassified than those belonging to the larger class(es). Another problem that impacts the model is missing bands. Indeed, each time an image is acquired in an AITS it is captured through one filter among a set of up to five or more channels. So, an image of a celestial object can be taken in many channels, but not necessarily at the same time. This results in missing bands for a given time of observation (see Figure 4.3). It is well known that the missing data negatively impacts the performance of the model if it is not dealt with. Gill et al. (2007) stated that an increasingly missing percentage of training data resulted in an increased testing error, which requires a solution to mitigate the impact of missing data.

4.3 METHODS

In this section, we propose a neural network based on a combination of convolution and self-attentions. The goal of the model is to handle the challenges that we mentioned previously, such as class imbalance, data sparsity, and missing observations. Figure 4.2 represents the general architecture of the ConvEntion model.

The model takes as its input the sequence of images that have been rearranged to embed the band information (See Section 4.3.1 and Figure 4.4). The sequence first passes through a 3DCNN to downsize its length. It allows for the reduction of the computation complexity of the model and also captures the local characteristics of the objects. The newly constructed sequence by the 3DCNN is fed to a convolutional BERT which then extracts the spatio-temporal features with high-level representation from the input. Finally, we pass the output of the convolutional BERT, which is a projection of our input into a high-level representation subspace, through a 3D max-pooling to downsample it, then it goes on to the final classifier to make the prediction. In the following subsections, we explain each component in depth.

4.3.1 Data modeling

First, we note that throughout the paper, vectors are given in bold capital letters, sizes in capital letters, and indices in lowercase. To start with the **missing data problem**, a network dedicated to image time series is usually fed a sequence of images $\mathbf{I} \in \mathbb{R}^{H \times W \times 5}$, where H and W are, respectively, the height and width of the image and 5 is the number of channels representing the bands (u, g, r, i, z). However, we know, as explained earlier, some bands are missing in the dataset. To fix this issue, instead of giving the model images with empty channels, thus introducing a bias to the network, we decided to separate the channels as individual images ($\mathbf{X} \in \mathbb{R}^{H \times W}$) simply by skipping the empty channels. As a consequence, the information about the type of filter, which holds a crucial value for the network to accurately discriminate between objects, is also eliminated. In an image with different channels, the order of the channels usually represents the type of filter (see

Figure 4.3).

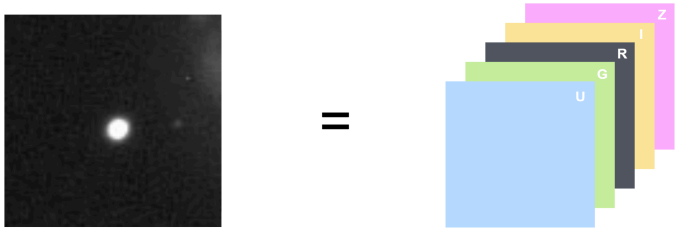

Figure 4.3: Each image has five filters (u, g, r, i, z). The black channel represents the missing observation.

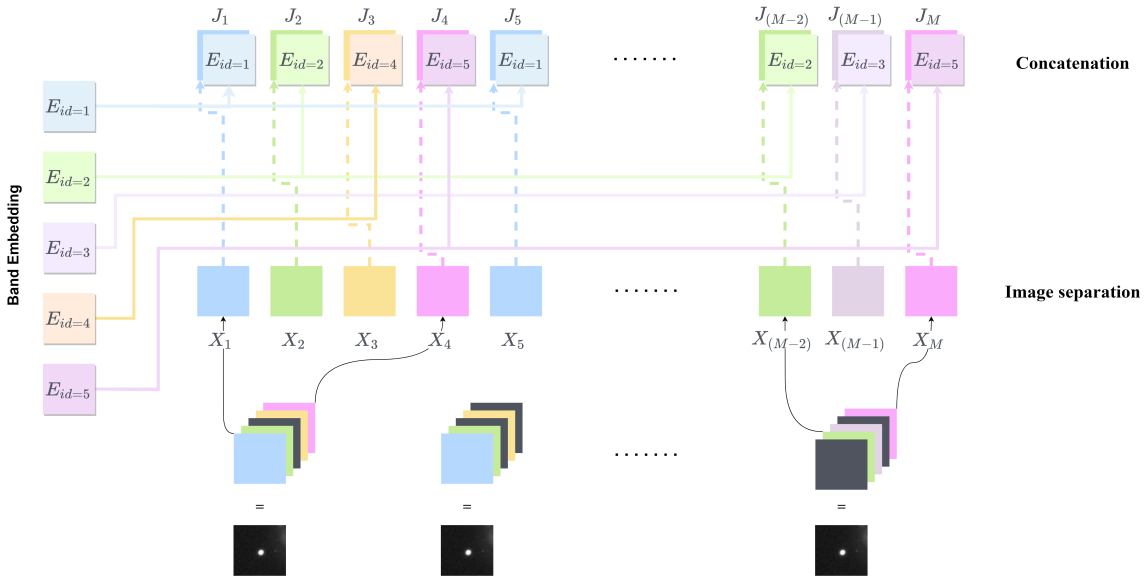

Figure 4.4: Illustration of the handling of missing information by separating the bands. The empty channels are dropped, then we concatenate each image with a 2D representation of the band used to capture the image. The band embedding contains five band representations. The black channel represents the missing observation

In order to preserve this valuable information, we should add the band type to the new 2D images **X**. Knowing that the information about the type of the filter is a categorical feature, thus we need to adapt it to the model 2D input representation. To do so, we propose using an embedding layer to encode the channel type before passing the input to the model. For each band (u, g, r, i, z), we assign a unique

number $id \in \{1, 2, 3, 4, 5\}$. Then, an embedding layer $BandEmbed$ converts the band type id, which is a categorical feature, into 2D dense representation E_{id} with $E_{id} \in \mathbb{R}^{H \times W}$ (see Figure 4.4):

$$E_{id} = BandEmbed(id). \tag{4.1}$$

The embedding layer is a fully connected layer that is reshaped to a 2D representation. The weights of $BandEmbed$ are learnable. After getting the band embedding, we concatenate it with the new image to get our new input $J \in \mathbb{R}^{M \times H \times W \times 2}$ that contains the band information, where M is the length of the sequence:

$$J_m = Concat(X_m, E_{id}), \qquad m \in \{1, .., M\}. \tag{4.2}$$

The problem of **class imbalance** is one of the major challenges for any machine learning project. Some tried to solve this problem by adding a new loss function to mitigate the impact of the class imbalance. For example Lin et al. (2017) proposed a loss function called focal loss which applies a modulating term to the cross-entropy loss in order to focus the learning on hard misclassified examples. However, this approach tends to produce a vanishing gradient during backpropagation Hossain et al. (2021). Other solutions propose the use of oversampling such as SMOTE Chawla et al. (2002). Those authors proposed an approach where they synthesize new samples of the minority class. However, this solution was proposed mainly for tabular data. Knowing that our data are images that contain a much higher number of features than tabular data, it appears obvious that using SMOTE may not be optimal in our case. Dablain et al. (2021) introduced a solution

based on SMOTE dedicated to images called DeepSMOTE. It is aimed at generating new images for the minority class. Once again, this approach is unsuitable in our case as our dataset is not composed of images, but of a sequence of images, and it is too expensive to generate a whole new sequence. So, instead of generating a new one, we used data augmentation and weighted random sampling (WRS) Efraimidis (2015) on our database. We oversampled the dataset, which translates to simply altering the dataset to remove such an imbalance by increasing the number of minority classes and undersampling the data by decreasing the majority classes until we have reached a balanced dataset. In our case, the WRS was applied on a batch level. We generate balanced batches based on the probability of a sample being selected. We weighted each sample according to the inverse frequency of its label's occurrence and then sampled mini-batches from a multinomial distribution based on these weights. This means that samples with high weights are sampled more often for each mini-batch. The same sample can be reused in other mini-batches of the same epoch to increase the minority class, but with a data augmentation applied to it. Different methods of data augmentation were used: for example, a random drop of some steps from the whole sequence to create a new one or a sequence rotation, horizontal and vertical flips, and sequence shifting, where we construct a smaller sequence from the original one which has a bigger length than the input length of ConvEntion.

In our implementation, we recall the dataset at every epoch, the transforms operation (augmentation) is executed and then we get different augmented data. Using this oversampling approach has drastically improved the performance of the model. We used the function $WeightedRandomSampler$ from PyTorch Paszke et al. (2019) as an implementation of WRS.

4.3.2 3D Convolution Network:

In several deep learning applications, large transformer models have demonstrated fantastic success in obtaining state-of-the-art results. However, because the original transformer's self-attention mechanism consumes $O(M^2)$ time and space with respect to the sequence length, M, training the model for a long sequence is so expensive, it causes the problem called "attention bottleneck" Wang et al. (2020); Choromanski et al. (2021). The problem is more severe for us because we use convolutions and 3D tensors inside the attention mechanism; for instance, the attention map is of size $H \times W$, so the complexity of the attention will be $O(M^2 \times H \times W)$. Thus, our model would then be prohibitively expensive to train. In the last few years, there have been numerous proposals aimed at solving this issue. Wang et al. (2020) demonstrated that a low-rank matrix could approximate the self-attention mechanism. They suggested a new self-attention method that minimizes total self-attention complexity. Choromanski et al. (2021) presented a novel transformer architecture that uses linear space and time complexity to estimate regular (softmax) full-rank-attention Transformers with proven accuracy. However, all these propositions remain irrelevant in our case because we do not use the standard self-attention mechanism, as the convolutions make it an arduous task. So, the solution we preferred to go with is to reduce the length of the sequence before feeding it to the transformer block. Reducing the sequence must be done without losing relevant information. Thus, we propose using a 3D convolution neural network (3D CNN). A 3D CNN is an improved type version of CNN first proposed by Tran et al. (2014), where it applies a 3D filter to the dataset and the filter moves in three directions to calculate the low-level feature representations. Their output shape is in a 3D volume space. We applied $3DCNN$ where we input the sequence

Layer	Layer Parameters
Conv3d + BN3d	$11 \times 11 \times 3 \times 64, 64$
Conv3d + BN3d	$5 \times 5 \times 3 \times 128, 128$
Conv3d + BN3d	$3 \times 3 \times 3 \times 64, 64$
Conv3d + BN3d	$3 \times 3 \times 3 \times 64, 64$

Table 4.2: 3D CNN architecture where Conv3D is a 3D convolutional element and BN3d is a 3D batch normalization element.

J to get the reduced new sequence **S** following the equation:

$$S_n = \text{3DCNN}(J_{(n-1)*K+1}, .., J_{n*K}), \quad n \in \{1, .., N\}. \tag{4.3}$$

We let M be the length of the series, J and we fed K inputs of J to the 3DCNN to generate one entry, S for our transformer. So, in the end, the new sequence, S will be $S \in \mathbb{R}^{N \times H' \times W' \times D}$, where $N = M/K$, D is the number of channels and H' and W' are the new height and width. By using the 3DCNN, we reduced the length of the sequence by a factor of K, which also reduced the complexity of the model. The 3DCNN does not just reduce the length of the input sequence, it also captures local spatio-temporal low-level features. The 3DCNN captures these particulate features due to its focus on the local characteristics (space and time) of the sequence, while the transformer focuses on the global characteristics. On the whole, we have reduced the computation without losing essential information that is important for classification. Table 4.2 summarizes the architecture used inside the 3DCNN.

4.3.3 Convolutional BERT

After getting the output S of the 3DCNN, it is time to feed it to what we call convolutional BERT which stands for Convolutional Bidirectional Encoder Representations from Transformers. Transformer and self-attention have become one of the main models that revolutionize deep learning in the last few years, especially in neural language processing (NLP). Self-attention Bahdanau et al. (2014), also known as intra-attention, is an attention mechanism that connects different positions in a single sequence to compute a representation of the sequence. Here, "attention" refers to the fact that in real life, when viewing a video or listening to a song, we frequently pay more attention to certain details while paying less attention to others, based on the importance of the details. Deep learning uses a similar flow for its attention mechanism, giving particular parts of the data more focus as it is processed. Our intention in using this mechanism is for the model to focus more on the changes happening in the image sequence to better discriminate between astronomical objects. Self-attention layers are the foundation of the transformer block design. Transformers were first introduced by Vaswani et al. (2017), using model-based attention dispensing with recurrence and convolutions entirely. Their work inspired others who used the concept of transformers to achieve even better results. For example, in BERT Devlin et al. (2019) the authors used only the encoder block by stacking many of them. Even though transformers were widely used in NLP in the last two years, people started implementing these blocks in other domains like image classification. Dosovitskiy et al. (2021a) presented a model free from convolutions by using only a transformer to classify images. Garnot et al. (2019) also suggested that they are able to extract temporal characteristics using a custom neural architecture based on self-attention instead

of recurrent networks. Their use was not limited to image classification; action recognition was also investigated as in Sharir et al. (2021), where the authors used a transformer-based approach inspired by the work of Dosovitskiy et al. (2021a). Liu et al. (2021b) did propose a new transformer where they added convolution to the attention mechanisms, making it able to apply convolutions while extracting the temporal features.

4.3.3.1 Positional encoding

Because transformers have no recurrence throughout the thumbnail sequence, some information about each thumbnails relative or absolute position must be injected into the feature map obtained by the 3DCNN to inform the model about the order in the sequence. Similarly to the original transformer paper Vaswani et al. (2017), we use positional encoding at each layer in the encoder to achieve this. The only difference is that our positional encoding is a 3D tensor, where $P \in \mathbb{R}^{N \times H' \times W' \times D}$. Because the positional encoding and the new feature maps have the same dimension, they can be added together. We use sine and cosine functions to encode the position Vaswani et al. (2017):

$$P_{(n,2i)} = sin(n/10000^{2i/D}), \tag{4.4}$$

$$P_{(n,2i+1)} = cos(n/10000^{2i/D}), \tag{4.5}$$

where n denotes the position in the sequence of length, N, and i is the channel dimension, while D represent the total number of channel gotten by the 3DCNN. The sinusoidal positional encoding is chosen to make it easy for the model to learn

to attend to relative positions. To get the new input for the convolutional BERT, we conducted an element-wise addition between the positional encoding and the feature maps obtained from 3DCNN to obtain the new tensor $F \in \mathbb{R}^{N \times H' \times W' \times D}$:

$$F_n = S_n + P_n, \qquad n \in \{1, .., N\}. \tag{4.6}$$

In this study, we only used information about the position of the image in a sequence. While the observation date could be used as an alternative to the position, this would require adjusting the positional encoding function. Our experiments on the SDSS dataset did not reveal any improvement in the model when using the observation date, as opposed to just using the position. This can be understood because we do the training and the test with the same observation sequence and the network can therefore learn this sequence. On the other hand, not incorporating any information regarding the order of the sequence greatly degraded the performance of the model. As a result, we ultimately chose to use only the position in our model (see Section 4.4.2 for a discussion).

