FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

ASSIGNMENT OF MASTER’S THESIS

Title:	Neural Networks Based Domain Adaptation in Spectroscopic Sky Surveys
Student:	Bc. Ondřej Podsztavek
Supervisor:	RNDr. Petr Škoda, CSc.
Study Programme:	Informatics
Study Branch:	Knowledge Engineering
Department:	Department of Applied Mathematics
Validity:	Until the end of winter semester 2020/21

Instructions

The goal of this thesis is the analysis of the impact of domain adaptation in astronomical archives with a focus on neural networks that would allow using labeled data from one ground-based telescope or space mission archive to discover knowledge in another archive. Current astronomy has been the primary customer of scalable Big data handling and analysis requirements due to its petabyte-scale archives, where an advanced machine learning is an indispensable part of workflows leading to new discoveries.

1. Survey the current state of domain adaptation using neural network models in machine learning and with a focus on astronomical applications.
2. Select a suitable dataset of astronomical spectra for experiments.
3. Investigate the structure of data space in selected datasets.
4. Apply domain adaptation to the selected data.
5. Prepare the visualization of results.
6. Discuss the precision performance and scalability of various solutions and suggest future improvements.

References

Will be provided by the supervisor.

Ing. Karel Klouda, Ph.D.
Head of Department

doc. RNDr. Ing. Marcel Jiřina, Ph.D.
Dean

Prague August 15, 2019

FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

Master's thesis

Neural Networks Based Domain Adaptation in Spectroscopic Sky Surveys

Bc. Ondřej Podsztavek

Department of Applied Mathematics
Supervisor: RNDr. Petr Škoda, CSc.

January 9, 2020

Acknowledgements

I would like to thank Dr. Petr Škoda for his enthusiasm, domain expertise and encouragemenent throughout my research. This project would have been impossible without the support of Professor Tvrdík. Thanks to my girlfriend Kateřina. You supported me greatly. Thanks to Petr and Vojtěch for their friendship in the lab.

Furthermore, I would like to acknowledge the Sloan Digital Sky Survey and the Large Sky Area Multi-Object Fiber Spectroscopic Telescope for providing their data.

Funding for the Sloan Digital Sky Survey IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions. SDSS-IV acknowledges support and resources from the Center for High-Performance Computing at the University of Utah. The SDSS web site is `www.sdss.org`.

Guoshoujing Telescope (the Large Sky Area Multi-Object Fiber Spectroscopic Telescope) is a National Major Scientific Project built by the Chinese Academy of Sciences. Funding for the project has been provided by the National Development and Reform Commission. LAMOST is operated and managed by the National Astronomical Observatories, Chinese Academy of Sciences.

Declaration

I hereby declare that the presented thesis is my own work and that I have cited all sources of information in accordance with the Guideline for adhering to ethical principles when elaborating an academic final thesis.

I acknowledge that my thesis is subject to the rights and obligations stipulated by the Act No. 121/2000 Coll., the Copyright Act, as amended. In accordance with Article 46(6) of the Act, I hereby grant a nonexclusive authorization (license) to utilize this thesis, including any and all computer programs incorporated therein or attached thereto and all corresponding documentation (hereinafter collectively referred to as the “Work”), to any and all persons that wish to utilize the Work. Such persons are entitled to use the Work for non-profit purposes only, in any way that does not detract from its value. This authorization is not limited in terms of time, location and quantity.

In Prague on January 9, 2020 …………………

Czech Technical University in Prague
Faculty of Information Technology

Citation of this thesis

Podsztavek, Ondřej. *Neural Networks Based Domain Adaptation in Spectroscopic Sky Surveys.* Master's thesis. Czech Technical University in Prague, Faculty of Information Technology, 2020.

Abstrakt

Představujeme analýzu dopadu doménové adaptace založené na neuronových sítích v astronomické spektroskopii. Doménové adaptace řeší problém použití dříve získaných znalostí na nová data. Analýzu ukazujeme na problému identifikace kvasarů v přehlídce Large Sky Area Multi-Object Fiber Spectroscopic Telescope pomocí anotovaných dat z přehlídky Sloan Digital Sky Survey. Pro experimenty jsme vybrali čtyři modely založené na neuronových sítích pro doménovou adaptaci: Deep Domain Confusion, Deep Correlation Alignment, Domain-Adversarial Network and Deep Reconstruction-Classification Network. Výsledky experimentů ukázaly, že tyto modely nejsou schopné vylepšit klasifikační přesnost v porovnání s konvoluční neuronovou sítí, která doménovou adaptaci nebere v potaz. S využitím redukce dimensionality, statistik zmíněných metod a chyb v klasifikaci ukazujeme, že zvolené metody doménové adaptace nejsou dostatečně robustní, abychom je mohli aplikovat na komplexní a nevyčištěná astronomická data.

Klíčová slova doménová adaptace, neuronové sítě, hluboké učení, astronomická spektroskopie, astronomie

Abstract

We present an analysis of the impact of neural-based domain adaptation in astronomical spectroscopy. Domain adaptation addresses the problem of applying prior knowledge to a new data of interest. Therefore, we selected a problem of quasar identification in the Large Sky Area Multi-Object Fiber Spectroscopic Telescope survey using labelled data from the Sloan Digital Sky Survey. We choose to experiment with four neural models for domain adaptation: Deep Domain Confusion, Deep Correlation Alignment, Domain-Adversarial Network and Deep Reconstruction-Classification Network. However, our experiments reveal that these model cannot improve classification performance in comparison to a convolutional neural network that does not consider domain adaptation. Using dimensionality reduction, statistics of the selected methods and misclassifications, we show that the domain adaptation methods are not robust enough to be applied to the complex and dirty astronomical data.

Keywords domain adaptation, neural networks, deep learning, astronomical spectroscopy, astronomy

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this thesis, we ask the question of what is the impact of domain adaptation based on neural networks in astronomical spectroscopy. Having two different spectroscopic sky surveys, are we able to extract knowledge from one and apply it to another with a neural network. To be concrete, we will experiment with the identification of quasars in the Large Sky Area Multi-Object Fiber Spectroscopic Telescope survey with quasars identified in the Sloan Digital Sky Survey.

Previous research in machine learning has shown that domain adaptation can overcome the problem of analysing data from different distributions in one experiment while having a sufficient amount of data. Nowadays, astronomy is facing an avalanche of data, and exponential growth of data transforms science in general. The vast data sources provide immense potential for discoveries. However, we need to adapt existing methods or develop new sophisticated automatic approaches for data analysis. Therefore, we see potential in domain adaptation by exploiting knowledge extracted in previous problems.

To explore the potential of domain adaptation based on neural networks, we set the goal of this thesis to be the analysis of the impact of domain adaptation using data from the Sloan Digital Sky Survey for identification of quasars in the Large Sky Area Multi-Object Fiber Spectroscopic Telescope survey.

In the second chapter 2, we firstly introduce astronomical spectroscopy, describe the aforementioned spectroscopic surveys, and we demonstrate why they are suitable for our problem. In chapter 3, we provide a summary of the theory of domain adaptation in the context of the machine and transfer learning, survey domain adaptation based on neural network and give an overview of applications of domain adaptation in astronomy. In the chapter 4 with experiments, we analyse the impact of domain adaptation on four models based on neural networks. We suggest future improvements and conclude this thesis in the last chapter 5.

Chapter 2

Spectroscopic Sky Surveys

We start with a brief introduction to astronomical spectroscopy to understand the data used in this thesis. Then, we describe quasars because they are the object we would like to identify in the Large Sky Area Multi-Object Spectroscopic Telescope survey using the data and labels from the Sloan Digital Sky Survey. Therefore, we introduce the two large spectroscopic surveys in the following section. We end this chapter with a comparison of the two surveys to show that they are different, thus suitable for the problem of domain adaptation.

2.1 Astronomical Spectroscopy

Almost all we know about the universe outside our solar system is based on the analysis of light. For example, we observe its flux or time variations. [1]. Light is *electromagnetic* (EM) radiation. Spectroscopy started when Issac Newton experimented with light decomposition through a prism in 1666. Next, Thomas Young has shown that light behaves like a wave. However, EM radiation exhibits also particle nature.

The particles of EM radiation are photons. Photons have no mass but transport energy and have momentum. Every photon has an associated frequency ν of the corresponding EM wave giving it energy by the Planck–Einstein relation:

$$E = h\nu, \tag{2.1}$$

where h is the Planck constant and ν is the frequency of the photon. Therefore, a higher frequency means higher energy. All photons in vacuum move in the speed of light c. The frequency ν is related to the wavelength λ:

$$\nu = \frac{c}{\lambda}, \tag{2.2}$$

Radiation type	Wavelength (m)
γ ray	10^{-12}
X-ray	10^{-10}
Ultraviolet	10^{-8}
Visible	0.5×10^{-6}
Infrared	10^{-5}
Microwave	10^{-2}
Radio	10^{3}

Table 2.1: Parts of electromagnetic spectrum

where c is the speed of light. We see that the energy of a photon is inversely proportional to wavelength λ by combining equations 2.1 and 2.2 gives [2]:

$$E = h\frac{c}{\lambda}. \tag{2.3}$$

Secondly, EM radiation has the wave nature. EM wave also propagates in the speed of light c and transfers energy. Again the higher the frequency, the more energy it carries. [3]

EM wave can be decomposed (by a prism or a diffraction grating) as a function of its wavelength λ. The complete decomposed spectrum of EM radiation is called the EM spectrum. The visible light is only a tiny part of the complete spectrum of electromagnetic radiation. The complete spectrum consists of γ rays, X-rays, ultraviolet, visible light, infrared radiation, microwaves, radio waves (see Table 2.1).