The newly obtained sequence F is fed to a multi-head convolutional attention, which is an improved self-attention that has convolution. Then the multi-head convolutional attention is followed by the second component which is a tiny feed-forward network (FFN) that has convolutions applied to every attention map. Its primary purpose is to transform the attention map into a form acceptable by the next convolutional BERT layer, with the FFN consisting of two convolutional layers with ReLU activation in between.

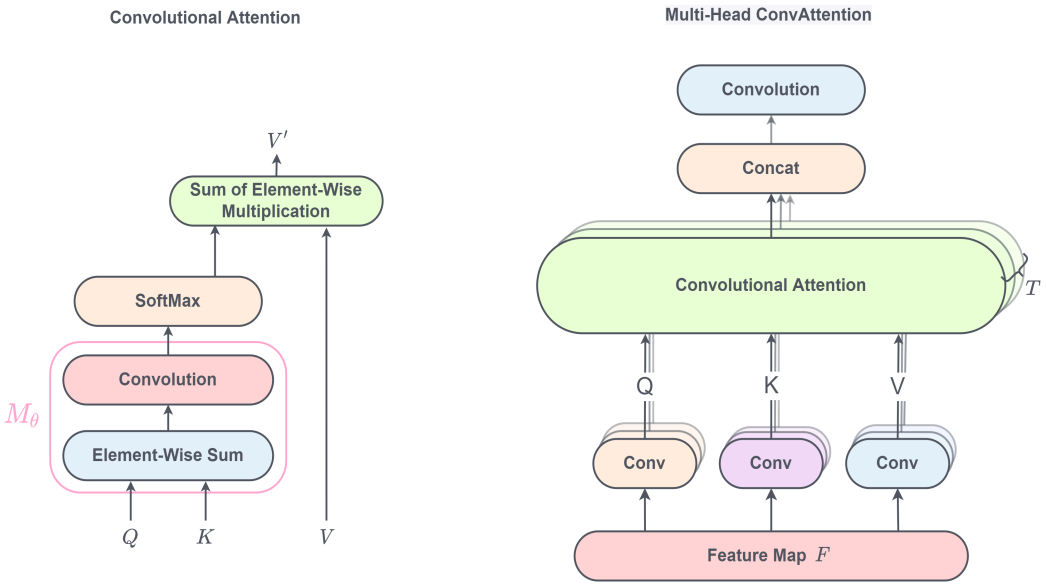

Figure 4.5: Convolutional attention (left). Multi-head convolutional attention (right). To obtain the query, key, and value maps, we applied a convolution layer on the feature map obtained from 3DCNN.

4.3.3.2 *Multi-head convolutional self-attention*

For this process, we used the model proposed by Liu et al. (2021b), with a few modifications where we replaced the last linear layer with a convolution layer. We believe that convolution in self-attention is better than the dot product between the query and the key because the convolution will accurately calculate the similarity, especially when we have 3D feature maps. A query map and a set made up of a pair of key maps and value maps that are encoded to an output using convolutional self-attention. The query map, key maps, value maps, and output are all 3D tensors. Figure 4.5 represent the general architecture of the multi-head ConvAttention.

We used a convolution layer to generate the attention models query, value, and key. The input to the attention model is $F \in \mathbb{R}^{N \times H' \times W' \times D}$. We pass each map through a convolution layer to get $\{Q, K, V\} \in \mathbb{R}^{N \times H' \times W' \times D'}$, where $D' = D/T$

and T represent the number of attention heads. Then we applied a subnetwork, M_θ, on the query and the key maps, which consists of an element-wise sum of the query and the key maps followed by another convolution layer to generate our attention map $H_{(n,m)} \in \mathbb{R}^{H' \times W' \times 1}$:

$$H_{(n,m)} = M_\theta(Q_n, K_m), \qquad n, m \in \{1, .., N\}. \tag{4.7}$$

After getting all the map attentions, $H_n = \{H_{(n,1)}, H_{(n,2)}, ..., H_{(n,N)}\}$, where $H_n \in \mathbb{R}^{H' \times W' \times N}$, we applied a softmax operation along the third dimension of size, N. Then we conducted an element-wise product between the attention map and the value map following the equation:

$$V'_n = \sum_{m=1}^{N} SoftMax(H_n)_{(n,m)} V_m. \tag{4.8}$$

We concatenated the new value representation, V'_n, obtained from the different attention heads. The multi-head attention is used to attend to input from various representation subspaces jointly:

$$MultiHead(Q, K, V) = Concat(V'_{n_1}, ..., V'_{n_T}). \tag{4.9}$$

Finally, we applied a convolution layer for merging the output of the multi-head and obtaining a high-level representation that groups all the heads. At the end of the network, we pass the encoded sequence to 3D max-pooling and finally to the classifier to make a prediction.

4.3.4 Evaluation metrics

Accuracy is the probability that an object will be correctly classified. It is defined as the sum of the true positives plus true negatives divided by the total number of individuals tested:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{4.10}$$

where TP, TN, FP, and FN are, respectively, the true positive, true negative, false positive, and false negative.

The F1 score is a classification accuracy metric that combines precision and recall. It is a suitable measure of models tested with imbalanced datasets:

$$Precision = \frac{TP}{TP + FP}, \tag{4.11}$$

$$Recall = \frac{TP}{TP + FN}, \tag{4.12}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \tag{4.13}$$

4.4 EXPERIMENTS

4.4.1 Implementation details

The supernovae in our data are not all spectroscopically confirmed, which means that the unconfirmed ones might contain some misclassified objects due to errors from the photometric typing. The model may not generalize due to this data bias. To ensure that our model performs a generalization only on spectroscopically con-

firmed data, we split up the training process into two steps. We divided the data into two datasets. The first one contains only the photometrically typed data and the second contains spectroscopically confirmed data. We trained the model at first with the photometrically typed data, then we used transfer learning to fine-tune the model on only spectroscopically confirmed data (Table 4.3 summarizes the partition of the data). The models are trained using cross-validation of five folds and three ensembles in each fold. All the architectures presented in this paper follow this same process and are implemented using PyTorch Paszke et al. (2019).

Class	Train	Fine-tune	Test
AGN	362	362	182
SNIa	1448	400	99
Variable	1290	1290	645
SNOther	2041	72	17

Table 4.3: Count of every object in a dataset of each step in training protocol. Train contains only photometrically typed data, "fine-tune" and "test": contain only spectroscopically confirmed data.

We performed an extensive hyperparameter tuning of over 20 models to specify the best hyperparameters for our architecture, which contains 1.3 Million parameters. We conducted a hyperparameter optimization using only a non-confirmed dataset with different parameters, such as sequence length, M, learning rate, lr, 3DCCN sub-sequence length, K, classifier layers' size, number of ConvBERT layers, L, number of Multi-head ConvAttention, T, batch size, and dropout. We used an Adam optimizer Kingma & Ba (2017), with a value of the learning rate of 10^{-3}, and we trained the model with cross-entropy loss and a dropout of 0.3. Hyperparameter tuning involves the number of images K that feed the 3DCNN and the maximum length of the sequence. The best values were $K = 3$ and $M = 99,$ which

means the number of sequences for the convolutional BERT is $N = 33$. The batch size was 128 sequences which we ran over 100 epochs. We chose the number of convolutional BERT layers to be $L = 2$ and the number of attention heads $T = 4$. Also, the images were normalized band-wise, as each band has different characteristics. We used only four classes (AGN, SNIa, Variable, SNOther) to train all the models. The class marked as "unknown" has not been considered in the study. It corresponds to noisy or very sparse data. It can easily be tagged from sparsity or noise in the image metrics and we do not expect any improvement in the classification if such objects are added to the training. We trained all models with 4 GPUs GeForce RTX 2080 Ti. Each model takes about three hours to complete training. The implementation will be released upon publication in our Github page [1].

4.4.2 Results

This section provides studies on SDSS comparing the accuracy and F1 score of our proposed solution with other works. Table 4.4 summarizes the result of **different models from different deep learning areas** to diversify our benchmark as it contains RNN architectures (SuperNNova, LSTM), CNN-based models such as SCONE, Hybrid models that have CNN and RNN such as Carrasco-Davis et al. (2019) and Gómez et al. (2020), and, finally, a transformer-based model. Also, we compared the result using two types of datasets: first, the image dataset and, second, the same dataset object but with the light curves; the goal is to highlight the advantage of using images instead of light curves. Moreover, the different works mentioned in Table 4.4 were initially proposed for different datasets with different classes and training protocols. Hence, the results do not reflect the quality of these

[1] https://github.com/DaBihy/ConvEntion

Model	Bands	Type of data	Accuracy	F1 Score	Num params
ConvEntion (Ours)	ugriz	Images	**79.83**	**70.62**	1.253M
CNN+GRU Gómez et al. (2020)	ugriz	Images	66.39	63.22	1.993M
ConvEntion (Ours)	g	Images	76.89	63.20	1.253M
CNN+GRU Gómez et al. (2020)	g	Images	63.67	61.00	1.992M
CNN+LSTM Carrasco-Davis et al. (2019)	ugriz	Images	64.08	60.65	2.190M
CNN+LSTM Carrasco-Davis et al. (2019)	g	Images	63.00	60.00	2.189M
SuperNNova (Bayes) Möller & de Boissière (2020)	ugriz	Light curves	65.54	55.40	-
SITS-BERT Yuan & Lin (2021c)	ugriz	Light curves	67.43	51.60	0.596M
SCONE (CNN) Qu et al. (2021)	ugriz	Light curves	62.57	50.43	22.2K
SuperNNova (RNN) Möller & de Boissière (2020)	ugriz	Light curves	56.30	42.60	-
LSTM	ugriz	Light curves	55.24	40.33	60K

Table 4.4: Performance comparison in terms of average F1 score and the average of the accuracy of five folds of cross-validation. This table includes only experiments on a dataset with four classes.

Model	Bands	Accuracy	F1 Score
ConvEntion (Ours)	ugriz	**83.90**	**75.77**
ConvEntion (Ours)	g	79.47	72.38
CNN+GRU Gómez et al. (2020)	g	74.84	68.95
CNN+LSTM Carrasco-Davis et al. (2019)	g	73.94	67.29

Table 4.5: Performance comparison in terms of average F1 score and the average of the accuracy of five folds of cross-validation. This table includes only experiments on a dataset with three classes (AGN, SN, Variable).

works on other datasets. The goal of the comparison is to give visibility into the performance of our model from a deep learning standpoint and the importance of using image time series from an astronomy perspective.

Overall, our model ConvEntion obtains the highest accuracy of 79.83% and F1 score of 70.62%, 13 points higher in accuracy than the best results on images by Gómez et al. (2020) and 12 points higher in accuracy than the best model using light curves. This confirms the advantage of using images over light curves.

This advantage can be explained by the fact that the image contains more information than a single value of flux in a light curve. Hence, a model can learn robustly with the existence of more high-level feature maps. Also, ConvEntion performed better compared to the other image-based models, such as Carrasco-Davis et al. (2019). Additionally, transformers give a remarkable computational advantage because they avoid recursion and allow for parallel computation, thus reducing the training time. Our model took only three hours to train, compared to other image-based models which took five hours of training on our GPUs. Our model achieved better results using fewer parameters, compared to the other mod-

Model	Accuracy	F1 score ˆ	Run time
ConvEntion	**79.83**	**70.62**	1.5
No Oversampling	79.36	64.23	1.5
No Band Embedding	70.74	59.85	1.5
Fixed Band Embedding	78.45	65.73	1.5
2D CNN	77.38	62.25	4.5

Table 4.6: Different ablation experiments to show the impact of each component in our model

els trained on image sequences. The main benefit of using a transformer is that it reduces the drop in performance due to long dependencies. Transformers do not rely on past hidden states to capture dependencies with previous features such as RNNs. They instead process a sequence as a whole. Therefore, there is no risk of losing past information. Also, the integration of a spatio-temporal feature extraction helped in getting a better high-level representation of the sequence, in comparison to separating the spatial features from the temporal ones. The two types of features have correlations that may help the model to better discriminate between objects. We can also highlight the importance of separating the band to mitigate the impact of missing observations. Our model performed well, in comparison to that of Gómez et al. (2020) which uses multiple bands, which shows that separating the bands and adding band embedding works better than feeding the network with empty bands.