Firstly, EM radiation is produced by either heating up matter or by exciting atoms. The blackbody is a physical model of spectral radiation $B(\lambda, T)$. Max Planck derived the spectral distribution of a black body. Heating transforms into emission of EM radiation at all wavelengths with an energy distribution as a function of a wavelength which only depends on the temperature and is described by Planck's law:

$$B(\lambda, T) = \frac{2hc^2}{\lambda^5}\frac{1}{e^{\frac{hc}{\lambda k_\mathrm{B} T}} - 1}, \tag{2.4}$$

where k_B is the Boltzmann constant and h is the Planck constant. This phenomenon does not depend on the composition of the body but only on its temperature. The Wien's displacement law gives the wavelength of maximum intensity λ_max:

$$\lambda_\mathrm{max} = \frac{b}{T}, \tag{2.5}$$

where b is the Wien's displacement constant. Accordingly, we can derive the temperature T of an object with known wavelength of maximum intesity $\lambda_{\max}$:

$$T = \frac{b}{\lambda_{\max}}, \tag{2.6}$$

where b is the Wien's displacement constant. [3]

We consider stars to be black bodies because they are objects hotter than its environment and emit electromagnetic radiation. Therefore, all EM radiation of a star is determined only by its temperature. The other way around, with the Wien's displacement law, we can estimate the absolute temperature T of a star. [2]

Secondly, EM radiation can be emitted by exciting atoms. Therefore, EM radiation carries information about stars and planets made of matter across the universe. The energy carried by EM radiation interacts with matter in the following ways:

- *emission* occurs when EM radiation propagates through a gas because the atoms of the gas might excite and emit EM radiation;
- *absorption* happens when a gas absorbs wavelengths of the EM radiation.

Fraunhofer was also one of the first to observe dark lines in the solar spectrum and David Brewster postulated that the dark lines correspond to absorption from gas in the way of light travelling towards us. Robert Wilhelm Bunsen and Gustav Kirchhoff showed that each chemical element has its own set of spectral lines and Kirchhoff established the three famous laws describing the three types of spectra:

- the spectrum of a conventional light bulb is a continuous rainbow (called *continuous spectrum*);
- if a cloud of gas lies between a detector and a bulb, the cloud can absorb specific wavelength making what an *absorption-line spectrum*;
- if a cloud emits light itself, its spectrum is called an *emission-line spectrum.*

Real astronomical spectra are usually a combination of these types. Therefore, spectroscopy can be used for chemical analysis of the matter. Then, at the beginning of the twentieth century, the invention of quantum mechanics help us to understand the origin of spectral lines.

We model matter as made of atoms. An atom has a specific number of electrons which are place in particular orbits of the atom. Electrons in an orbit have a specific energy level. We know from quantum mechanics that an electron can change energy level by an exchange of energy in the form of a photon.

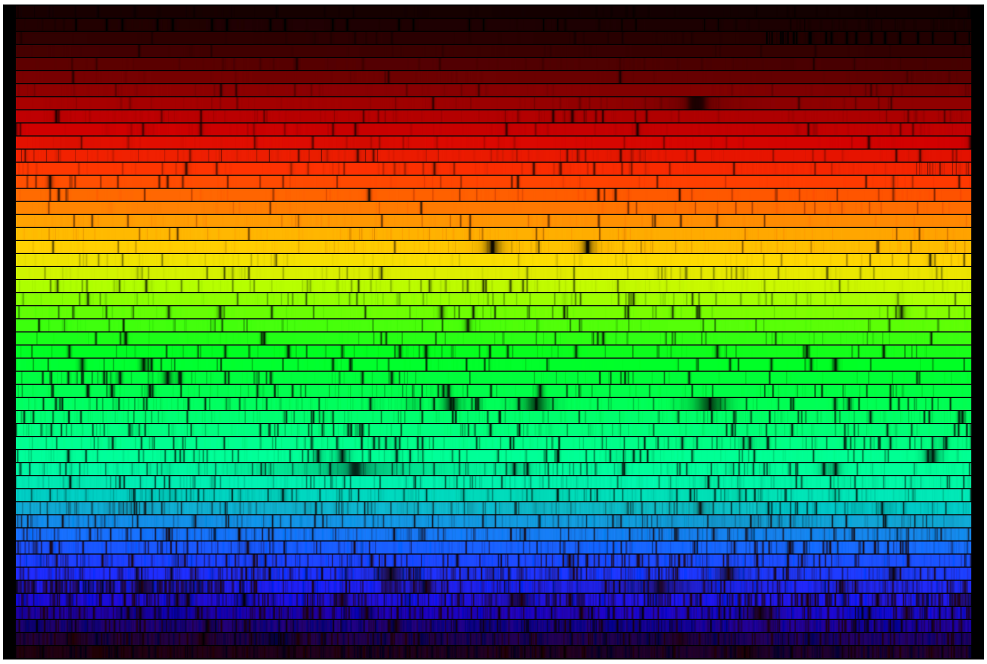


Figure 2.1: The two-dimensional spectrum of the Sun by NASA.

The energy transfer is not continuous. The energy has to correspond to the difference in the energy levels precisely. Therefore, the change produces EM radiation with an energy E in the wavelength λ according to Planck–Einstein relation in Equation 2.3. Therefore, the specific set of energy levels of an atom determines if photons are either absorbed or emitted by it, which is a direct consequence for spectroscopy. [3]

The fact that each atom, ion or molecule possesses a unique set of energy levels causes emission and absorption lines at specific wavelengths in spectra. Spectral lines correspond to the wavelengths of light absorbed by chemicals on the surface of the star. Therefore, positions of emission and absorption lines can tell us objects composition. We display spectra as bands of light that is a projection of light that passes through a prism on a wall (see Figure 2.1) called *two-dimensional* spectra. [3]

More reasonable is to display spectra as graphs of intensities of the light as the vertical axis and wavelengths as the horizontal axis: called a *one-dimensional spectral profile* (see Figure 2.2). [3, 4] This representation of an astronomical spectrum can be seen as a one-dimensional image.

Joseph von Fraunhofer was the first to observe spectra of stars by using a spectroscope in combination with a telescope. Nowadays, new technologies have advanced spectroscopic observations (CCD detectors, optical fibres and computing power). [3] Telescopes are giant eyes than can collect much more light that the eye of a human. A telescope is composed of mirrors and lenses

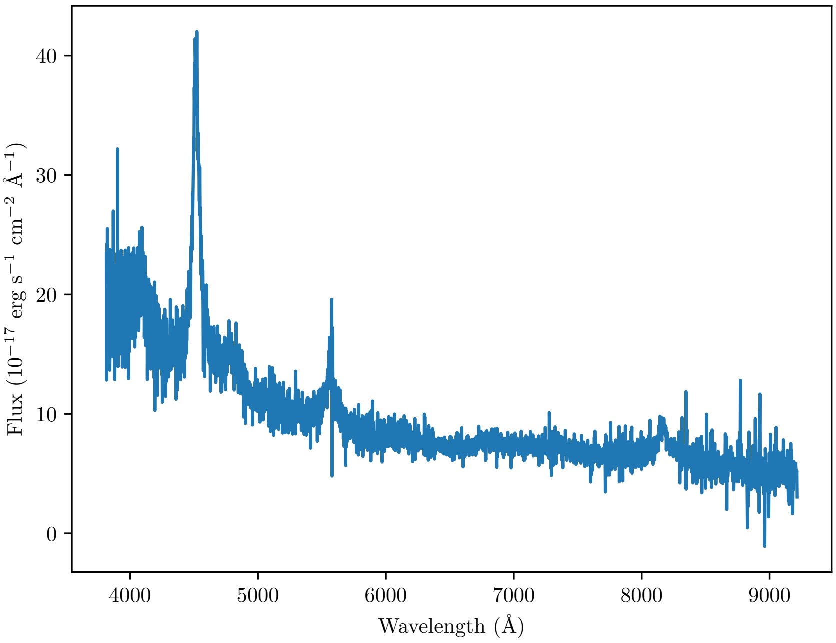


Figure 2.2: The one-dimensional spectral profile of the first-ever discovered quasar 3C 273.

that lead light into a spectrograph. A spectrograph contains a diffraction grating and a *charge-coupled device* (CCD) camera. Fraunhofer invented diffreaction grating based on the wave nature of light. It can disperse the light collected by a telescope into a spectrum while allowing more excellent dispersion than prisms. Diffraction gratings are one of the essential parts of modern spectroscopes. Then, the dispersed spectrum reveals objects composition, speed, temperature and more. [4]

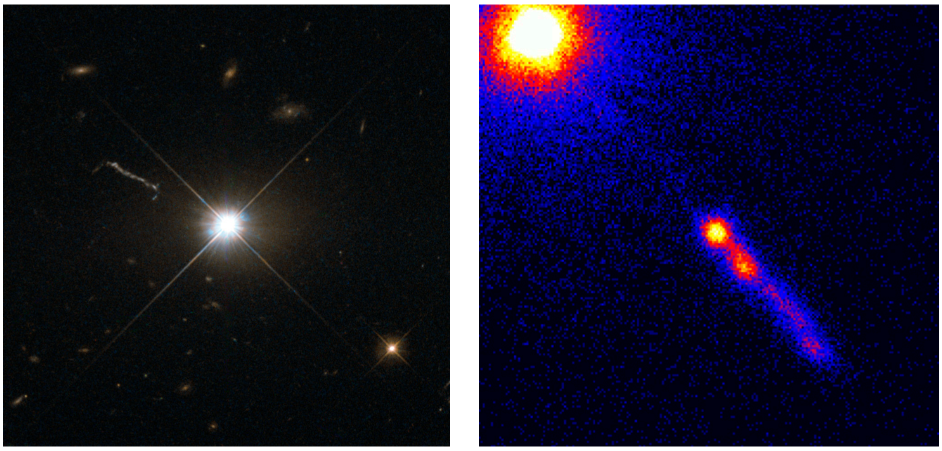
Photons carry information about observed objects to a pixel of a CCD camera in a telescope. CCD cameras require the particle nature of light (electromagnetic radiation). [2]

An important parameter of a telescope is its *field of view* and important parameters of a spectrograph are *spectral resolving power*, *signal-to-noise ration* and *full width at half maximum* for our purposes.