In the study of Carrasco-Davis et al. (2019), the authors trained their model on a dataset that only has a "g" band and they noted that the model can be adapted to classify the image sequence combining information using multiple bands. For the sake of comparison, we trained the image models with all the bands "ugriz" at first and then with only one "g" band. Our model achieved an accuracy of 76.89%

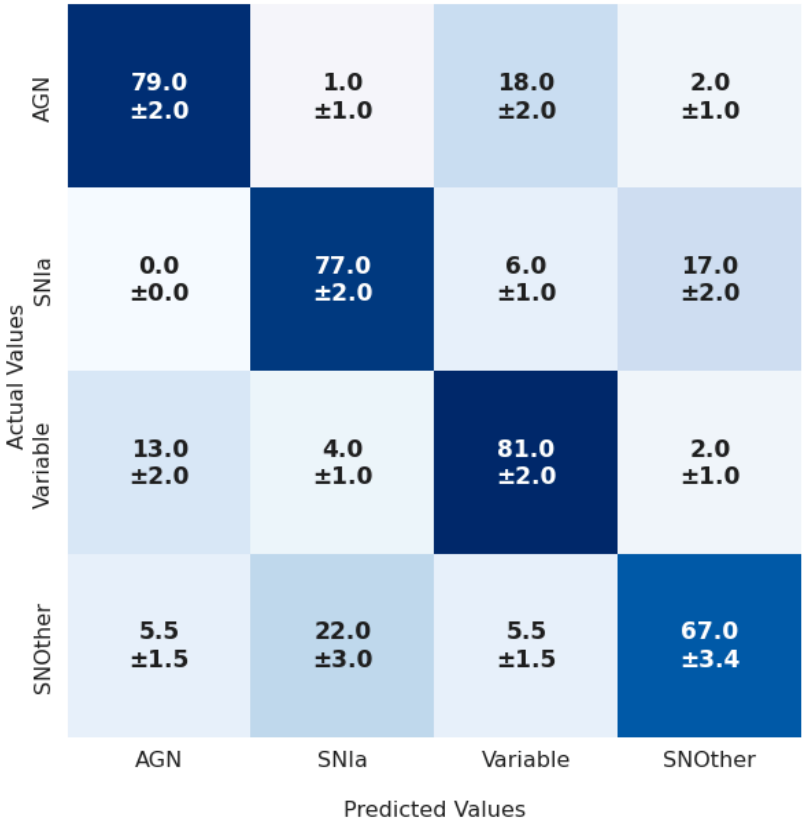

Figure 4.6: Confusion matrix showing the average accuracy and standard deviation of the predictions generated by ConvEntion over cross-validation of five folds on test data.

and 63.20% in the F1 score using one band ("g") which dropped 7% in comparison to using multiple bands. Meanwhile, Carrasco-Davis et al. (2019) achieved 63% in accuracy and 60% in their F1 score. This shows that our model is more efficient when using multiple bands. This also highlights the impact of band separation to mitigate the impact of the missing observations.

Figure 4.6 illustrates the obtained confusion matrix by ConvEntion and it shows that the model has well classified the supernovas. Most of the misclassified SNIa are associated with SNOther and vice versa, which is not a serious error. This is even an expected behavior, especially since all types of supernovas share a lot of

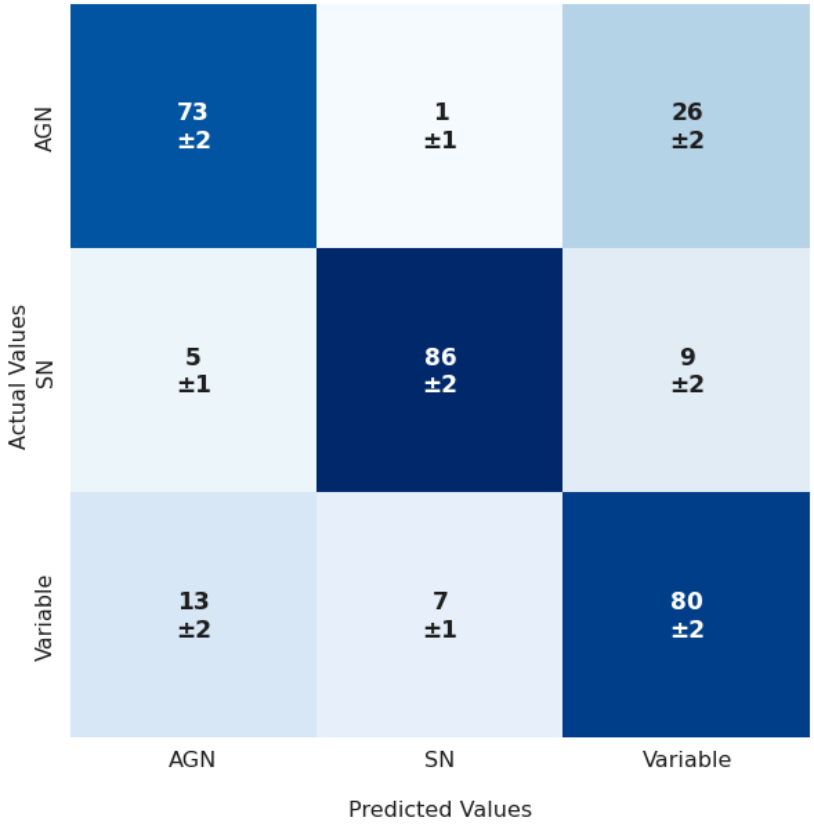

Figure 4.7: Confusion matrix of three classes showing the average accuracy and standard deviation of the predictions generated by ConvEntion over cross-validation of five folds on test data.

similarities which may confuse the model. Additionally, with a small dataset like ours, it is normal to have such behavior because the model does not have enough samples to totally discriminate among objects. Meanwhile, variables were the best-classified class in our dataset, with just a bit of confusion with the AGN; this misclassification between AGN and variable can be explained by the class imbalance in our dataset based on the knowledge that the number of variables is higher than in the other classes.

Table 4.5 summarizes the results of different models trained only on three classes (AGN, SN, Variable), where classes SNIa and SNOther are combined into a single

class. The goal of this experiment is to see the behavior of our model in discriminating between transient and non-transient objects. We got the best results with an accuracy of 83.90% with an F1 Score of 75.77%. The model was able to classify the SN accurately, with a score of 86% (as shown in Figure 4.7).

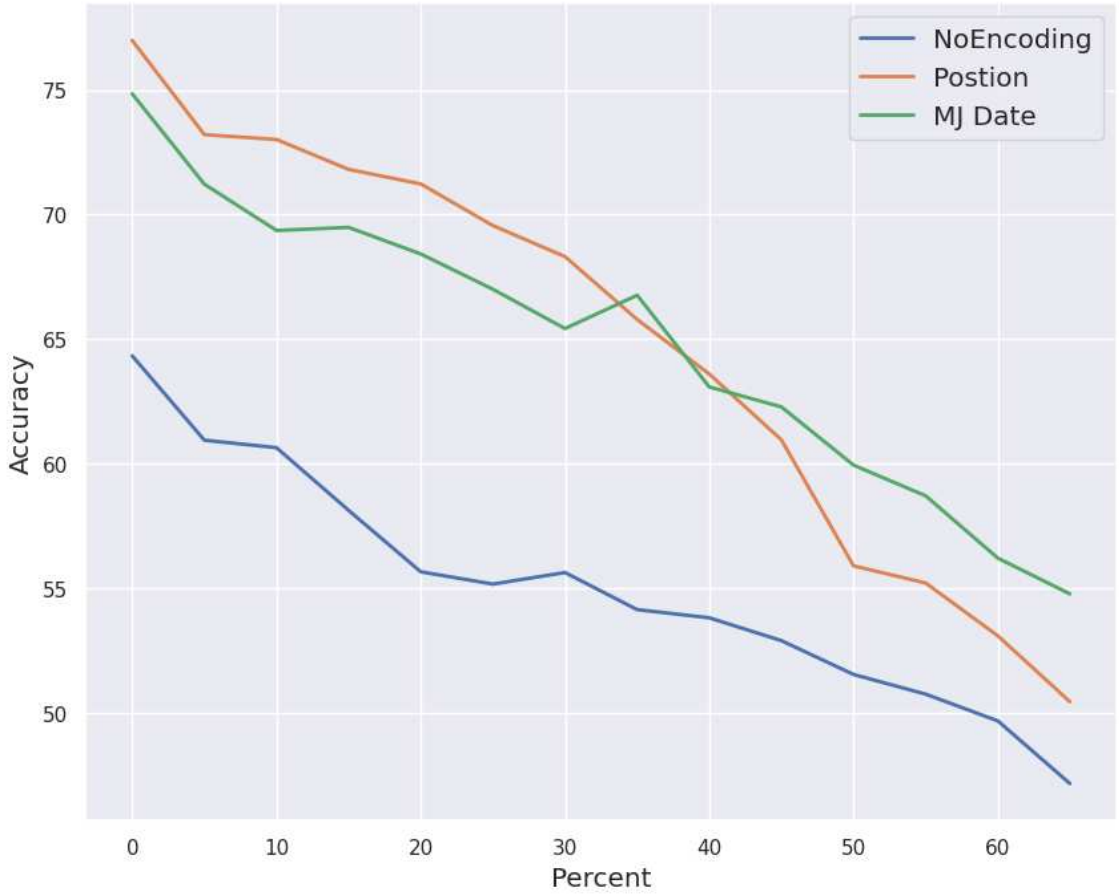

Figure 4.8: Comparison of model accuracy as a function of missing observation percentage: Positional Encoding, MJD Encoding, and No Encoding. As the percentage of missing observations increases, all models exhibit a decline in accuracy, but models with Positional and MJD encoding perform significantly better than the model with no encoding, demonstrating their effectiveness in dealing with missing data

The ablation study 4.6 provides a comprehensive understanding of the effects of various modifications made to the ConvEntion model. By making certain changes

to the model, and comparing the resultant performance metrics, we can infer the contributions of each component to the model's performance. The table reveals that ConvEntion, which is the full-featured model, outperforms the other variations in terms of both accuracy (79.83%) and F1 score (70.62%). It also shares the same run time (1.5) as all other variants, except the 2D CNN, indicating it maintains its efficiency despite providing superior performance.

Looking closely at the performance of other models, the implementation of oversampling, a technique often used to counteract class imbalance, in the ConvEntion model appears to play a significant role in enhancing the model's precision and recall, as demonstrated by the F1 score. Specifically, the "No Oversampling" variant exhibits only a modest reduction in accuracy relative to ConvEntion (0.47% drop), but a markedly larger decrease in the F1 score (6.39% drop). This underscores that while the absence of oversampling doesn't drastically influence overall accuracy, it considerably impacts the balance between precision and recall, highlighting the crucial role of oversampling in addressing class imbalance and thereby optimizing model performance. On the other hand, "No Band Embedding" undergoes a significant decrease in both accuracy and F1 score (9.09% and 10.77% drop respectively), suggesting that band embedding is a crucial feature for model performance.

The "Fixed Band Embedding" variant exhibits a more modest reduction in both accuracy and F1 score compared to ConvEntion, with a drop of 1.38% and 4.89% respectively. This suggests that having a dynamic or adaptable band embedding, as opposed to a fixed one, contributes to better performance. It's worth noting that the Fixed Band Embedding is static compared to the original Band Embedding, which is a trainable dense layer. This difference in flexibility could account for the

performance disparity. Lastly, the 2D CNN version, despite having a longer run time (4.5), performs less effectively, with a reduction in both accuracy and F1 score compared to ConvEntion (2.45% and 8.37% respectively). This finding reinforces the good performance of ConvEntion not only in metrics but also in computational efficiency.

Based on the provided data in figure 4.8, it appears that as the percentage of missing observations increases, the accuracy of all the models decreases. This trend is expected as missing observations would likely reduce the effectiveness of the models. However, its also clear that the models using positional and MJD encodings perform better compared to the model with no encoding. MJD encoding, or Modified Julian Date encoding, represents the continuous count of days since the beginning of the Julian Period on November 17, 1858. It's often used in astronomical studies to pinpoint the date of an object's observation. Specifically, the model with positional encoding performs best when the missing observation percentage is low. Meanwhile, the model with MJD encoding maintains better performance as the percentage of missing observations increases. Its interesting to note that the MJD model shows a relatively stable trend in accuracy as the missing observation percentage increases, suggesting that it may handle missing data more effectively. Overall, while all models suffer from increasing missing observation percentages, those with positional and MJD encoding still outperform the model without any encoding.

The model is able to effectively process a given survey without any loss in performance and without the requirement of providing it with the time information for each image. However, when there is a covariate shift, or a mismatch, between the training set and the test set as when using a different dataset with a different

observation sequence), incorporating the time information can improve the results.

4.5 CONCLUSION

In this chapter, we present a method for efficient astronomical image time series classification that is entirely based on the combination of convolutional networks and transformers. Inspired by action recognition and satellite image time series classification, we propose a model ConvEntion that utilizes convolutions and transformers jointly to capture complex spatio-temporal dependencies between distinct steps, leading to accurate predictions based on different observations of an object. The accuracy of our model is better with a high margin of 13%, in comparison to state-of-the-art methods using image data – and even better compared to approaches using light curves.

Our model achieves good results on the SDSS dataset, while also being faster thanks to using fewer parameters and parallel computational processes, making it a good candidate for latency-sensitive applications such as the real-time thumbnail classifier of astronomical events. Meanwhile, our benchmark stands as clear evidence of the importance of images in the domain of astronomy. Indeed, the images contain more information than the normal light curves, even if they present more difficulties.

CHAPTER 5

Semi-Supervised Image Time Series Representation Learning Using Convolutional Transformers

5.1 INTRODUCTION

Deep learning has made significant strides in various domains, including natural language processing and computer vision. Among the models that have emerged from this field, the ConvEntion model has demonstrated its capabilities in image classification tasks. However, it faces a challenge with class imbalance, a common issue that can notably impact the performance of a model. This limitation becomes more evident when dealing with image sequences, a key area of focus for ConvEntion.

The issue of class imbalance becomes even more prominent when working with small datasets. Large datasets often provide a more balanced representation of classes, but when the dataset is small, the class imbalance can skew the model's learning, leading to less than optimal performance. This is a challenge that we face in our research, as we are working with small datasets.

In recent years, self-supervised learning has emerged as a potential solution to these challenges. Methods such as BYOL (Grill et al., 2020) and DINO (Caron et al., 2021) have shown the potential of self-supervised learning in handling large datasets. However, their performance on small datasets could be improved.

This is where our research comes in. Drawing inspiration from self-supervised learning methods like BYOL and DINO, we propose a semi-supervised learning technique that is specifically designed to handle class imbalance in small datasets. This approach, which we call the Semiconformer, leverages the capabilities of vi-

sion transformers to provide a robust solution for image time series representation learning.

The Semiconformer is not just an enhancement to the ConvEntion model. It is a versatile solution that can be applied to any model grappling with class imbalance, particularly when dealing with small datasets. By leveraging the principles of self-supervised learning (Grill et al., 2020; Caron et al., 2021; Bao et al., 2022; Li et al., 2021; Yuan & Lin, 2021a), the Semiconformer offers a new pathway in the field of deep learning.

In the following sections, we will delve deeper into the design and implementation of the Semiconformer. We will discuss how it mitigates the impact of class imbalance, minimizes the intra-class variance of the model, and enhances the results when dealing with a small dataset, without the need for additional unlabeled data.

Our research is a novel approach that seeks to exploit the potential of Vision Transformers in a semi-supervised setting. It opens up new pathways in the field of deep learning, offering a robust solution to the challenges of class imbalance and small datasets.

In conclusion, this chapter will provide a comprehensive introduction to the Semiconformer, a semi-supervised learning technique that promises to address the challenges of class imbalance in small datasets.

5.2 OUR APPROACH

5.2.1 Description of Semiconformer

In this section, we detail the architecture and operations of our proposed Semi-Convolutional transFormer (SemiConFormer), a semi-supervised learning frame-

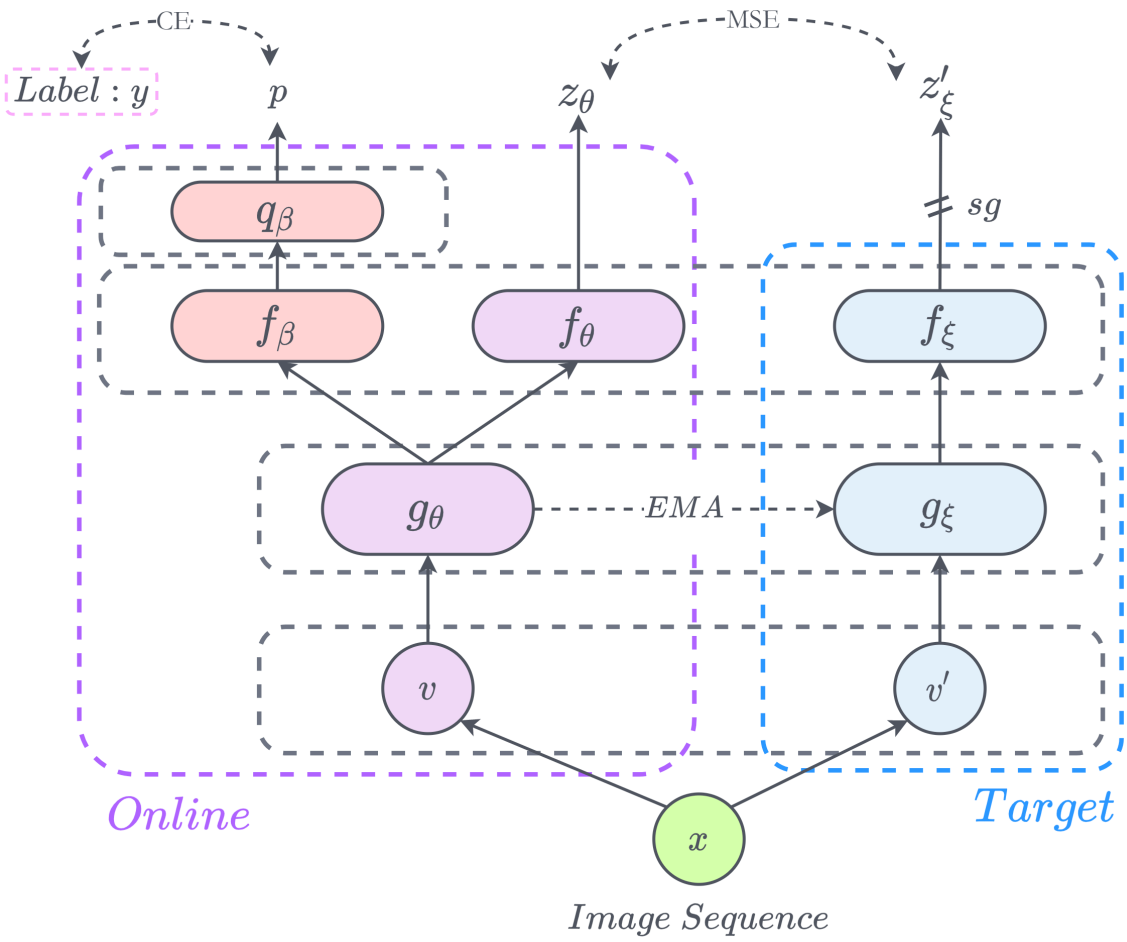

Figure 5.1: The general architecture of the Semiconformer. Semiconformer's architecture involves minimizing the difference between z_θ and z'_ξ using mean squared error, also minimizing the cross entropy between y and p , where θ represents trained weights, ξ represents an exponential moving average of θ, and sg means stop-gradien.

work specially tailored to address the challenges associated with small data sets and class imbalances. The structure of SemiConFormer bears similarities to recent self-supervised learning approaches (Grill et al., 2020; Caron et al., 2021; Bao et al., 2022; Li et al., 2021; Yuan & Lin, 2021a), but the core difference lies in its design to work effectively with smaller data sets, a feat not easily achievable by traditional self-supervised methods.

The SemiConFormer, like contrastive methods (Chen et al., 2020b; He et al., 2019), employs an integrated data augmentation for representation learning. The process commences by creating two augmented image sequences (views) (See figure 4.2), namely v and v', from the original sequence. These views are generated using a combination of data augmentation techniques such as dropping random steps from the sequence, rotating the sequence, and flipping it both horizontally and vertically. Additionally, the sequence is shifted to create a shorter sequence from the original, contributing to the diversity of views.

Within our SemiConFormer framework, we utilize two neural networks, specifically an online network and a target network, a concept akin to the "Bootstrap Your Own Latent" (BYOL) method (Grill et al., 2020). The online network, defined by a set of weights θ, β, is comprised of three stages.

In the first stage, we employ an encoder g_θ based on convolution transformers. The primary role of this encoder is representation learning. The second stage includes two projection heads f_β and f_θ, which project the representations obtained from the encoder. The first projection f_β is leveraged for classification using ground truth labels, while the second projection f_θ minimizes the distance between the online and target networks. By separating these projections, we circumvent the learning collapse that may be encountered in methods such as Mean Teacher (Tarvainen & Valpola, 2017), which is attributed to the disparate nature of the loss functions involved in classification and representation learning. In the third stage, a prediction layer q_β is used for class prediction.

Conversely, the target network, defined by a set of weights ξ, encompasses two stages - an encoder g_ξ for representation learning and a projection head f_ξ. It should be noted that g_θ, f_θ share the same architecture respectively with g_ξ and f_ξ.

Upon generating the two augmented views v and v' of the image sequence, we feed v' into the target network. It passes through the encoder and the projector, producing an output z'_ξ. Simultaneously, v is fed into the online network, generating a projection z_θ and a class prediction p. These projections are then normalized, resulting in $\tilde{z}_\theta \triangleq z_\theta/||z_\theta||_2$ and $\tilde{z'}_\xi \triangleq z'_\xi/||z'_\xi||_2$, which are utilized to calculate the similarity loss using a Mean Squared Error (MSE) function:

$$\mathcal{L}_{MSE} \triangleq ||\tilde{z}_\theta - \tilde{z}'_\xi||_2^2 = 2 - 2 \cdot \frac{\langle z_\theta, z'_\xi \rangle}{||z_\theta||_2 \cdot ||z'_\xi||_2} \tag{5.1}$$

$\langle z_\theta, z'\xi \rangle$ is the dot product of $z\theta$ and z'_ξ, which measures the similarity between the two vectors. The MSE loss measures the mean squared difference between the normalized projections of the online and target networks. It is minimized during training to bring the projections of the online and target networks closer together, thereby aligning the learned representations of the two networks.

Further, the classification loss is computed using a Cross-Entropy (CE) function between the ground truth and the online network prediction p:

$$\mathcal{L}_{CE} = -\sum_{j=1}^{C} y_j \log(p_j) \tag{5.2}$$

Where C is the number of classes.

To enhance the robustness of the learning, both the MSE loss $\mathcal{L}_{MSE}$ and the CE loss $\mathcal{L}_{CE}$ are symmetrized by feeding v' to the online network and v to the target network, yielding $\tilde{\mathcal{L}}_{MSE}$ and $\tilde{\mathcal{L}}_{CE}$. This is done to ensure that the learning process is not biased towards any particular view and that the model learns to generate consistent representations regardless of the specific augmentation applied. This symmetrization helps to make the learning process more robust and stable. It also

ensures that the model learns to recognize the same underlying content even when it is presented in different ways due to the augmentations. The overall loss function for our framework is then computed as:

$$\mathcal{L} = \mathcal{L}_{MSE} + \mathcal{L}_{CE} + \tilde{\mathcal{L}}_{MSE} + \tilde{\mathcal{L}}_{CE} \tag{5.3}$$

We use a stop-gradient (sg) operator on the target network to ensure that the gradients are only propagated through the online network. This means that during the training process, when the gradients are backpropagated through the network to update the weights, the gradients do not propagate through the target network. This is done because the target network's weights are not directly updated through backpropagation. Instead, they are updated as an exponential moving average of the online network's weights. Unlike traditional knowledge distillation, we do not rely on pre-trained weights ξ. Instead, the weights ξ are updated as follows:

$$\xi \leftarrow \tau\xi + (1-\tau)\theta. \tag{5.4}$$

In this relation, $\tau \in [0, 1]$ represents the decay rate, which determines the extent to which the online network parameters influence the updating of the target network parameters. This approach ensures that the target network continually learns from the online network, maintaining the network's dynamism and promoting effective learning

The use of two networks, namely the online and target networks, instead of one network is a common strategy in self-supervised learning methods. This dual network structure serves to stabilize the learning process and improve the performance of the model. The online network is the one actively learning and updat-

ing its parameters, continuously adapting to new data, hence the name "online". The target network, on the other hand, provides the regression targets to train the online network. Its parameters are an exponential moving average of the online parameters, making it a slowly changing version of the online network. This slow adaptation helps to stabilize the learning process by providing a consistent target for the online network to learn from. Furthermore, the use of two networks allows for an asymmetric architecture, which has been shown to improve performance in self-supervised learning tasks.

BYOL (Grill et al., 2020) method differs from the SemiConFormer in its approach to loss calculation, specifically in relation to the use of ground truth labels. While SemiConFormer uses the projection head f_θ, to calculate the Cross-Entropy (CE) loss using ground truth labels, BYOL does not incorporate this step. This is because BYOL is a self-supervised learning approach, which means it learns representations from the data itself without the need for explicit labels or ground truth. This fundamental difference in the use of ground truth labels for loss calculation underscores the distinct methodologies employed by BYOL and SemiConFormer in their respective learning frameworks.

5.2.2 Network architecture and Backbone

In the implementation of our SemiConFormer framework, we employ as the backbone for g, the $ConvEntion$ model $ConvEntion$ (Bairouk et al., 2023) that we previously devised. The ConvEntion model merges convolutional and self-attention mechanisms, addressing challenges linked to image time series classification, and missing observations.

The ConvEntion model processes an input sequence of images, denoted as

$J \in \mathbb{R}^{M \times H \times W \times C}$, where M signifies the length of the sequence and H, W, and C respectively correspond to the image's height, width, and number of channels. Initially, a 3D Convolutional Neural Network (CNN) processes this sequence, simultaneously diminishing the sequence's length and capturing local object characteristics. This operation yields a new sequence $S \in \mathbb{R}^{N \times H' \times W' \times D}$, where $N = M/K$ indicates the reduced length, D stands for the number of channels, and H' and W' are the newly computed height and width.

The reduction factor, represented by K, influences the operation of the 3D CNN. The CNN takes K inputs of J and generates a single entry of S, resulting in a shortened sequence length. After this reduction, the sequence S is passed through a convolutional-BERT architecture. This operation aids in the extraction of high-level spatial-temporal features from the input, enriching the representation learned by the network.

The projection heads of our model, denoted as f_β, f_θ, and f_ξ, each consist of a pair of fully connected layers. A Rectified Linear Unit (ReLU) activation function is utilized within these layers, accompanied by a Dropout mechanism to prevent overfitting and ensure robust learning.

Within the online network, we incorporate a prediction layer, q_β, that consists of a singular fully connected layer. This layer accepts the output of f_β and generates an output whose dimensions correspond to the number of classes in the dataset. By linking the output to the number of classes, the prediction layer can effectively predict class labels for the given inputs, facilitating effective classification.

5.3 EXPERIMENTS

5.3.1 Dataset

We have employed the same data set as in chapter 4 for the ConvEntion Model. We used the Sloan Digital Sky Survey (SDSS) and its SDSS Supernova Survey, which included a vast range of celestial objects imaged repeatedly over 20 years. The data set provided comprehensive coverage of various astronomical transient events, including Galactic Variable Stars, Active Galactic Nuclei (AGN), and Supernovae (SNe). We supplemented this rich image database with a spectroscopic follow-up program for more precise identification and classification.

Considering the challenges of class imbalance and similarities among different class features, we took a two-step approach to training our model. Initially, we used photometrically typed data, acknowledging possible misclassifications. To address these, we then applied transfer learning to fine-tune the model with spectroscopically confirmed data, which enhanced the model's generalization capabilities. This process of separating and training on two data sets, as summarized in Table 4.3, ensured reliable outcomes and helped us to overcome the inherent hurdles of this complex data set.

5.3.2 Implementation details

We train the model with the adamw Loshchilov & Hutter (2018) optimizer and a batch size of 128, distributed over 4 GPUs. The learning rate was set to 2×10^{-3} and a dropout of 0.3 . We decay the learning rate with a cosine schedule Loshchilov & Hutter (2016). The decay rate for EMA is set to be $\tau = 0.99$. For the backbone $ConvEntion$ we used the same parameters used in the original paper Bairouk et al.

(2023) where $K = 3$ and $M = 99$ which means the number of sequences for the convolutional BERT is $N = 33$. The number of Conv-BERT layers is set to be $L = 2$ and the number of attention heads $T = 4$. The models are trained using cross-validation of five folds. All the architectures presented in this paper follow this same process and are implemented using PyTorch Paszke et al. (2019).

5.3.3 Results

In this section, the results of using the proposed framework are presented, and a comparison is made to using models without the framework. Additionally, various figures are presented to evaluate the effect of the proposed solution on the challenges discussed earlier.

Model	Accuracy	F1 Score
Semiconformer	**82.18**	**75.33**
ConvEntion (Transformer based)	79.83	70.62
CNN+GRU	66.39	63.22
CNN+LSTM	64.08	60.65

Table 5.1: Performance comparison in terms of average F1 Score and the average of the Accuracy of 5 folds of cross-validation.