Spectral resolving power R expresses the capacity of a telescope to observe details of a spectrum and is defined as:

$$R = \frac{\lambda}{\Delta\lambda}, \tag{2.7}$$

where λ is considered wavelength and $\Delta\lambda$ is the smallest visible detail.



(a) Image of the bright quasar 3C 273 by ESA/Hubble is licensed under CC BY 4.0.

(b) The jet of the quasar 3C 273 by NASA/CXC/SAO/H. Marshall et al.

Figure 2.3: The left image demonstrate the star-like apperant of the QSO 273 while the right image by the Chanda X-ray Obsevatory shows important details in the powerful jet shooting from the quasar 3C 273.

Signal-to-noise ratio (SNR) determines how much we can trust our measurement concenring the power of the signal in comparison to noise. Full width at half maximum (FWHM) is the measurement of the width of a spectral line at the half of its maximum intensity measured from the continuum. FWHM is determined by the width of a slit which makes the broadening of a line. A perfect instrument would have an infinitely thin line. [3]

2.2 Quasi-Stellar Objects

Quasi-stellar (star-like) objects (also known as *quasars* and abbreviated *QSO*) are the most luminous *active galactic nuclei* (AGN). [5]

The physical model is a supermassive black hole surrounded by a gaseous accretion disk and jets (see Figure 2.3). A QSO generates energy by stress and friction in the disk outside of the black hole because no light can escape the *event horizon*. The energy is in the form of EM radiation is the strongest in the ultraviolet band. Moreover, QSOs exhibit significant cosmological redshift.

QSOs were common in the early universe probably because galaxies have run out of matter: they stop to be lumionous. Therefore, QSOs help us to study the early universe.

There are different types of QSOs: radio-loud, radio-quiet, red, broad absorption-line, type II, optically violent variable, weak emission-line.

A typical spectrum of a QSO is redshifted and contains a characteristical combination of broad and narrow emission lines. A spectrum of a QSO from SDSS is shown in Figure 2.2.

2.3 Large Spectroscopic Surveys

Since the discovery of the first QSO, there has been massive progress in spectroscopy, allowing observing a vast amount of spectra and QSOs. It started with the Bright Quasar Survey, Large Bright Quasar Survey (LBQS) and 2dF Quasar Redshift Survey. Their significant successors are the *Sloan Digital Sky Survey* (SDSS) and the *Large Sky Area Multi-Object Fiber Spectroscopic Telescope* (LAMOST) that already contain millions of spectra. We choose LAMOST and SDSS surveys for our experiments because they offer a large volume of data suitable for machine learning and for training neural networks for domain adaptation.

In the two following subsections, we introduce the parameters of their instruments, their recent data releases and corresponding catalogues of QSOs.

2.3.1 Sloan Digital Sky Survey

SDSS is in operation since 2000, and its telescope is designed to provide both a photometrically and astrometrically calibrated imaging survey and a spectroscopic survey of galaxies and QSOs. [6]

The SDSS survey uses a 2.5 m telescope located at the Apache Point Observatory, New Mexico (see Figure 2.4). The telescope has 3° field of view due to its mirror. The original spectrograph of the telescope was able to obtain 640 spectra with a wavelength coverage 380–920 nm simultaneously with spectral resolution $R \sim 1\,800$. [6, 7]

In 2009, the original spectrograph was upgraded for *Baryon Oscillation Spectroscopic Survey* (BOSS). The upgraded BOSS spectrograph covers a wavelength range 356–1 040 nm with resolving power $R \sim 2\,000$ and is capable to observe 1 000 spectra at once. [8]

Recent SDSS Data Release 14 (SDSS DR14) which corresponds to the latest catalogue of QSOs, contains more than one-third of the entire celestial sphere. The total number of optical spectra in the catalogue is 4 851 200.

The SDSS Data Release 14 Quasar (SDSS DR14Q) catalogue described in [9] contains 526 356 quasars (contamination is estimated to be about 0.5%). SDSS provides calibrated spectra covering the wavelength range 3 610–10 140 Å at a spectral resolution $1\,300 < R < 2\,500$ for all the QSOs.

The catalogue defines a QSO as an object with a certain luminosity in a specific distance and either displaying FWHM > 500 km s^{-1} for least one emission line having interesting absorption features.

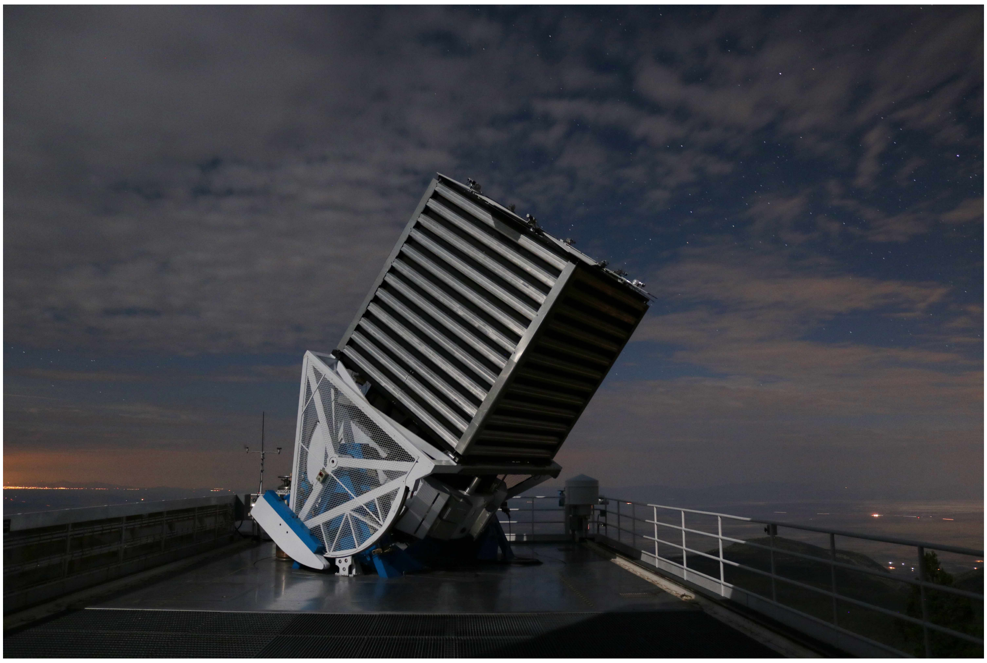


Figure 2.4: The SDSS telescope at night located in the Apache Point Observatory, New Mexico, by Patrick Gaulme is licensed under CC BY 4.0.

2.3.2 Large Sky Area Multi-Object Fiber Spectroscopic Telescope

LAMOST survey was launched in 2012. Its two primary scientific goals are to explore both extragalactic and intragalactic phenomenons. Therefore, unlike SDSS, LAMOST also observes a large volume of stars. However, the other scientific goal of LAMOST is the extragalactic spectroscopic survey of the large scale structure of the universe and the physics of galaxies. The goal includes a spectroscopic survey of nearly 10 million galaxies and *quasars* that will contribute to the study of the accretion process of massive black holes in AGNs besides other things. [10]

The LAMOST is located in Xinglong Station of National Astronomical Observatory, China. The telescope is a special telescope with a primary mirror made of 37 hexagonal spherical mirrors of total size 6.67 m times 6.05 m. The large primary mirror makes a field of view of 5°. The focal surface has 4 000 fibres connected to 16 spectrographs with 32 CCD cameras. Therefore, the telescope is capable of observing up to 4 000 spectra simultaneously in wavelength coverage of 370–900 nm with spectral resolution $R = 1\,000$ or $R = 1\,500$ depending on gratings and camera positions. [10]

LAMOST Data Release 5 v3 (LAMOST DR5) is the lastest data release which has corresponding catalogue of QSOs. The LAMOST DR5 contains 9 026 365 optical spectra, and the catalogues of QSOs of interest are DR1 [11],

Parameter of an instrument	SDSS	BOSS	LAMOST
Wavelength coverage (nm)	380–920	356–1 040	370–900
Spectral resolution R	1 800	2 000	1 250
Field of view	3°	3°	5°

Table 2.2: We list the parameters of the LAMOST and SDSS instruments to compare their different parameters. The instrument of SDSS has higher spectral resolution and wider wavelength coverages while LAMOST can observer bigger areas of the sky.

DR2&3 [12] and DR4&5 [13] that in total contains 42 552 spectra of QSOs.