The table in reference 5.1 illustrates a comparison of the results of using the proposed Semiconformer to using only the ConvEntion model. The proposed solution achieved an accuracy of 82.18% and an F1 Score of 75.33%. From the table, we can see that there is a clear improvement in the F1 score when using the proposed Semiconformer model, compared to the result achieved by the ConvEntion model, which was only 70.62%. This can be attributed to the fact that Semiconformers are better equipped to handle class imbalance. By training on two augmentations in

each step, the model is able to learn the distinct characteristics of each object, allowing for more accurate discrimination between classes. This is further highlighted by the confusion matrices in figures 5.2 and 5.3. It can be seen that the accuracy of the "SNIa" and "SNOther" classes has been improved without any negative impact on the majority "Variable" class.

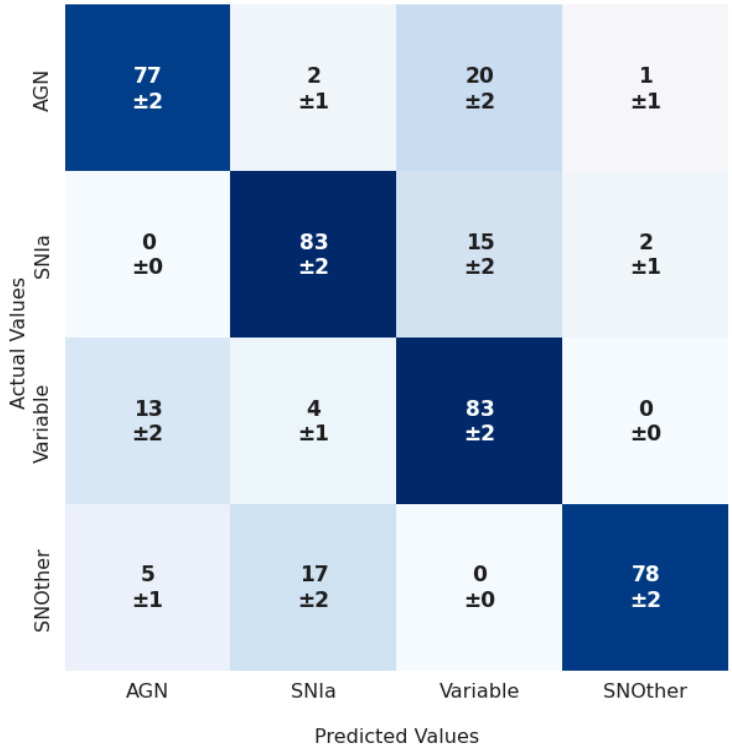

Figure 5.2: Confusion matrix showing the average accuracy and standard deviation of the predictions generated by Semiconformer over cross-validation of five folds on test data.

Figure 5.5 illustrates a comparison of the ROC curves for the classes "SNIa" and "SNOther" between the proposed Semiconformer framework and the ConvEntion model. The results for both classes have been improved, despite the similarity in their characteristic levels. This further supports the effectiveness of the proposed solution in handling the challenges of class imbalance and increasing classification

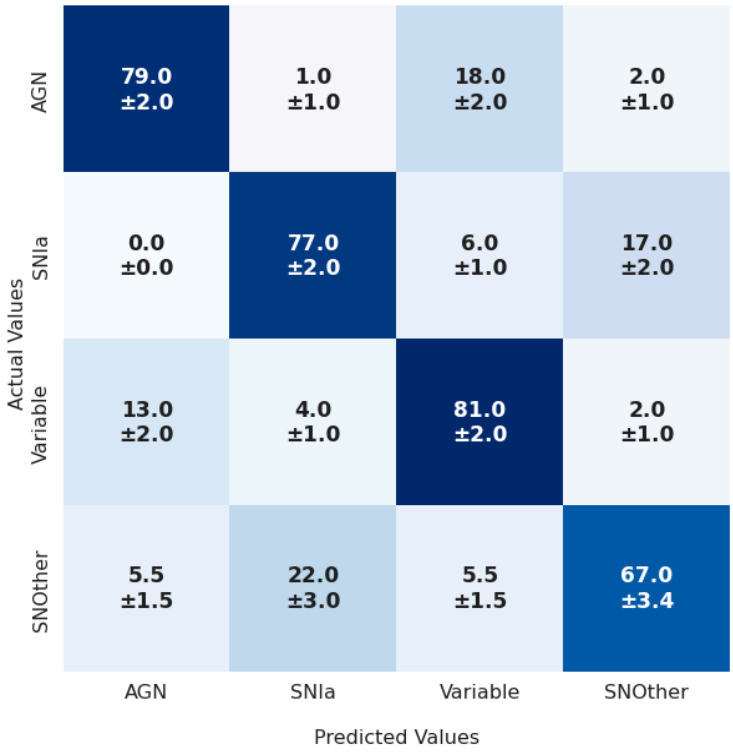

Figure 5.3: Confusion matrix showing the average accuracy and standard deviation of the predictions generated by ConvEntion over cross-validation of five folds on test data.

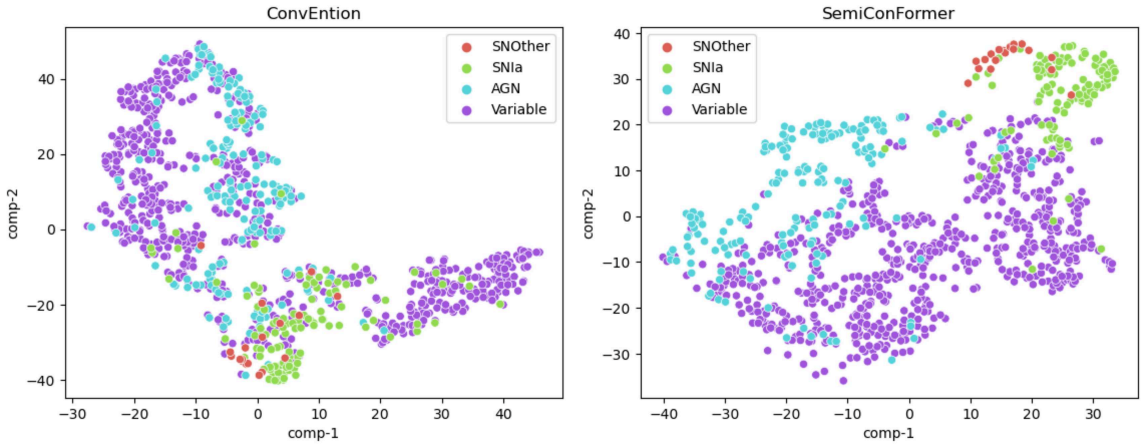

Figure 5.4: The t-distributed stochastic neighbor embedding (t-SNE) projections with features extracted from the projection f_β of the last fully connected layer of ConvEntion and Semiconformer

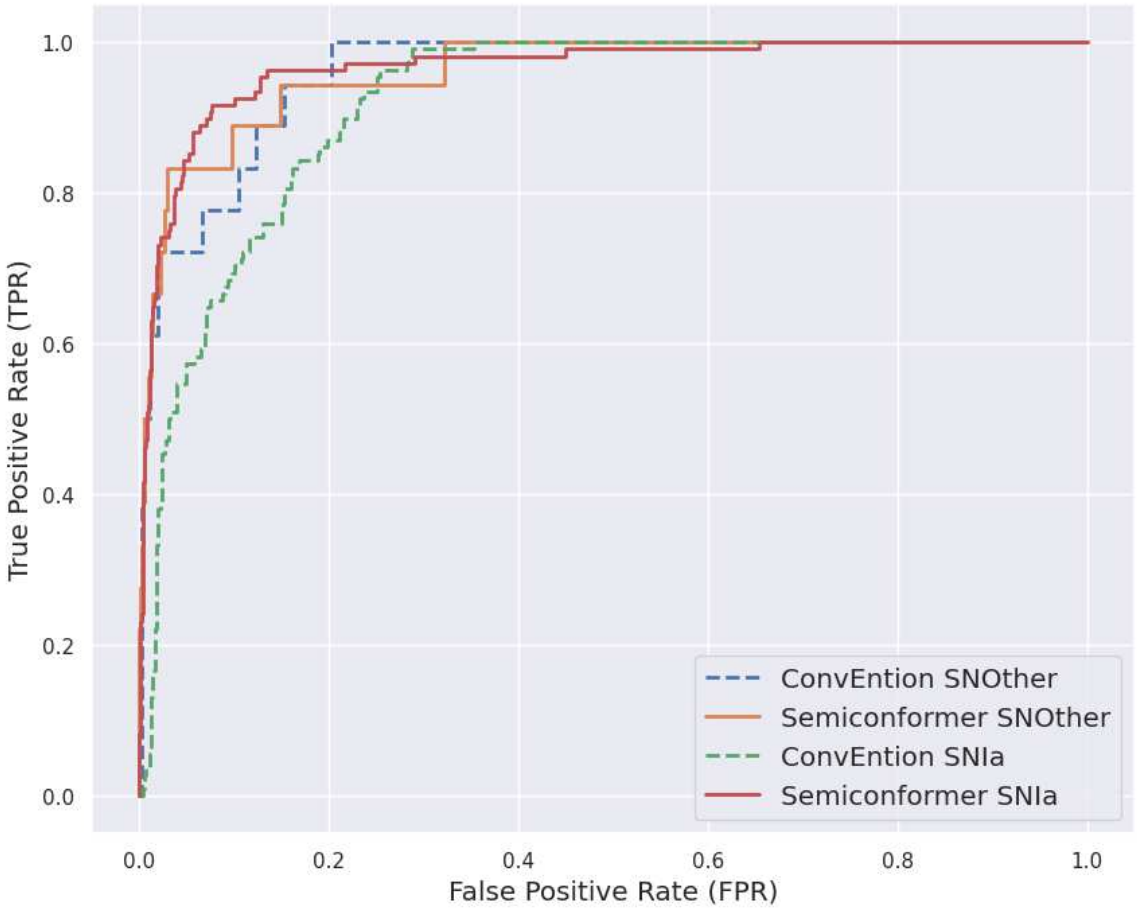

Figure 5.5: Comparison of ROC curves of classes "SNIa" and "SNOther" between the new proposed framework Semiconformer and the model ConvEntion

accuracy. This improvement in the accuracy of the classes "SNIa" and "SNOther" can be attributed to the calculated similarity loss (MSE) which is used to minimize the distance between different views of the data. This effectively reduces the intra- class variance, increasing their separability. This is supported by the results shown in Figure 5.4, which illustrates the t-SNE projection of the classes "SNIa" and "SNOther" for both the Semiconformer and the ConvEntion model. It can be observed that in the Semiconformer projection, the classes are well-localized, whereas in the ConvEntion projection, the classes are more spread out, particu-

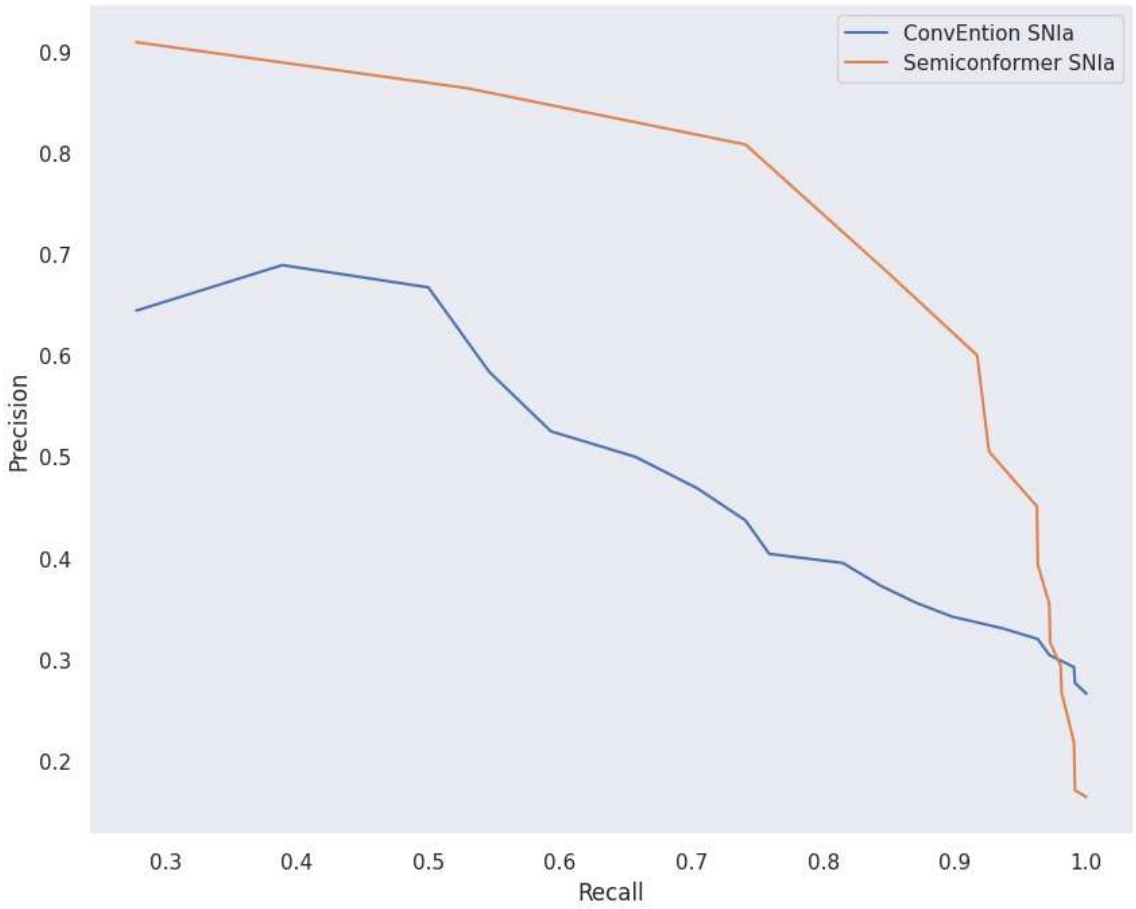

Figure 5.6: Precision-Recall (PR) analysis of Supernova Ia identification models 'ConvEntion SNIa' and 'Semiconformer SNIa'. The 'Semiconformer SNIa' model consistently achieves higher precision across various levels of recall, indicating its superior performance in balancing the trade-off between precision and recall.

larly the minority class "SNOther".