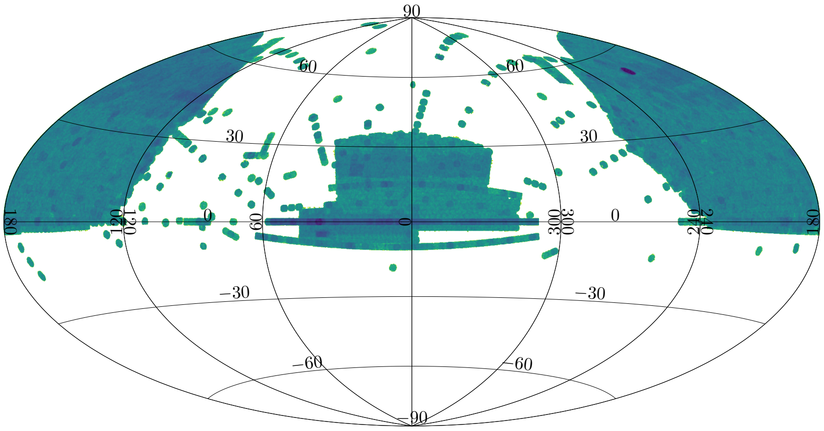
2.3.3 Comparison of the Spectral Data

Now, we compare the SDSS and LAMOST survey to prove their suitability for domain adaptation. The surveys are mainly different in term of instruments, sky coverage and targeting strategy.

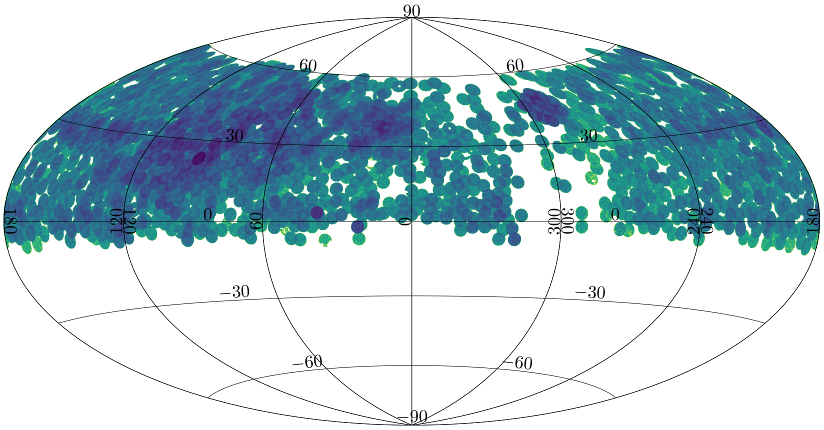
We summarise the main parameters of instruments of the surveys in Table 2.2. We see that LAMOST has lower resolution and shorter wavelength coverage than SDSS. However, SDSS has a smaller field of view.

Sky coverage is very connected to the targeting strategy or scientific goals. Figure 2.5 displays sky coverage of both SDSS and LAMOST. We see that SDSS does not observe our Milky Way galaxy. On the other hand, LAMOST observes everywhere on the northern hemisphere but is not able to observe close to zenith due to its construction limits. From the perspective of coverage of QSOs depicted in Figure 2.6, LAMOST did not observe QSOs in some part of the sky where QSOs are abundant according to SDSS.

We conclude that the two surveys seem to be suitable for domain adaptation because their instruments create spectra with wavelength range and different resolution. Moreover, observations of SDSS and LAMOST survey has different distributions.

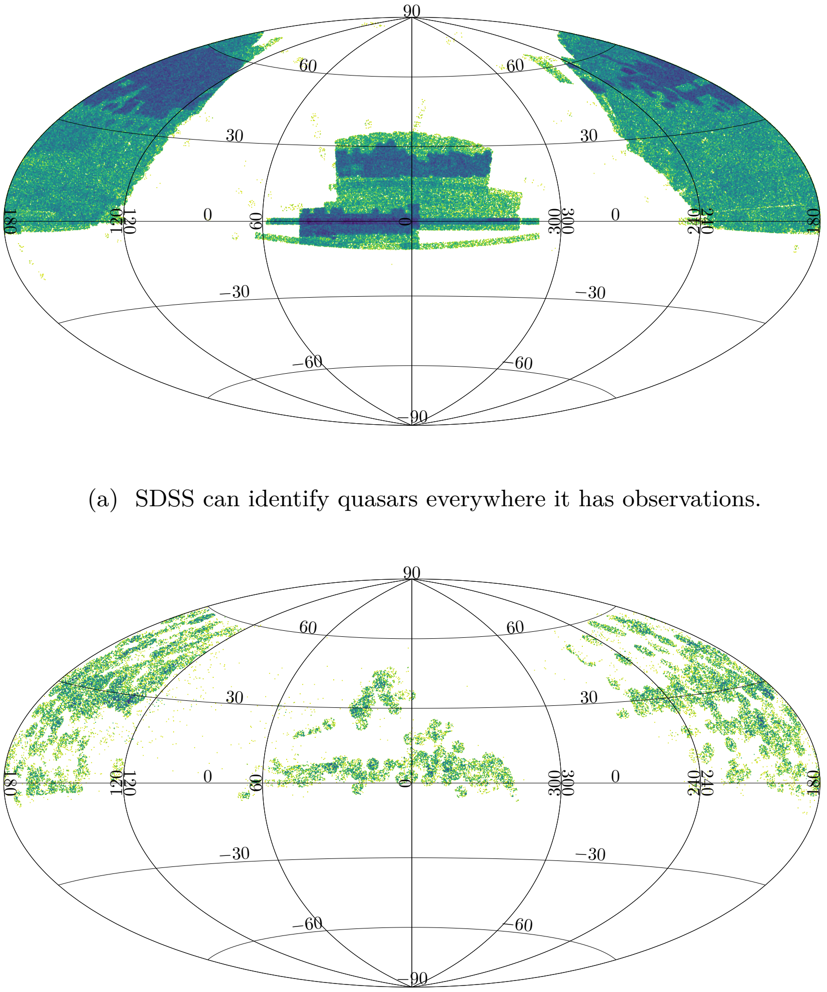


(a) Sky coverage of SDSS corresponds to its primary scientific goal to observe galaxies and QSOs. Therefore, it has almost no observations in the area of our Milky Way galaxy.



(b) LAMOST observes both the extragalactic and intragalactic objects. Therefore, its distribution of observations is almost uniform on the northern hemisphere. An exception is the surroundings of the zenith due to the telescope construction limits.

Figure 2.5: Comparison of sky coverage of SDSS and LAMOST.



(a) SDSS can identify quasars everywhere it has observations.

(b) Although LAMOST has observations in similar areas as SDSS, its catalogues contain an only small amount of QSOs in comparison to SDSS.

Figure 2.6: Sky positions of QSOs listed in either SDSS or LAMOST catalogues.

Chapter **3**

Domain Adaptation

In this chapter, we survey the current state of domain adaptation using neural network models in the context of the machine and transfer learning. We formally define the domain adaptation scenario and categorise it according to recent surveys by Wang and Deng [14] and Csurka [15]. From each category, we select a representative method for our experiments. Then, we introduce in detail selected methods. We conclude the chapter with applications of domain adaptation in astronomy.

3.1 Domain Adaptation in the Context of Machine Learning

Domain adaptation is a subfield of transfer learning, which is part of machine learning. Machine learning has a common assumption that training and test data are *independent and identically distributed* (IID) that means data are drawn from the same feature space and the same distribution. [16] When this assumption does not hold transfer learning or domain adaptation come into play. Moreover, from the biological point of view the assumption seems to limit because humans seem to have natural ways to transfer knowledge from previous experience to new challenges. [17]

Transfer learning is defined in most papers regarding the survey by Pan and Yang [18]. There is a more recent survey by Weiss et al. [19] that has the benefit of containing newer methods than the survey by Pan and Yang [18]. However, its definition of transfer learning and domain adaptation is the same.

Pan and Yang [18] define transfer learning as the ability of a system to recognise and apply knowledge and skill learned in previous problems to novel problems, and they introduce the notion of a *domain* and a *task*.

A domain consists of a *feature space* X and a *marginal probability distribution* $P(X)$ (marginal expresses that it is summed over a label space Y).

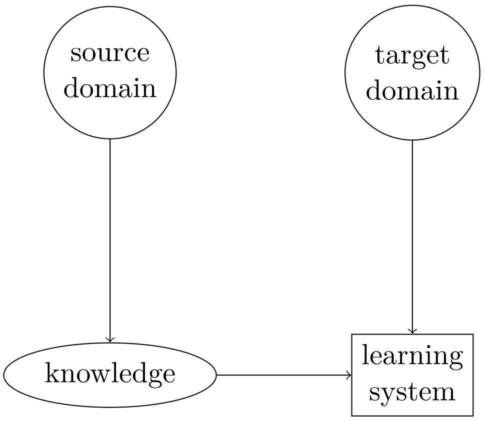


Figure 3.1: Schematic diagram of domain adaptation depicts the fundamental intuition behind it. A domain adaptation model has to extract as much knowledge as possible from the source domain and apply it to the target data.

Therefore, a domain is a tuple $\mathcal{D} = [X, P(X)]$ of a d-dimensional feature space $X \subset \mathbb{R}^d$ and a marginal probability distribution $P(X)$.

Given a specific domain $\mathcal{D}$, a task consists of a *label space* and an *objective predictive function* which is not observed but learned from training data. Therefore, a task is also defined as a tuple $\mathcal{T} = [Y, P(Y|X)]$ where Y is a *label space* and $P(Y|X)$ a *conditional probability distribution* which we like to model as closely as possible.