The Precision-Recall (PR) curve shown in the figure 5.6 provides a visual depiction of the performance of Supernova Ia identification in the two models: 'ConvEntion' and 'Semiconformer'. It is immediately evident that the 'Semiconformer' model outperforms the 'ConvEntion' model across nearly the entire recall spectrum. The model shows a higher precision at almost every recall value, indicating that it can identify a greater proportion of true positive results (Supernova Ia), and

at the same time, it generates fewer false positive identifications.

When analyzing the figure more deeply, we can see that both models maintain relatively high precision at the start, indicating that both are capable of delivering high-quality identifications with a low false discovery rate. However, as recall increases, the precision of the 'ConvEntion' model deteriorates at a faster pace than the 'Semiconformer' model. This suggests that while both models can initially avoid false positives, the 'ConvEntion' model struggles to maintain this performance as it attempts to capture more true positives (increasing recall). Thus, the 'Semiconformer' model offers a better balance between precision and recall, and as such, it might be more effective for practical use, especially in scenarios where both high recall and precision are desired.

The proposed Semiconformer method has yielded promising results, however, the training process is quite costly, taking three times longer than the training of the model without the proposed framework. To overcome this, further research could focus on improving the training methodology to reduce computational costs.

5.4 CONCLUSION

In conclusion, the proposed Semiconformer framework for handling image time series classification using convolutional transformers has been shown to be effective in increasing classification accuracy. The results of using the proposed Semiconformer model were compared to using only the ConvEntion model, and it was found that the proposed solution achieved an accuracy of 82.18% and an F1 Score of 75.33%. The improvement in F1 score is attributed to the semiconformer's ability to better represent the classes in the latent space which is a good property in order to be less sensitive to class imbalance, allowing for more accurate discrimination

between classes. However, the proposed Semiconformer method has the cost of a longer training process, which could be a focus for future research to overcome.

CHAPTER 6

Conclusions and Perspectives

6.1 CONCLUSION

This dissertation has made significant strides in the field of astronomical image time series classification. It has underscored the importance of using image sequences instead of light curves for the classification of celestial objects. The use of image sequences allows for the extraction of richer contextual information, which can lead to more accurate and robust classification results. This approach is particularly beneficial in the context of astronomical data, where observations are often sparse and noisy.

A key contribution of this dissertation is the ConvEntion model, a novel approach for classifying different types of space objects directly using images. The ConvEntion model leverages the power of convolutions and transformers to process astronomical image time series. It has shown impressive performance in mitigating the problem of missing observations, a common issue in astronomical data. By integrating spatiotemporal features, the ConvEntion model has achieved a significant improvement in classification accuracy, with an increase of 13% compared to state-of-the-art approaches that use image time series and a 12% increase compared to approaches that use light curves.

Another significant contribution is the Semiconformer framework, a semisupervised learning framework for image time series representation. The Semiconformer framework leverages self-supervised approaches to reduce the effect of class imbalance with data augmentation, minimize intra-class variance, and improve results on small datasets without the need for extra labeled data. It has

demonstrated superior performance compared to a conventional model, achieving an accuracy of 82.18% and an F1 Score of 75.33%.

In conclusion, this dissertation has made substantial contributions to the field of astronomical image time series classification. It has demonstrated the potential of deep learning techniques in handling complex astronomical data and has developed innovative models that significantly improve the accuracy and efficiency of classification tasks. The findings of this dissertation pave the way for future research in this area and provide a strong foundation for the further development of advanced machine learning models for astronomical image analysis.

6.2 PERSPECTIVES

Looking forward, there are several promising directions for future research and development in the field of astronomical image time series classification.

Expanding the number of classes on which the ConvEntion model is trained can significantly broaden its practical applicability in astronomical research. Currently, the model is configured to identify four classes - AGN, SNIa, Variable, and SN Other. This limited number of classes, although effective for specific studies, might restrict a comprehensive understanding of the cosmos. By training the model on a greater number of classes, it can be more inclusive of the diverse range of astronomical phenomena. With a more extensive classification spectrum, researchers could glean more nuanced insights into celestial events, making for a deeper and more detailed exploration of the universe.

However, increasing the number of classes isn't without challenges. It implies a need for more diverse and voluminous data for training, validation, and testing. Moreover, the model would need to maintain its existing robustness and accu-

racy while accommodating a broader array of classes. This task would involve refining the underlying algorithm and might necessitate more computational resources. Nevertheless, the potential insights gained from a more comprehensive classification framework could vastly outweigh these challenges, underscoring the potential of extending the ConvEntion model's classification spectrum.

Transforming ConvEntion into an online operation mode, instead of its current batch-processing mode, presents a compelling prospect for real-time astronomical object classification. With live classification, it is possible to detect and categorize space objects as they are observed, marking a substantial shift from the conventional approach of analyzing image sequences. This transformation would enable an unprecedented level of immediacy and responsiveness, making the ConvEntion model an even more effective tool for astronomers. Real-time alerts could provide valuable indications of unexpected or particularly noteworthy events, potentially accelerating discovery and research in the field of astronomy.

To implement this idea effectively, Bayesian change point detection (Adams & MacKay, 2007) could be employed to handle alert generation when classifying objects. Bayesian change point detection, a statistical methodology, is adept at identifying 'change points' or significant shifts in a data sequence, representing the onset of an event. In this context, a 'change point' would indicate the detection of an object of interest like a supernova or an asteroid. The integration of this statistical approach would allow the model to not just classify objects but also determine the precise moment when an object of interest appears, triggering an alert for astronomers to swiftly analyze the event.

This proposed modification to the ConvEntion model seamlessly complements its core strengths. As the ConvEntion model already proves effective in handling

missing observations and leveraging spatiotemporal features for enhanced classification, the addition of real-time detection and alerting capabilities could further refine its robustness and accuracy. The capacity to classify and react to observations in real-time could bring about an increase in accuracy and timeliness in astronomical research, providing more comprehensive and immediate insights to scientists.

For the Semiconformer framework, there is potential to transition to a fully self-supervised model if a more structured unlabeled dataset becomes available. Self-supervised learning has shown great promise in various fields of machine learning, and it could be particularly beneficial in the context of astronomical image time series classification, where labeled data can be scarce.

One of the major challenges in the field of astronomical image time series classification is the lack of a standard dataset that can be used by all scientists to compare and benchmark their models. The absence of such a dataset slows the integration of machine learning into this part of astronomy and does not motivate machine learning experts to dive into this domain. Therefore, a significant contribution to the field would be the creation of a comprehensive, high-quality, and publicly available dataset of astronomical image sequences. This would provide a common benchmark for researchers and could spur further advancements in the field.

In conclusion, while this dissertation has made significant contributions to the field of astronomical image time series classification, there is still much work to be done. The perspectives outlined above provide a roadmap for future research and development efforts. By continuing to innovate and push the boundaries of what is possible, we can further enhance our understanding of the universe and make even more exciting discoveries.

CHAPTER 7

Résumé en français: Classification des Séries Temporelles d'Images Astronomiques Utilisant l'Apprentissage Profond

7.1 RÉSUMÉ

7.1.1 Contexte et Motivation

La quête de compréhension de l'univers et de son évolution, incluant son expansion et la formation de grandes structures, est une préoccupation centrale de la cosmologie moderne. cette quête repose on partie sur l'identification et l'étude de divers corps célestes qui servent de reference dans l'univers distant. Parmi ces corp element, les supernovae de type Ia sont particulièrement intéressantes en raison de leur mécanisme d'explosion standard (produisant une énergie de flux presque identique pour chaque supernova), qui permet de déduire la distance lumineuse de la supernova. Cependant, la détection de ces objets est une tâche difficile en raison des comportements similaires exhibés par de nombreux autres corps célestes. Cela nécessite la mise en oeuvre de méthodes robustes d'apprentissage automatique.

L'analyse de ces corps célestes est principalement réalisée à l'aide d'informations photométriques, qui impliquent la construction d'une série temporelle, appelée "courbe de lumière", représentant les valeurs d'amplitude d'un objet dans une bande spécifique au fil du temps. Malgré de nombreuses études sur ces courbes de lumière, les résultats de classification restent insuffisants pour mener la science astrophysique. C'est là que le rôle de l'apprentissage automatique devient inestimable. Les techniques d'apprentissage automatique ont été utilisées dans divers aspects de la recherche en astrophysique, y compris la détection et la classifica-

tion d'objets célestes, l'examen de la formation et de l'évolution des galaxies, et la détermination de paramètres cosmologiques fondamentaux.

L'avènement du big data en astronomie a transformé le domaine, avec des observatoires du monde entier qui génèrent des téraoctets de données chaque nuit. Cela présente des opportunités uniques d'examiner le cosmos avec une précision inégalée. Cependant, extraire des informations significatives de ces volumes énormes de données est un défi primordial. Les méthodes statistiques conventionnelles se sont avérées inadéquates pour gérer la complexité et l'échelle des jeux de données astronomiques contemporains. En conséquence, les chercheurs se sont tournés vers les algorithmes d'apprentissage automatique, qui ont montré un succès exceptionnel dans le traitement de données à grande échelle et complexes.

L'objectif de cette thèse est d'étendre ce travail en utilisant directement des images de corps célestes à différents moments, c'est-à-dire des séquences d'images centrées sur l'objet d'intérêt. L'objectif est de construire un modèle basé sur l' apprentissage profond capable d'extraire des informations contextuelles et ainsi d'améliorer les performances par rapport aux résultats obtenus à partir de l'utilisation de courbes de lumière. C'est une tâche complexe en raison du bruit intrinsèque et de la grande plage dynamique des données astrophysiques, ainsi que du problème de discordance entre les statistiques des bases de données d'entraînement et de test. Le phénomène de décalage vers le rouge, qui modifie drastiquement l'image perçue, complique encore cette tâche.

Dans ce contexte, la motivation pour cette recherche est de développer et d' appliquer des techniques avancées d'apprentissage automatique pour améliorer la précision et l'efficacité de la classification des séries temporelles d'images astronomiques, débloquant ainsi pleinement le potentiel de l'ère du big data en as-

tronomie.

7.1.2 But et Défis

L'objectif de cette recherche se concentre sur la classification des séries temporelles d'images astronomiques, une tâche essentielle dans l'astronomie contemporaine car elle permet l'identification et l'étude de phénomènes dépendants du temps. Cette tâche est particulièrement difficile en raison des vastes et complexes jeux de données générés par les relevés astronomiques modernes. Les défis principaux dans ce domaine peuvent être largement catégorisés en défis liés aux données, défis méthodologiques et défis computationnels.

Les défis liés aux données découlent de la complexité inhérente et de la variabilité des phénomènes astronomiques. Les données utilisées dans ce domaine, qui incluent les séries temporelles d'images astronomiques, sont intrinsèquement très bruyantes et ont une grande plage dynamique. La base de données est également sujette au problème de discordance, où les statistiques des bases de données d'entraînement et de test sont différentes. Ce phénomène s'explique par le fait que les observations constituant la base de données d'entraînement sont principalement des objets astrophysiques "proches", ce qui implique que leur spectre est moins décalé vers le rouge que les objets distants. Ce phénomène de décalage vers le rouge modifie radicalement l'image perçue et, par conséquent, la forme et la nature des images dans la base de données de test.

Les défis méthodologiques découlent du besoin d'une préparation minutieuse des données, du défi de la gestion des données incomplètes ou irrégulières, et de la complexité et de l'intensité computationnelle de certains modèles de réseaux neuronaux. L'utilisation de séquences d'images synthétiques, comme proposée

par certains chercheurs, est une nouvelle approche pour contourner le problème de la rareté des jeux de données. Cependant, les données synthétiques doivent être suffisamment proches des données du monde réel pour un entraînement et des tests efficaces, ce qui peut être difficile en raison de la complexité inhérente et de la variabilité des phénomènes astronomiques.

Les défis computationnels sont associés au développement d'algorithmes avancés et d'outils computationnels pour traiter, analyser et interpréter efficacement les informations contenues dans ces immenses jeux de données. Ces algorithmes ont le potentiel de révéler des relations et des modèles auparavant inconnus dans les données, conduisant finalement à de nouvelles découvertes et à une compréhension plus profonde de l'univers. Cependant, les exigences computationnelles des réseaux neuronaux avancés et la nécessité de gérer les complexités de ces réseaux présentent des défis significatifs.

Pour répondre à ces défis, des efforts continus de recherche et de développement seront nécessaires et pourraient également nécessiter une plus grande collaboration et un partage des ressources au sein de la communauté de recherche astronomique. L'objectif ultime reste le développement de modèles de classification robustes, efficaces et précis capables de gérer les vastes et complexes jeux de données générés par les relevés astronomiques modernes.

7.1.3 Contributions

Cette thèse apporte plusieurs contributions significatives au domaine de la classification des séries temporelles d'images astronomiques. La contribution principale est le développement d'une nouvelle approche pour classer différents types d'objets spatiaux directement à l'aide d'images. Cette approche, nommée Con-

vEntion (CONVolutional attENTION), tire parti de la puissance des convolutions et des transformateurs pour traiter les séries temporelles d'images astronomiques. ConvEntion intègre des caractéristiques spatio-temporelles et peut être appliqué à divers types de jeux de données d'images avec n'importe quel nombre de bandes. L'implémentation de ConvEntion a conduit à des améliorations substantielles de la précision de classification, avec une augmentation de 13% par rapport aux approches de pointe qui utilisent des séries temporelles d'images et une augmentation de 12% par rapport aux approches qui utilisent des courbes de lumière.

Une contribution secondaire de cette thèse est la proposition d'un cadre d'apprentissage de représentation de séries temporelles d'images semi-supervisées de bout en bout. Ce cadre tire parti des approches auto-supervisées pour réduire l'effet de déséquilibre des classes avec l'augmentation de données, minimiser la variance intra-classe, et améliorer les résultats sur de petits jeux de données sans avoir besoin de données supplémentaires étiquetées. Le cadre proposé, nommé Semiconformer, a démontré des performances supérieures par rapport à un modèle conventionnel, atteignant une précision de 82.18% et un score F1 de 75.33%.

Ces contributions représentent des avancées significatives dans le domaine de la classification des séries temporelles d'images astronomiques. Elles améliorent non seulement la précision et l'efficacité des tâches de classification mais ouvrent également la voie à de futures recherches dans ce domaine. En démontrant le potentiel des techniques d'apprentissage profond pour gérer des données astronomiques complexes, cette thèse fournit une base pour l'exploration et le développement ultérieurs de modèles avancés d'apprentissage automatique pour l'analyse d'images astronomiques.