When we are given a transfer learning problem, we have to identify a source domain and a source learning task, a target domain and a target learning task. Then, transfer learning aims to help improve the learning of the target predictive function in the target domain using knowledge from the source domain and the source task, where the domains are different or the tasks are different. As we will see, domain adaptation is the case when the source and target domains are different, while the source and target tasks are the same. [18]

For example, we have introduced in Chapter 2 two domains. The source domain is the SDSS DR14 and the LAMOST DR5 is the target domain. We have shown that the domain have different data distribution mainly because of their targeting strategies and instruments. The task to identify QSOs is the same for both SDSS DR14 and LAMOST DR5 as we defined it.

Lastly, Torrey and Shavlik in [17] warn that sometimes transfer learning can be harmful. Performance of our machine learning algorithm might suffer when the source and target domains or tasks are not sufficiently related. When the usage of source data degrades the performance, the situation is called a *negative transfer*. On the other side, when the performance is improved, we talk about a *positive transfer*.

3.2 Theory and Formalization of Domain Adaptation

Now, we introduce the crucial concept of our work: *domain adaptation* (DA). DA is a particular case of transfer learning that leverages data from the source domain to learn a classifier for a target domain while the tasks are the same. It is assumed that source and target domains are not identical but related. If the domains were identical, it would be a standard machine learning problem. Therefore, there is a distribution discrepancy between the source and target domains. [15]

More formally, DA is the scenario of transfer learning when the source $\mathcal{D}^s = [X^s, P(X^s)]$ and target $\mathcal{D}^t = [X^t, P(X^t)]$ domains are different ($\mathcal{D}^s \neq \mathcal{D}^t$), but the source $\mathcal{T}^s$ and target $\mathcal{T}^t$ tasks are the same ($\mathcal{T}^s = \mathcal{T}^t$). The first condition implies that $X^s \neq X^t$, $P(X^s) \neq P(X^t)$ or both are true. [18] In our case, the domains are different because the two archives have different target selection strategies, and the instruments are different (see Section 2.3).

Based on the condition $\mathcal{D}^s \neq \mathcal{D}^t$, we categorise DA into *homogeneous* and *heterogeneous*. Homogeneous DA is setting in which the source X^s and target X^t feature spaces are the same ($X^s = X^t$) while in *heterogeneous* DA, the source and target spaces are different ($X^s \neq X^t$). [15]

Our case is the homogeneous DA, where the feature spaces are identical ($X^s = X^t$). Still, the source and target data have different distributions ($P(X^s) \neq P(X^t)$). Therefore, further, we focus on the homogeneous DA because our spectra can be prepared into the same feature space $\mathbb{R}^{3659}$.

The second categorisation of DA is according to availability of labels in the target domain. Wang and Deng [14] distinguish *supervised*, *semi-supervised* and *unsupervised* DA:

- supervised DA: labelled data in the target domain are present;
- semi-supervised DA: we have only a minimal amount of labelled data in the target domain while most of the data is unlabelled;
- unsupervised DA: no labelled data are available in the target domain.

We phrase our problem as *unsupervised DA* because the LAMOST survey uses different criteria for identification of QSOs than SDSS (see Section 2.3).

Now, we have all the tools to define our homogeneous unsupervised DA problem. In correspondence with [20], we consider a classification problem within an input space X with L possible labels from a set $Y = \{0, \dots, L-1\}$. The source $\mathcal{D}_s$ and target $\mathcal{D}_t$ domains have different distributions over $X \times Y$. From the source and target domains we have *labelled source sample* S drawn IID from $\mathcal{D}_s$ and *unlabelled target sample* T drawn IID from $\mathcal{D}_t^X$ (the marginal distribution of $\mathcal{D}_t$ over X):

$$S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n} \sim (\mathcal{D}_s)^n, \tag{3.1}$$

$$T = \{\mathbf{x}_i\}_{i=n+1}^{N} \sim (\mathcal{D}_t^X)^{n'}, \tag{3.2}$$

where $N = n + n'$ is the total number of examples (note that the examples are numbered from 1 to N). Concretely, the input space X is equal to $\mathbb{R}^{3659}$, $y_i \in Y$ which is equal to $\{0, 1\}$ implying $L = 2$ (0 stands for a non-QSO object and 1 for a QSO) and a vector $\mathbf{x}_i$ is an astronomical spectrum (its fluxes in specified wavelengths).

Further categorisation of methods of DA will help us to separate methods based on neural networks. Csurka in [15] divides DA into two categories: *shallow DA* methods and *deep DA*. Shallow DA methods are not based on neural networks but rather on statistical theory. Deep DA methods are based on neural networks augmented for DA.

Furthermore, according to Wang and Deng [14], shallow DA methods can be categorised into two classes. The first class is *instance-based* DA (the discrepancy is reduced by reweighting the source instances) and the second is *feature-based* DA (tries to learn a common shared space in which the discrepancy diminishes). However, shallow DA methods are not interesting for us because they do not utilise neural networks directly. Still, one possibility is to use a neural network as a feature extractor. [15]

The following section introduces the other category of deep DA, which takes advantage of neural networks.

3.3 Neural Networks in Domain Adaptation

Already, Donahue et al. state in the paper on DeCAF [21] that deep neural networks learn more transferable features that can help with transfer learning. However, the DeCAF paper also shows that the performance is still affected by the domain shift. Therefore, there is enough space for specialised deep architectures for DA.

The group of DA methods using neural networks is called *deep DA*. Wang & Deng define it in [14] as methods that utilise deep neural networks to enhance the performance of DA. Deep DA architectures can be trained with *backpropagation*. Such architectures can be trained with backpropagation learning domain invariant representation discriminative for a common task. For example, a network is extended with a particular loss, a domain classifier or an autoencoder as we will see next.

Approaches to deep DA were initially divided by Csurka in [15] into three categories according to training loss. Then, Wang and Deng further improve and detail the categorisation into [14]:

- *discrepancy-based* DA fine-tunes the neural network with target data to diminish the domain shift;

- *adversarial-based* DA encourages domain confusion by using discriminators with an adversarial objective; and
- *reconstruction-based* DA uses data reconstruction auxiliary task to learn domain invariant features.

We summarise the deep DA methods categorisation in the mind map of Figure 3.2. The mind map also gives further subcategories with some example methods. Next, we detail the three categories of DA according to the survey by Wang and Deng [14].

3.3.1 Discrepancy-based Deep Domain Adaptation

The first category is discrepancy-based deep DA methods that use either labelled or unlabelled target data to fine-tune a deep neural network to diminish the discrepancy between the source and target domains.

The discrepancy-based DA methods are subdivided based on *class*, *statistic*, *architecture* and *geometric* criterion by [14]. *Class criterion* methods use label information to do DA. For example, the knowledge is transferred in the form of soft labels or pseudo labels. *Statistic criterion* methods align some statistic of the source and target distribution. The most used methods reduce domain shift with the maximum mean discrepancy, correlation alignment or Kullback-Leibler divergence. Methods with *architecture criterion* adjust the network topology in order to learn more transferable features. The last subcategory is *geometric criterion* methods which want to diminish the difference between the source and target distributions based on their geometrical properties.

We choose to focus on the statistic criterion category in this work because its methods can be used in an unsupervised fashion, focus straight on the different source and target distributions and perform very well on two-dimensional images. We think class criterion subcategory and architecture criterion subcategories are unsuitable. They focus mostly on supervised DA, and the geometric criterion gives much worst result in comparison to methods based on statistic criterion.

The most successful statistic criterion methods are Deep Domain Confusion (DDC) [22], Deep Adaptation Network (DAN) [23], Joint Adaptation Network (JAN) [24] and Deep Correlation Alignment (Deep CORAL) [25]. DDC is the fundamental method of the statistic criterion subcategory, and both DAN and JAN are its extensions. Deep CORAL is very similar to DDC, but it aligns the second-order statistics. To keep things simple, we selected to experiment with DDC and Deep CORAL. We describe the theory behind them next.

Deep Domain Confusion (DDC) [22] finds domain invariant representation while learning to predict class labels. The intuition behind DDC is that learning representation that minimises the distance between the source and target

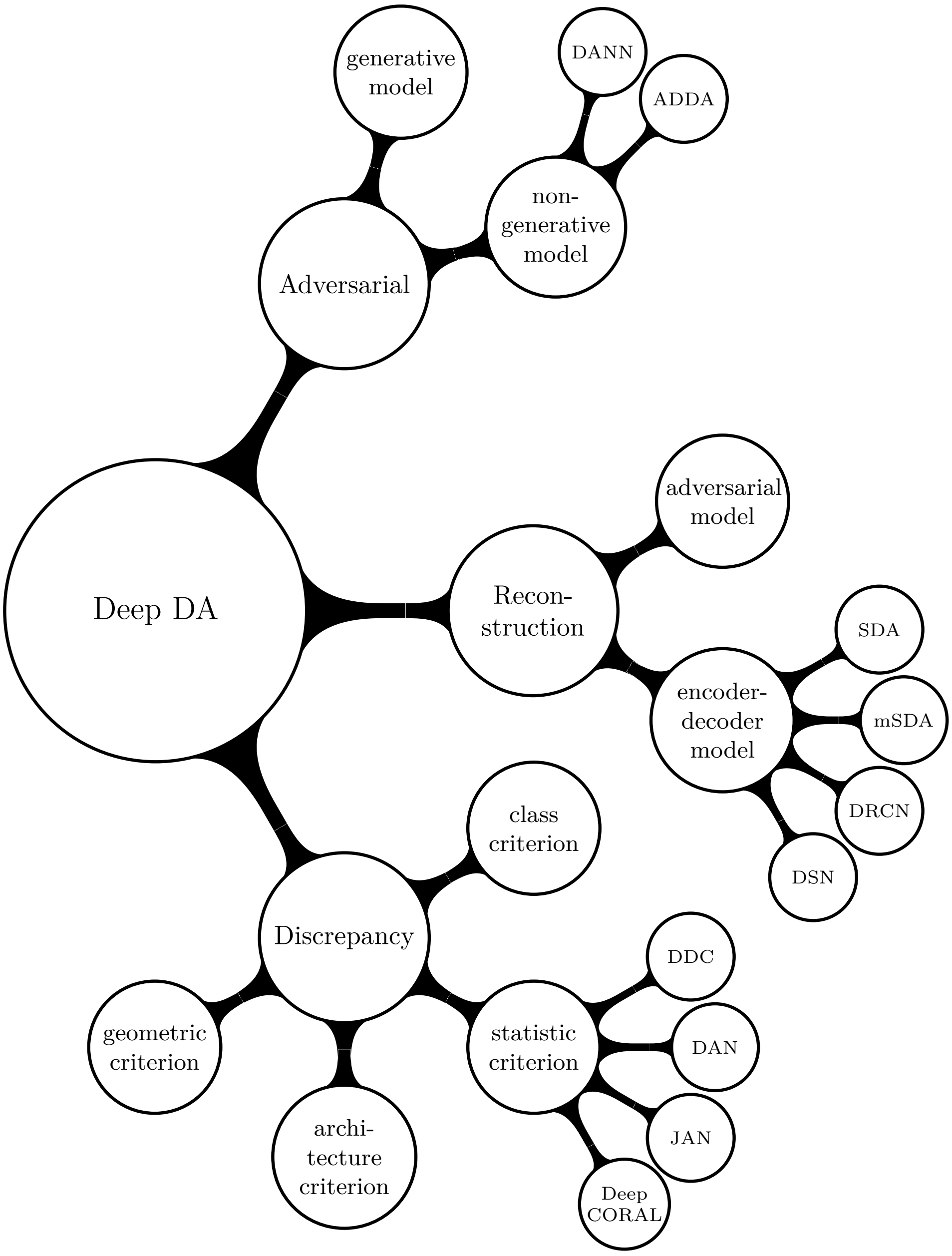


Figure 3.2: This mind map presents categorisation of deep DA divided as we described it. Moreover, we also show concrete deep DA methods that we have mentioned.

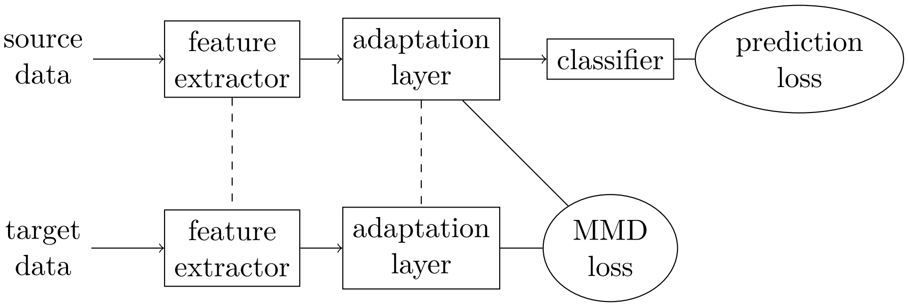


Figure 3.3: Schema of Deep Domain Confusion

distributions will help a classifier trained on source labelled data to be directly applied to the target domain.

DDC optimises loss function that includes both prediction error and *domain confusion* loss to learn domain invariant representation. The invariant representation is achieved by incorporating an *adaptation layer* into a deep *convolutional neural network* (CNN) with a domain confusion loss computed via *maximum mean discrepancy* (MMD). MMD is a standard distribution distance metric which is empirically approximated by:

$$\mathrm{MMD}(S,T) = \left\| \frac{1}{n}\sum_{i=1}^{n}\phi(\mathbf{x}_i) - \frac{1}{n'}\sum_{i=n+1}^{N}\phi(\mathbf{x}_i) \right\|, \tag{3.3}$$

where $\phi : X \to Z$ is a feature extractor from the data space X to a feature space Z that operates on both source and target data points $\mathbf{x}_i \in X$.

Using the adaptation layer with the domain loss function, DDC claims to learn a representation that is both domain invariant but still offers strong semantic separation enabling to learn a robust label classifier. Therefore, the goal of DDC is to minimise a loss that incorporates both domain confusion loss and classification loss:

$$\mathcal{L}_{\mathrm{DDC}} = \mathcal{L}_C(S) + \lambda \mathrm{MMD}^2(S,T), \tag{3.4}$$

where $\mathcal{L}_C$ stands for classification loss on the labelled source data S and to control the strength of domain confusion, DDC introduces a hyperparameter λ ($\lambda = 0.25$ in experiments of DDC).

In comparison with DDC, *Deep Correlation Alignment* (Deep CORAL) uses *correlation alignment* (CORAL) instead of MMD. CORAL aligns second-order statistics of the source and target distributions. Then, Deep CORAL is a deep neural network that incorporates the differentiable CORAL loss.

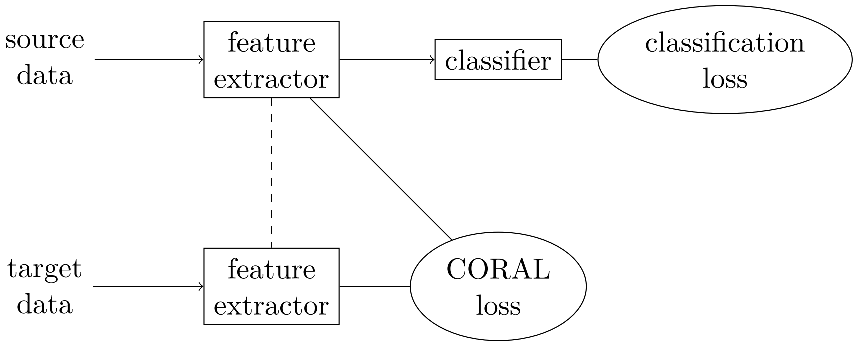


Figure 3.4: Schema of Deep Correlation Alignment

CORAL loss is defined as the distance between covariances of the source and target features extracted from a layer of a deep neural network:

$$\mathcal{L}_{\text{CORAL}}(S,T) = \frac{1}{4d^2}\|C(S) - C(T)\|_F^2, \tag{3.5}$$

where $\|\cdot\|_F^2$ is the squared matrix Frobenius norm and the function C return a covariance matrix of a given set:

$$C(D_A) = \frac{1}{|A|-1}(D_A^\top D_A - \frac{1}{|A|}(\mathbf{1}^\top D_A)^\top(\mathbf{1}^\top D_A)), \tag{3.6}$$

where A is a set of examples that are either the source sample S or the target sample T and D_A is a design matrix corresponding to the set that is a matrix where each row is an example. The $\mathbf{1}$ is a column vector with all elements equal to 1.

The final composed loss consist of a classification loss $\mathcal{L}_C$ and the CORAL loss:

$$\mathcal{L}_{\text{Deep CORAL}} = \mathcal{L}_C(S) + \lambda\mathcal{L}_{\text{CORAL}}(S,T), \tag{3.7}$$

where λ is a trade-off hyperparameter similar to the one in the DDC loss in Equation 3.4.

Note that Deep CORAL does not introduce an adaptation layer but uses a layer that is already in the network. The CORAL loss can be even applied to several layers.

3.3.2 Adversarial-based Deep Domain Adaptation

Adversarial-based deep DA utilises a domain discriminator that tries to distinguish whether a sample comes from the source or target domain. If we can confuse the discriminator, we will also achieve domain confusion through the adversarial objective.

The adversarial deep DA methods are divided into *generative* and *non-generative* models by [14]. *Generative models* create synthetic target data according to source data while keeping source data labels. These models are usually based on Generative Adversarial Network (GAN) [26]. Rather than generating synthetic examples, *non-generative models* learn via an adversarial objective a feature extractor that produces domain invariant representation of an example.

We explore the second non-generative subcategory because we think that proper representation is satisfactory, and we consider synthetic data generation as overkill for our task.

Domain-Adversarial Neural Network (DANN) [20] is the fundamental algorithm of non-generative deep DA. Other methods like Adversarial Discriminative Domain Adaptation (ADDA) [27] built on the idea of DANN. Therefore, we are convinced that DANN is the right model to try to our problem in our hands.

Domain-Adversarial Neural Network (DANN) [20] is an adversarial representation learning approach for domain adaptation. It is based on the idea that useful features for DA cannot discriminate between source and target domains while maintaining discriminative properties for a classification task.

DANN combines a feature extractor, label predictor and domain classifier is a single architecture and can be trained with a standard backpropagation learning algorithm. Because of that, DANN proposes a *gradient reversal layer* (GRL) that can be incorporated in almost any feed-forward neural network. Using the GRL, DANN jointly optimises feature extractor and two discriminative classifiers. The first discriminative classifier is a label predictor that predicts classes and the second is a domain classifier that discriminates between source and target domains.

The feature extractor is trained jointly to minimise a classification loss and maximise the loss of the domain classifier. Thus, the domain classifier and feature extractor are learning in adversarial fashion to encourage domain invariant features.