BIBLIOGRAPHY

Adams, R. P., & MacKay, D. J. C. (2007). Bayesian online changepoint detection.

Astier, P., Hage, P. E., Guy, J., Hardin, D., Betoule, M., Fabbro, S., Fourmanoit, N., Pain, R., & Regnault, N. (2013). Photometry of supernovae in an image series: methods and application to the SuperNova legacy survey (SNLS). *Astronomy & Astrophysics*, *557*, A55.

Badue, C., Guidolini, R., Carneiro, R. V., Azevedo, P., Cardoso, V. B., Forechi, A., Jesus, L., Berriel, R., Paixão, T., Mutz, F., Veronese, L., Oliveira-Santos, T., & Souza, A. F. D. (2019). Self-driving cars: A survey.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. https://arxiv.org/abs/1409.0473.

Bai, Z., Zhang, H., Yuan, H., Carlin, J. L., Li, G., Zhao, Y., & Cao, Z. (2017). Cosmic ray removal in fiber spectroscopic image. *arXiv preprint arXiv:1705.02084*.

Bairouk, A., CHAUMONT, M., FOUCHEZ, D., PASQUET, J., COMBY, F., & BAUTISTA, J. (2023). Convention: Astronomical image time series classification using convolutional attention. *Astronomy & Astrophysics Journal*.

Bao, H., Dong, L., Piao, S., & Wei, F. (2022). BEit: BERT pre-training of image transformers. https://openreview.net/forum?id=p-BhZSz59o4.

Bellm, E. C., Kulkarni, S. R., Graham, M. J., Dekany, R., Smith, R. M., Riddle, R., Masci, F. J., Helou, G., Prince, T. A., Adams, S. M., et al. (2019). The zwicky transient facility: System overview, performance, and first results. *Publications of the Astronomical Society of the Pacific*, *131*(995), 018002.

Bertin, E., & Arnouts, S. (1996). Sextractor: Software for source extraction. *Astronomy and Astrophysics Supplement Series*, *117*, 393–404.

Blanchard, A., Héloret, J.-Y., Ili, S., Lamine, B., & Tutusaus, I. (2023). λcdm is alive and well.

Boone, K. (2019). Avocado: Photometric classification of astronomical transients with gaussian process augmentation. *The Astronomical Journal*, *158*(6), 257.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers.

Carrasco-Davis, R., Cabrera-Vives, G., Förster, F., Estévez, P. A., Huijse, P., Protopapas, P., Reyes, I., Martínez-Palomera, J., & Donoso, C. (2019). Deep learning for image sequence classification of astronomical events. *Publications of the Astronomical Society of the Pacific*, *131*(1004), 108006.

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 6299–6308).

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321357.

Chen, S., Lui, M., Heigold, G., Dehghani, M., Arnab, A., & Schmid, C. (2019). Videobert: A joint model for video and language representation learning. *arXiv preprint arXiv:1904.01766*.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020a). A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning*, (pp. 1597–1607).

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020b). A simple framework for contrastive learning of visual representations.

Chen, Y., Fan, L., Zhang, P., Pu, S., & Wang, Y. (2021). Transvg: An end-to-end framework for visual grounding with transformers. *arXiv preprint arXiv:2103.16553*.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Choromanski, K. M., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J. Q., Mohiuddin, A., Kaiser, L., Belanger, D. B., Colwell, L. J., & Weller, A. (2021). Rethinking attention with performers. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Ua6zuk0WRH.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling.

Dablain, D., Krawczyk, B., & Chawla, N. V. (2021). Deepsmote: Fusing deep learning and smote for imbalanced data.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (pp. 4171–4186).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Doersch, C., Gupta, A., & Efros, A. A. (2015). Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*.

Donahue, J., Anne Hendricks, L., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021a). An image is worth 16x16 words: Transformers for image recognition at scale.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2021b). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*.

Efraimidis, P. S. (2015). Weighted random sampling over data streams.

Einstein, A. (1916). The foundation of the general theory of relativity. *Annalen der Physik*, *354*(7), 769–822.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Filippenko, A. V. (1997). Optical spectra of supernovae. *Annual Review of Astronomy and Astrophysics*, *35*(1), 309–355.

Fortunato, M., Blundell, C., & Vinyals, O. (2017). Bayesian recurrent neural networks. *ArXiv*, *abs/1704.02798*.

Freedman, W. L., Madore, B. F., Gibson, B. K., Ferrarese, L., Kelson, D. D., Sakai, S., Mould, J. R., Kennicutt Jr, R. C., Ford, H. C., Graham, J. A., et al. (2001). Final results from the hubble space telescope key project to measure the hubble constant. *The Astrophysical Journal*, *553*(1), 47–72.

Frieman, J. A., Bassett, B., Becker, A., Choi, C., Cinabro, D., DeJongh, F., Depoy, D. L., Dilday, B., Doi, M., Garnavich, P. M., Hogan, C. J., Holtzman, J., Im, M., Jha, S., Kessler, R., Konishi, K., Lampeitl, H., Marriner, J., Marshall, J. L., McGinnis, D., Miknaitis, G., Nichol, R. C., Prieto, J. L., Riess, A. G., Richmond, M. W., Romani, R., Sako, M., Schneider, D. P., Smith, M., Takanashi, N., Tokita, K., Heyden, K. v. d., Yasuda, N., Zheng, C., Adelman-McCarthy, J., Annis, J., Assef, R. J., Barentine, J., Bender, R., Blandford, R. D., Boroski, W. N., Bremer, M., Brewington, H., Collins, C. A., Crotts, A., Dembicky, J., Eastman, J., Edge, A., Edmondson, E., Elson, E., Eyler, M. E., Filippenko, A. V., Foley, R. J., Frank, S., Goobar, A., Gueth, T., Gunn, J. E., Harvanek, M., Hopp, U., Ihara, Y., Ivezi, ., Kahn, S., Kaplan, J., Kent, S., Ketzeback, W., Kleinman, S. J., Kollatschny, W., Kron, R. G., Krzesiski, J., Lamenti, D., Leloudas, G., Lin, H., Long, D. C., Lucey, J., Lupton, R. H., Malanushenko, E., Malanushenko, V., McMillan, R. J., Mendez, J., Morgan, C. W., Morokuma, T., Nitta, A., Ostman, L., Pan, K., Rockosi, C. M., Romer, A. K., Ruiz-Lapuente, P., Saurage, G., Schlesinger, K., Snedden, S. A., Sollerman, J., Stoughton, C., Stritzinger, M., SubbaRao, M., Tucker, D., Vaisanen, P., Watson, L. C., Watters, S., Wheeler, J. C., Yanny, B., & York, D. (2007). The sloan digital sky survey-ii supernova survey: Technical summary. *The Astronomical Journal*, *135*(1), 338347.

Frizzi, S. (2016). Convolutional neural network for video fire and smoke detection. *IECON 2016 - 42nd Annual Conference of the IEEE Industrial Electronics Society*.

Fukugita, M., Ichikawa, T., Gunn, J. E., Doi, M., Shimasaku, K., & Schneider, D. P. (1996). The Sloan Digital Sky Survey Photometric System. , *111*, 1748.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, *36*(4), 193–202.

Gabruseva, T., Zlobin, S., & Wang, P. (2019). Photometric light curves classification with machine learning.

Garnot, V. S. F., Landrieu, L., Giordano, S., & Chehata, N. (2019). Satellite image time series classification with pixel-set encoders and temporal self-attention.

Gehrels, N., & Razzaque, S. (2013). Gamma-ray bursts in the swift-fermi era. *Frontiers of Physics*, *8*(6), 661–676.

Gers, F., & Schmidhuber, J. (2000). Recurrent nets that time and count. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, vol. 3, (pp. 189–194 vol.3).

Gers, F. A., & Schmidhuber, J. (2001). Lstm recurrent networks learn simple context free and context sensitive languages. *IEEE Transactions on Neural Networks*.

Gill, K., Asefa, T., Kaheil, Y., & McKee, M. (2007). Effect of missing data on performance of learning algorithms for hydrologic predictions: Implications to an imputation technique.

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

Graham, M. J., Kulkarni, S. R., Bellm, E. C., Kasliwal, M. M., Prince, T. A., Masci, M. C., Dekany, R. G., Smith, R. M., Riddle, R., & Holoien, T. W.-S. (2019). The zwicky transient facility: Science objectives. *Publications of the Astronomical Society of the Pacific*, *131*(1001), 078001.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., & Valko, M. (2020). Bootstrap your own latent: A new approach to self-supervised learning.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., et al. (2018). Recent advances in convolutional neural networks. In *Proceedings of the IEEE*, vol. 106, (pp. 779–797). IEEE.

Guth, A. H. (1981). Inflationary universe: A possible solution to the horizon and flatness problems. *Physical Review D*, *23*(2), 347.

Gómez, C., Neira, M., Hernández Hoyos, M., Arbeláez, P., & Forero-Romero, J. E. (2020). Classifying image sequences of astronomical transients with deep neural networks. *Monthly Notices of the Royal Astronomical Society*, *499*(3), 31303138.

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2019). Momentum contrast for unsupervised visual representation learning.

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 770–778).

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. Wiley.

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

Holtzman, J. A., Marriner, J., Kessler, R., Sako, M., Dilday, B., Frieman, J. A., Schneider, D. P., Bassett, B., Becker, A., Cinabro, D., DeJongh, F., Depoy, D. L., Doi, M., Garnavich, P. M., Hogan, C. J., Jha, S., Konishi, K., Lampeitl, H., Marshall, J. L., McGinnis, D., Miknaitis, G., Nichol, R. C., Prieto, J. L., Riess, A. G., Richmond, M. W., Romani, R., Smith, M., Takanashi, N., Tokita, K., van der Heyden, K., Yasuda, N., & Zheng, C. (2008). The sloan digital sky survey-ii: Photometry and supernova ia light curves from the 2005 data. *The Astronomical Journal*, *136*(6), 23062320.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*(8), 2554–2558.

Hossain, M. S., Betts, J. M., & Paplinski, A. P. (2021). Dual focal loss to address class imbalance in semantic segmentation. *Neurocomputing*, *462*, 69–87.

Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. (2017). Densely connected convolutional networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 4700–4708).

Hubble, E. (1929). A relation between distance and radial velocity among extragalactic nebulae. *Proceedings of the National Academy of Sciences*, *15*(3), 168–173.

Hubble, E., & Tolman, R. C. (1935). Two methods of investigating the nature of the nebular redshift. *The Astrophysical Journal*, *82*, 302.

Hughes, D. W. (2006). The introduction of absolute magnitude (1902 - 1922). *Journal of Astronomical History and Heritage*, *9*(2), 173–179.

Ivezić, Ž., Kahn, S. M., Tyson, J., Abel, B., Acosta, E., Allsman, R., Alonso, D., Alsayyad, Y., Anderson, S. F., Andrew, J., Angel, J. R. P., Angeli, G. Z., Ansari, R., Antilogus, P., Araujo, C., Armstrong, R., Arndt, K. T., Astier, P., Aubourg, É., Auza, N., Axelrod, T. S., Bard, D. J., Barr, J. D., Barrau, A., Bartlett, J. G., Bauer, A. E., Bauman, B. J., Baumont, S., Bechtol, E., Bechtol, K., Becker, A. C., Becla, J., Beldica, C., Bellavia, S., Bianco, F. B., Biswas, R., Blanc, G., Blazek, J., Blandford, R. D., Bloom, J. S., Bogart, J., Bond, T. W., Booth, M. T., Borgland, A. W., Borne, K., Bosch, J. F., Boutigny, D., Brackett, C. A., Bradshaw, A., Brandt, W. N., Brown, M. E., Bullock, J. S., Burchat, P., Burke, D. L., Cagnoli, G., Calabrese, D., Callahan, S., Callen, A. L., Carlin, J. L., Carlson, E. L., Chandrasekharan, S., Charles-Emerson, G., Chesley, S., Cheu, E. C., Chiang, H.-F., Chiang, J., Chirino, C., Chow, D., Ciardi, D. R., Claver, C. F., Cohen-Tanugi, J., Cockrum, J. J., Coles, R., Connolly, A. J., Cook, K. H., Cooray, A., Covey, K. R., Cribbs, C., Cui, W., Cutri, R., Daly, P. N., Daniel, S. F., Daruich, F., Daubard, G., Daues, G., Dawson, W., Delgado, F., Dellapenna, A., Peyster, R. D., de Val-Borro, M., Digel, S. W.,

Doherty, P., Dubois, R., Dubois-Felsmann, G. P., Durech, J., Economou, F., Eifler, T., Eracleous, M., Emmons, B. L., Fausti Neto, A., Ferguson, H., Figueroa, E., Fisher-Levine, M., Focke, W., Foss, M. D., Frank, J., Freemon, M. D., Gangler, E., Gawiser, E., Geary, J. C., Gee, P., Geha, M., Gessner, C. J., Gibson, R. R., Gilmore, D., Glanzman, T., Glick, W., Goldina, T., Goldstein, D. A., Goodenow, I., Graham, M. L., Gressler, W. J., Gris, P., Guy, L. P., Guyonnet, A., Haller, G., Harris, R., Hascall, P. A., Haupt, J., Hernandez, F., Herrmann, S., Hileman, E., Hoblitt, J., Hodgson, J. A., Hogan, C., Howard, J. D., Huang, D., Huffer, M. E., Ingraham, P., Innes, W. R., Jacoby, S. H., Jain, B., Jammes, F., Jee, J., Jenness, T., Jernigan, G., Jevremović, D., Johns, K., Johnson, A. S., Johnson, M. W., Jones, R., Juramy-Gilles, C., Jurić, M., Kalirai, J. S., Kallivayalil, N. J., Kalmbach, B., Kantor, J. P., Karst, P., Kasliwal, M. M., Kelly, H., Kessler, R., Kinnison, V., Kirkby, D., Knox, L., Kotov, I. V., Krabbendam, V. L., Krughoff, K., Kubánek, P., Kuczewski, J., Kulkarni, S., Ku, J., Kurita, N. R., Lage, C. S., Lambert, R., Lange, T., Langton, J., Le Guillou, L., Levine, D., Liang, M., Lim, K.-T., Lintott, C. J., Long, K. E., Lopez, M., Lotz, P. J., Lupton, R. H., Lust, N. B., Macarthur, L. A., Mahabal, A., Mandelbaum, R., Markiewicz, T. W., Marsh, D. S., Marshall, P. J., Marshall, S., May, M., Mckercher, R., Mcqueen, M., Meyers, J., Migliore, M., Miller, M., Mills, D. J., Miraval, C., Moeyens, J., Moolekamp, F. E., Monet, D. G., Moniez, M., Monkewitz, S., Montgomery, C., Morrison, C. B., Mueller, F., Muller, G. P., Arancibia, F. M., Neill, D. R., Newbry, S. P., Nief, J.-Y., Nomerotski, A., Nordby, M., O'Connor, P., Oliver, J., Olivier, S. S., Olsen, K., O'Mullane, W., Ortiz, S., Osier, S., Owen, R. E., Pain, R., Palecek, P. E., Parejko, J. K., Parsons, J. B., Pease, N. M., Peterson, J., Peterson, J. R., Petravick, D. L., Petrick, M. L., Petry, C. E., Pierfederici, F., Pietrowicz, S., Pike, R., Pinto, P. A., Plante, R., Plate, S., Plutchak, J. P., Price, P. A., Prouza, M., Radeka, V., Rajagopal, J., Rasmussen, A. P., Regnault, N., Reil, K. A., Reiss, D. J., Reuter, M. A., Ridgway, S. T., Riot, V. J., Ritz, S., Robinson, S., Roby, W., Roodman, A., Rosing, W., Roucelle, C., Rumore, M. R., Russo, S., Saha, A., Sassolas, B., Schalk, T. L., Schellart, P., Schindler, R. H., Schmidt, S., Schneider, D. P., Schneider, M. D., Schoening, W., Schumacher, G., Schwamb, M. E., Sebag, J., Selvy, B., Sembroski, G. H., Seppala, L. G., Serio, A., Serrano, E., Shaw, R. A., Shipsey, I., Sick, J., Silvestri, N., Slater, C. T., Smith, J., Smith, R., Sobhani, S., Soldahl, C., Storrie-Lombardi, L., Stover, E., Strauss, M. A., Street, R. A., Stubbs, C. W., Sullivan, I. S., Sweeney, D., Swinbank, J. D., Szalay, A., Takacs, P., Tether, S. A., Thaler, J. J., Thayer, J. G., Thomas, S., Thornton, A. J., Thukral, V., Tice, J., Trilling, D. E., Turri, M., van Berg, R., Vanden Berk, D., Vetter, K., Virieux, F., Vucina, T., Wahl, W., Walkowicz, L., Walsh, B., Walter, C. W., Wang, D. L., Wang, S.-Y., Warner, M., Wiecha, O., Willman, B., Winters, S. E., Wittman, D., Wolff, S. C., Wood-Vasey, W., Wu, X., Xin, B., Yoachim, P., & Zhan, H. (2019). LSST: from Science Drivers to Reference Design and Anticipated Data Products. *Astrophys.J., 873*(2), 111.

Ivezic, Z., Tyson, J. A., Lupton, R. H., Juric, M., & Slater, C. T. (2019). Lsst: From science drivers to reference design and anticipated data products. *The Astrophysical Journal*, *873*(2), 111.

Ji, S., Xu, W., Yang, M., & Yu, K. (2013). 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 221–231.

Jing, L., & Tian, Y. (2020). Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Keller, P. M., Breedt, E., Hodgkin, S., Belokurov, V., Wild, J., García-Soriano, I., & Wise, J. L. (2021). Eclipsing white dwarf binaries in gaia and the zwicky transient facility. *arXiv preprint arXiv:2105.14028*.

Kingma, D. P., & Ba, J. (2017). Adam: A method for stochastic optimization.

Kirshner, R. P. (1999). Supernovae, an accelerating universe and the cosmological constant. *Proceedings of the National Academy of Sciences*, *96*(8), 4224–4227.

Krizhevsky, A., Sutskever, I., & Hinton, G. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90.

Lazio, J. (2009). The square kilometre array.

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., & Jackel, L. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, *1*(4), 541–551.

Levesque, M., & Lelievre, M. (2010). Evaluation of the accuracy of the dark frame subtraction method in ccd image processing. *Defence Research and Development Canada*.

Li, C., Yang, J., Zhang, P., Gao, M., Xiao, B., Dai, X., Yuan, L., & Gao, J. (2021). Efficient self-supervised vision transformers for representation learning.

Li, L. (2022). Dance art scene classification based on convolutional neural networks. *Scientific Programming*.

Li, W., Filippenko, A. V., Chornock, R., Berger, E., Berlind, P., Calkins, M. L., Challis, P., Fassnacht, C., Jha, S., Kirshner, R. P., Matheson, T., Sargent, W. L. W., Simcoe, R. A., Smith, G. H., & Squires, G. (2003). SN 2002cx: The most peculiar known type ia supernova. *Publications of the Astronomical Society of the Pacific*, *115*(806), 453–473.

Lin, J., Gan, C., & Han, S. (2019). Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (pp. 7083–7093).

Lin, T.-Y., Goyal, P., Girshick, R. B., He, K., & Dollár, P. (2017). Focal loss for dense object detection. *CoRR*, *abs/1708.02002*.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021a). Swin transformer: Hierarchical vision transformer using shifted windows.

Liu, Z., Luo, S., Li, W., Lu, J., Wu, Y., Sun, S., Li, C., & Yang, L. (2021b). Convtransformer: A convolutional transformer network for video frame synthesis.

Lorimer, D. R., Bailes, M., McLaughlin, M. A., Narkevic, D. J., & Crawford, F. (2007). A bright millisecond radio burst of extragalactic origin. *Science*, *318*(5851), 777–780.

Loshchilov, I., & Hutter, F. (2016). Sgdr: Stochastic gradient descent with warm restarts.

Loshchilov, I., & Hutter, F. (2018). Fixing weight decay regularization in adam. https://openreview.net/forum?id=rk6qdGgCZ.

Masci, F. J., Shupe, D. L., Laher, R. R., Carry, B., Helou, G., Bellm, E. C., Graham, M. J., Dekany, R. G., Smith, R. M., & Riddle, R. (2019). The zwicky transient facility: Data processing, products, and archive. *Publications of the Astronomical Society of the Pacific*, *131*(995), 018003.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, *5*(4), 115–133.

Minsky, M., & Papert, S. (1969). *Perceptrons: An introduction to computational geometry*. MIT press.

Mróz, P., Otarola, A., Prince, T. A., Dekany, R., Duev, D. A., Graham, M. J., Groom, S. L., Masci, F. J., & Medford, M. S. (2022). Impact of the spacex starlink satellites on the zwicky transient facility survey observations. *The Astrophysical Journal Letters*, *924*(2), L30.

Mészáros, P. (2006). Gamma-ray bursts. *Reports on Progress in Physics*, *69*(8), 2259–2322.

Möller, A., & de Boissière, T. (2020). Supernnova: an open-source framework for bayesian, neural network-based supernova classification. *Monthly Notices of the Royal Astronomical Society*, *491*(3), 42774293.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, *2*(1), 1–21.

Ng, J. Y.-H., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., & Toderici, G. (2015). Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 4694–4702).

Novikoff, A. B. (1962). On convergence proofs on perceptrons. In *Proceedings of the symposium on the mathematical theory of automata*, vol. 12, (pp. 615–622).

O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *Arxiv preprint arXiv:1511.08458*.

Osterbrock, D. E. (2001). Who really coined the word supernova? who first predicted neutron stars? *Bulletin of the American Astronomical Society*, *33*, 1501.

Pascanu, R., Mikolov, T., & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. *International Conference on Machine Learning*, (pp. 1310–1318).

Pasquet, J., Pasquet, J., Chaumont, M., & Fouchez, D. (2019). Pelican: deep architecture for the light curve analysis. *Astronomy & Astrophysics*, *627*, A21.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library.

Penzias, A. A., & Wilson, R. W. (1965). A measurement of excess antenna temperature at 4080-mc/s. *The Astrophysical Journal*, *142*, 419–421.

Perlmutter, S., Aldering, G., Goldhaber, G., Knop, R. A., Nugent, P., Castro, P. G., Deustua, S., Fabbro, S., Goobar, A., & Groom, D. E. (1999). Measurements of omega and lambda from 42 high-redshift supernovae. *The Astrophysical Journal*, *517*(2), 565–586.

PLAsTiCC-team, au2, T. A. J., Bahmanyar, A., Biswas, R., Dai, M., Galbany, L., Hloek, R., Ishida, E. E. O., Jha, S. W., Jones, D. O., Kessler, R., Lochner, M., Mahabal, A. A., Malz, A. I., Mandel, K. S., Martínez-Galarza, J. R., McEwen, J. D., Muthukrishna, D., Narayan, G., Peiris, H., Peters, C. M., Ponder, K., Setzer, C. N., Collaboration, T. L. D. E. S., Transients, T. L., & Collaboration, V. S. S. (2018). The photometric lsst astronomical time-series classification challenge (plasticc): Data set.

Prakash, B. S., Sanjeev, K. V., & K., R. P. (2019). A survey on recurrent neural network architectures for sequential learning. In *SocProS*, (pp. 51–60).

Qu, H., Sako, M., Möller, A., & Doux, C. (2021). Scone: Supernova classification with a convolutional neural network. *The Astronomical Journal*, *162*(2), 67.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*, 1–67.

Riess, A. G., Filippenko, A. V., Challis, P., Clocchiatti, A., Diercks, A., Garnavich, P. M., Gilliland, R. L., Hogan, C. J., Jha, S., Kirshner, R. P., Leibundgut, B., Phillips, M. M., Reiss, D., Schmidt, B. P., Schommer, R. A., Smith, R. C., Spyromilio, J., Stubbs, C., Suntzeff, N. B., & Tonry, J. (1998). Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant. , *116*, 1009–1038.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408.

Sako, M., Bassett, B., Becker, A. C., Brown, P. J., Campbell, H., Wolf, R., Cinabro, D., D'Andrea, C. B., Dawson, K. S., DeJongh, F., Depoy, D. L., Dilday, B., Doi, M., Filippenko, A. V., Fischer, J. A., Foley, R. J., Frieman, J. A., Galbany, L., Garnavich, P. M., Goobar, A., Gupta, R. R., Hill, G. J., Hayden, B. T., Hlozek, R., Holtzman, J. A., Hopp, U., Jha, S. W., Kessler, R., Kollatschny, W., Leloudas, G., Marriner, J., Marshall, J. L., Miquel, R., Morokuma, T., Mosher, J., Nichol, R. C., Nordin, J., Olmstead, M. D., Östman, L., Prieto, J. L., Richmond, M., Romani, R. W., Sollerman, J., Stritzinger, M., Schneider, D. P., Smith, M., Wheeler, J. C., Yasuda, N., & Zheng, C. (2014). The data release of the sloan digital sky survey-II supernova survey. *Publications of the Astronomical Society of the Pacific*, *130*(988), 064002.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*.

Seager, S., & Deming, D. (2010). Exoplanet atmospheres. *Annual Review of Astronomy and Astrophysics*, *48*, 631–672.

Sharir, G., Noy, A., & Zelnik-Manor, L. (2021). An image is worth 16x16 words, what is a video worth?

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., kin Wong, W., & chun Woo, W. (2015). Convolutional lstm network: A machine learning approach for precipitation nowcasting.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Springel, V., White, S. D. M., Jenkins, A., Frenk, C. S., Yoshida, N., Gao, L., Navarro, J., Thacker, R., Croton, D., Helly, J., et al. (2005). Simulations of the formation, evolution and clustering of galaxies and quasars. *Nature*, *435*(7042), 629–636.

Stephenson, F. R., & Green, D. A. (2003). *The Historical Supernovae*. Oxford University Press.

Sundermeyer, M., Schlüter, R., & Ney, H. (2012). Lstm neural networks for language modeling. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results.

Touvron, H., Caron, M., Joulin, A., & Alayrac, J.-B. (2021). Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2014). Learning spatiotemporal features with 3d convolutional networks. https://arxiv.org/abs/1412.0767.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, (pp. 4489–4497).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Vrskova, R., Hudec, R., Kamencay, P., & Sykora, P. (2022). Human activity classification using the 3dcnn architecture. *Applied Sciences*, *12*(2), 931.

Wang, S., Li, B. Z., Khabsa, M., Fang, H., & Ma, H. (2020). Linformer: Self-attention with linear complexity.

Wang, Y., Gu, Y., & Li, X. (2022). A novel low rank smooth flat-field correction algorithm for hyperspectral microscopy imaging. *IEEE Transactions on Medical Imaging*, *41*(12), 3862–3872.

Weinstein, G. (2016). Einstein's discovery of gravitational waves 1916-1918.

Woosley, S. E., Heger, A., & Weaver, T. A. (2002). The evolution and explosion of massive stars. *Reviews of Modern Physics*, *74*(4), 1015–1071.

Xu, W., Xu, Y., Chang, T., & Tu, Z. (2021). Co-scale conv-attentional image transformers.

Yang, Y., Krompass, D., & Tresp, V. (2017). Tensor-train recurrent neural networks for video classification.

Yuan, Y., & Lin, L. (2021a). Self-supervised pretraining of transformers for satellite image time series classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *14*, 474–487.

Yuan, Y., & Lin, L. (2021b). Self-supervised pretraining of transformers for satellite image time series classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *14*, 474–487.

Yuan, Y., & Lin, L. (2021c). Self-supervised pretraining of transformers for satellite image time series classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *14*, 474–487.

Zhang, Y., & Zhao, Y. (2015). Astronomy in the big data era. *Data Science Journal*, *14*, 11.

Zhou, G.-B., Wu, J., Zhang, C.-L., & Zhou, Z.-H. (2016). Minimal gated unit for recurrent neural networks. *International Journal of Automation and Computing*, *13*(3), 226–234.

Zhou, H., Zhang, L., Cheng, P.-T., Xiong, Y., Chang, S., & Song, W. (2021). Vivit: A video vision transformer. *arXiv preprint arXiv:2103.15691*.

Zhu, L. (2019). Faster recurrent networks for video classification. *ArXiv*.