To define the learning objective let $G_f(\cdot;\theta_f)$ be the neural network feature extractor with parameters θ_f, $G_y(\cdot;\theta_y)$ be the label predictor with parameters θ_y that takes features from G_f as inputs and outputs class probabilities, and $G_d(\cdot;\theta_d)$ be the domain classifier with parameters θ_d. Training DANN is optimising prediction loss and domain loss:

$$
\begin{aligned}
E(\theta_f, \theta_y, \theta_d) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_y(G_y(G_f(\mathbf{x}_i; \theta_f), \theta_y), y_i) \\
- \lambda \Bigg(\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_d(G_d(G_f(\mathbf{x}_i; \theta_f), \theta_d), d_i) \\
+ \frac{1}{n'} \sum_{i=n+1}^{N} \mathcal{L}_d(G_d(G_f(\mathbf{x}_i; \theta_f), \theta_d), d_i) \Bigg)
\end{aligned} \tag{3.8}
$$

by finding the saddle point $\hat{\theta}_f$, $\hat{\theta}_y$, $\hat{\theta}_d$ that satisfy:

$$
(\hat{\theta}_f, \hat{\theta}_y) = \operatorname*{arg\,min}_{\theta_f, \theta_y} E(\theta_f, \theta_y, \hat{\theta}_d), \tag{3.9}
$$

$$
\hat{\theta}_d = \operatorname*{arg\,max}_{\theta_d} E(\hat{\theta}_f, \hat{\theta}_y, \theta_d). \tag{3.10}
$$

where $\mathcal{L}_y$ is a classification loss, $\mathcal{L}_d$ is a domain loss, λ is a trade-off hyperparameter and d_i is a binary variable (a domain label):

$$
d_i = \begin{cases} 0 & \text{if } i \in \{1, \ldots, n\}, \\ 1 & \text{if } i \in \{n+1, \ldots, N\}. \end{cases} \tag{3.11}
$$

The DANN paper shows that saddle points can be found using gradient updates similar to *stochastic gradient descent* (SGD):

$$
\theta_f \leftarrow \theta_f - \mu \left(\frac{\partial \mathcal{L}_y^i}{\partial \theta_f} - \lambda \frac{\partial \mathcal{L}_d^i}{\partial \theta_f} \right), \tag{3.12}
$$

$$
\theta_y \leftarrow \theta_y - \mu \frac{\partial \mathcal{L}_y^i}{\partial \theta_y}, \tag{3.13}
$$

$$
\theta_d \leftarrow \theta_d - \mu \lambda \frac{\partial \mathcal{L}_d^i}{\partial \theta_d}. \tag{3.14}
$$

However, the update in Equation 3.12 has subtraction in it instead of an addition. DANN overcomes the substruction by incorporating the GRL $\mathcal{R}(\mathbf{z})$ between the feature extractor and domain classifier. The GRL has no parameters, and during forward propagation, the GRL acts as identity:

$$
\mathcal{R}(\mathbf{z}) = \mathbf{z}. \tag{3.15}
$$

However, in backpropagation, it negates the gradient:

$$
\frac{d\mathcal{R}}{d\mathbf{z}} = -\mathbf{I}, \tag{3.16}
$$

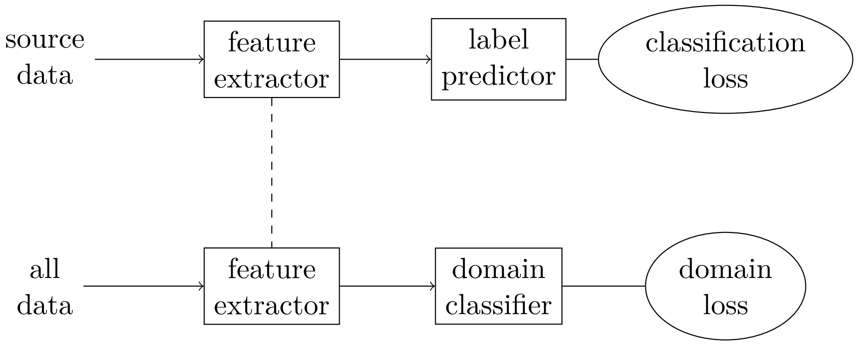


Figure 3.5: Schema of Domain-Adversarial Neural Network

where **I** is the identity matrix, and **z** is the representation produced by the feature extractor $G_f(\cdot, \theta_f)$. Now, DANN can seamlessly work with SGD because the objective to be optimised with gradient descent is transformed into:

$$\begin{aligned} E(\theta_f, \theta_y, \theta_d) = \frac{1}{n}\sum_{i=1}^{n} &\mathcal{L}_y(G_y(G_f(\mathbf{x}_i; \theta_f), \theta_y), y_i) \\ &- \lambda \left(\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_d(G_d(\mathcal{R}(G_f(\mathbf{x}_i; \theta_f)), \theta_d), d_i) \right. \\ &+ \left. \frac{1}{n'}\sum_{i=n+1}^{N} \mathcal{L}_d(G_d(\mathcal{R}(G_f(\mathbf{x}_i; \theta_f)), \theta_d), d_i) \right) . \end{aligned} \tag{3.17}$$

The GRL changes the update 3.12 to version fully compatible with SGD:

$$\theta_f \leftarrow \theta_f - \mu \left(\frac{\partial \mathcal{L}_y^i}{\partial \theta_f} + \lambda \frac{\partial \mathcal{L}_d^i}{\partial \theta_f} \right) . \tag{3.18}$$

3.3.3 Reconstruction-based Deep Domain Adaptation

The last group of methods is based on a data reconstruction auxiliary task. The reconstructor forces to find a shared representation of the source and target domains.

According to [14], subcategories are *encoder-decoder* and *adversarial* reconstruction. The more uncomplicated *encoder-decoder* category exploits stacked autoencoders [28] or stacked convolutional autoencoders [29] to do the reconstruction task. On the other hand, the methods based on *adversarial reconstruction* use cyclic mapping obtained via a GAN discriminator [26].

Furthermore, we explore the *encoder-decoder* methods that are a good fit for our problem because they are not that complicated as GAN-based method of the second subcategory.

It all started with Stacked Denoising Autoencoder (SDA) [30] and it extension marginalised SDA (mSDA) [31] for DA on text sentiment analysis data. While SDA and mSDA are base on fully-connected networks, Deep Reconstruction-Classification Network (DRCN) [32] uses stacked convolutional autoencoder for the reconstruction task on images. Moreover, Domain Separation Networks (DSN) [33] is based on the same idea as DRCN. However, it uses three separate encoders to model shared representation and also private specific representations of the source and target data. We choose to experiment with DRCN because it is designed for image data while being more straightforward than DSN.

Deep Reconstruction-Classification Network (DRCN) [32] is a CNN that learn both supervised source labelled classification and unsupervised target data reconstruction. The encoder is shared between both task, but decoding parameters are separate. The data reconstruction task is an auxiliary task supposed to help to learn good feature representation beneficial for the DA scenario. The intuition behind DRCN is that good representation for DA should capture both the properties for classification and the data structure (reconstruct well the target domain).

DRCN is composed of a label predictor for classification and a convolutional autoencoder for target data reconstruction. Let define $F_c : X \rightarrow Y$ as the label predictor and $F_r : X \rightarrow X$ is the data reconstructor. The two functions are composed of three components:

- an encoder feature mapping $G_e : X \rightarrow Z$;
- a decoder $G_d : Z \rightarrow X$; and
- a classificator $G_l : Z \rightarrow Y$;

where Z is a latent feature space to which DRCN encodes data. Given the above component, the classification pipeline is $F_c(\cdot; \theta_c) = G_l(G_e(\cdot; \theta_e); \theta_l)$, and the reconstruction pipeline is $F_r(\cdot; \theta_r) = G_d(G_e(\cdot; \theta_e); \theta_d)$ where $\theta_c = \{\theta_e, \theta_l\}$ are parameters of the classificator and $\theta_r = \{\theta_d, \theta_e\}$ are parameters of the reconstructor. Note that θ_e is shared between the label predictor and the autoencoder.

The model is jointly optimised for classification and reconstruction tasks. Therefore, the learning objective of DRCN contains classification and reconstruction loss, respectively:

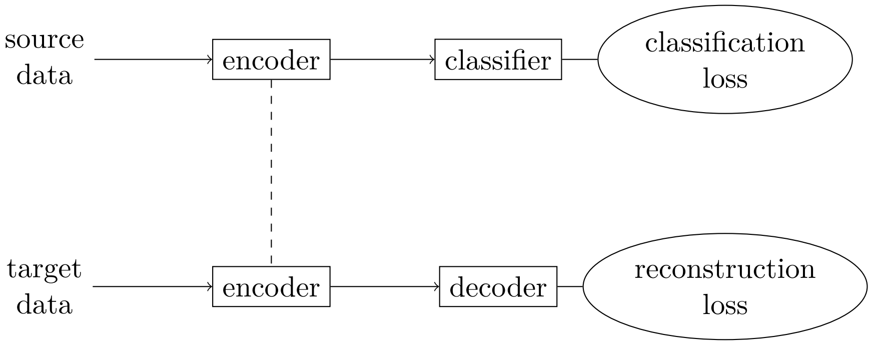


Figure 3.6: Schema of Deep Reconstruction-Classification Network

$$\mathcal{L}_c(S) = \sum_{i=1}^{n} l_c(F_c(\mathbf{x}_i; \theta_c), y_i), \tag{3.19}$$

$$\mathcal{L}_r(T) = \sum_{i=n+1}^{N} l_r(F_r(\mathbf{x}_i; \theta_r), \mathbf{x}_i), \tag{3.20}$$

where l_c is a form of classification loss and l_r is a reconstruction loss, for example, the squared loss:

$$l_r = \|\mathbf{x}_i - F_r(\mathbf{x}_i)\|_2^2. \tag{3.21}$$

By combining the terms 3.19 and 3.20, DRCN aims to minimise the following objective:

$$\mathcal{L}_{\text{DRCN}} = \lambda \mathcal{L}_c(S) + (1 - \lambda)\mathcal{L}_r(T), \tag{3.22}$$

where $\lambda \in [0, 1]$ is a hyperparameter of the trade-off between the two loss functions. The objective 3.22 can be optimised with SGD.

3.4 Previous Applications of Domain Adaptation in Astronomy

As we have shown in Section 2.3, DA is of great interest to astronomers because of different instruments, measurements and observation distribution. Therefore, we survey the current state of DA in astronomical applications in this section.

If we have a common set of observed stars in both archives (supervised DA), then we can map them and learn a transfer function. Ho et al. [34] did exactly that because they found a common set of 9 952 spectra in both APOGEE and LAMOST archive. Using the common set, they trained the Cannon method [35], and they used the model to transfer some stellar physical parameters from APOGEE to LAMOST.

In the case, when there is no common set (unsupervised DA) Gupta et al. [36] experimented with *subspace alignment* [37] and *kernel mean matching* [38] followed by *active learning* [39]. In the case of subspace alignment, the negative transfer occurred while the kernel mean matching seems very promising in the task of supernova classification. However, they observed that those shallow methods (subspace alignment and kernel mean matching) are not sufficient on their own, and the active learning phase is crucial. Nevertheless, active learning requires human expert interaction, so it is not automatic and depends on domain knowledge.

Then, Vilalta et al. [40] extended the work of Gupta et al. [36]. Vilalta et al. used a supervised *maximum a posteriori* (MAP) approach to learning a prior on the model parameters from a spectroscopic source domain and then use this prior distribution to learn a model in a photometric target domain. Concretely, Vilalta et al. put a prior on the number of layers of a neural network and then used active learning.

Richards et al. [41] faced a similar situation, as Gupta et al. Richards et al. introduce the problem as sample selection bias [42] or covariate shift [43] when different distributions generate the source and target data. That is precisely the problem we have defined as DA. Richards et al. experimented with random forest in combination with three DA methods: *importance weighting* [42], *co-training* [44] and *active learning* [39]. Active learning works best while importance weighting and co-training achieve negative transfer.

It seems that active learning is crucial to achieving good result on astronomical data when using shallow DA methods. We speculate that it is the nature of scientific data that makes shallow DA method unusable on its own.

Next, we describe a supervised deep DA approach that shown to work well on astronomical data. Therefore, we think neural network DA methods should be promising in astronomy. The term transfer learning has been recently used in the context of deep learning. However, transfer learning in the context of deep learning means something more concrete than what we defined as transfer learning previously. Transfer learning in the context of deep learning is the specific situation when a pre-trained deep neural network model is taken, and its last layers are retrained with the target domain data.

Ackermann et al. [45] employed a deep CNN with the transfer learning approach in the context of deep learning to detect galaxy merges from two-dimensional images. They took the pre-trained Xception CNN [46] and fine-tuned its last layers with images of galaxy merger labelled in a citizen science Galaxy Zoo project [47]. This transfer learning approach allowed them to lower the best error rate so far by 15%.

Next application of deep DA method in astronomy is *Deep Variational Transfer* (DVT) [48]. DVT is a semi-supervised model based on variational autoencoders [49] and mixture models. They encode the source and target data into a shared latent space and identify clusters with the labelled target data. They experimented with light curves (time series of light intensity) of stars.

This section shows that applications of deep DA in astronomy are minimal. Therefore, our work will be almost the first to explore the ideas of deep DA in astronomy with huge potential to discover either a new tool for astronomy or a new challenge for deep DA.

Chapter **4**

Experiments with Deep Domain Adaptation

In previous chapters, we have selected suitable astronomical data, surveyed and chosen suitable deep domain adaptation methods. Now, we carry out experiments with the DDC, Deep CORAL, DANN and DRCN on astronomical spectra from the SDSS and LAMOST spectroscopic sky surveys.

Firstly, we create the source and target datasets for experiments. Then, we use PCA, t-SNE and UMAP to reduce the data to two-dimensions, so we can visualise the data and investigate distributions of both the source and target data. Thirdly, we introduce a CNN baseline model which serves as a benchmark for comparison of the performance of the deep domain adaptation methods. Finally, we employ four deep domain adaptation methods and evaluate the adapted results to see if astronomical spectroscopy can benefit from deep domain adaptation.

4.1 Data Preparation

Our data of source domain consists of 4 851 200 optical spectra from the SDSS DR14 catalogue and the corresponding SDSS DR14Q catalogue of 649 791 spectra of 526 356 QSO objects (we have to distinguish an object and a spectrum because each astronomical object could be observed multiple times having multiple spectra). Both catalogues are introduced in Subsection 2.3.1. However, 20 279 spectra of QSOs cannot be identified because there is a bug in the SDSS DR14Q catalogue. Therefore, we can identify only 629 512 spectra of all QSOs. Next, we need to cross-match the SDSS DR14 and DR14Q catalogues to merge the data stored in individual FITS files with QSO labels. The cross-matching is based on a triplet of a *plate number*, a *Modified Julian Date of observation* and a *fibre number* that is unique to each spectrum. Additionally to the 20 279 spectra of QSOs lost due to the bug in the DR14Q catalogue, we were unable to cross-match 55 QSOs with the SDSS DR14 catalogue. Therefore, we have 629 457 spectra of QSOs for which we have actual data in FITS files and not only metadata in catalogues.

The complement to the source domain in domain adaptation is the target domain. We selected data from LAMOST DR5 to be target domain data for reasons described in Subsection 2.3.2. The LAMOST DR5 general catalogue contains 9 026 365 spectra, and the complete catalogue of QSOs has 42 552 spectra. Again, we cross-matched the LAMOST DR5 catalogue and the catalogue of QSOs according to a quartet of a *plan identifier*, a *local Modified Julian Date* (one less the Modified Julian Date), a *spectrograph identifier* and an *identifier of fibre*. We were able to cross-match 31 755 spectra of QSOs with the general catalogue effectively losing 10 797 spectra of QSOs. We believe that LAMOST has sound reasons for not including those spectra in the LAMOST DR5 catalogue.

However, the labels of QSOs from LAMOST are incompatible with labels from SDSS because the criteria of what is a QSO are different in SDSS and LAMOST (see Section 2.3). For us, the ground truths are labels of SDSS DR14Q catalogue while the labels of LAMOST serves only for evaluation purposes, not for training. Therefore, there might be spectra truly QSOs in LAMOST not yet identified by LAMOST biasing our performance metrics.

Having assigned labels of QSOs to individual spectra, we need to extract the spectra from individual FITS files because learning of neural networks requires datasets to be in the form of design matrices. A design matrix contains a different example (a spectrum) in each row. In contrast, each column of the design matrix corresponds to a different feature (a measurement of flux in a specific wavelength). [50]

Fortunately, SDSS and LAMOST spectra have a common wavelength grid in logarithmic wavelengths evenly space by 0.0001. Although all spectra have a common wavelength grid, the minimal and maximal wavelengths are different for each spectrum. Figure 4.1 displays histograms of minimal and maximal wavelengths of all LAMOST spectra. In this work, we aim to find QSOs in the LAMOST DR5. Therefore, we would like to keep as many spectra as possible from the LAMOST DR5. To keep all spectra from the LAMOST DR5, we have to select wavelength range starting at 3 839.7244 Å (3.5843 in logarithmic wavelength) which is the maximum from minimal wavelengths and ending at 8 914.597 Å (3.9501 in logarithmic wavelength) which is the minimum from maximal wavelengths. The selection gives us a wavelength grid of 3659 logarithmically-spaced wavelengths (each spectrum is a real vector of $\mathbb{R}^{3659}$).

Given the selected grid of wavelengths, we will lose some SDSS spectra because not all of them have all measurements in the range. Figure 4.2 shows the cumulative histogram of how many spectra we will keep for cuts in different wavelengths. We see that significant drops are behind the selected minimal and maximal wavelengths that mean we will keep most of the spectra.

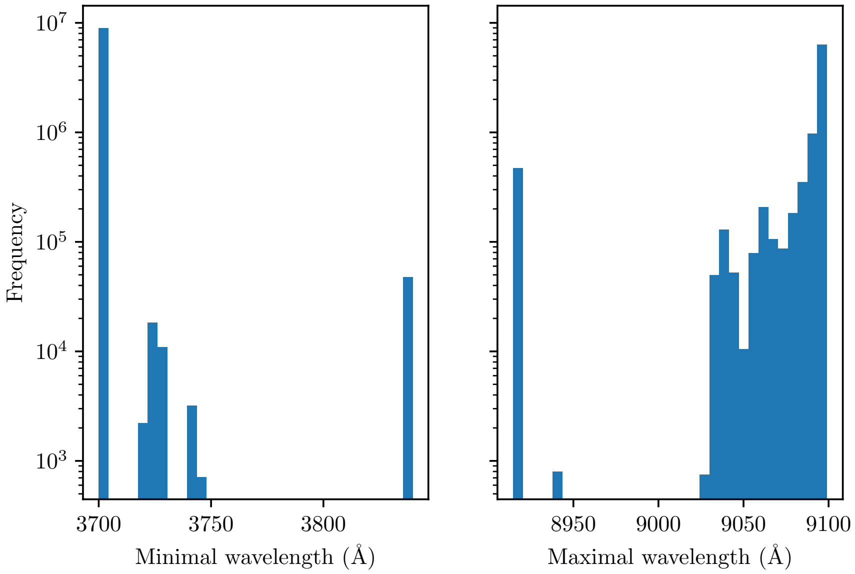


Figure 4.1: Two plots showing histograms of minimal and maximal wavelengths of all LAMOST spectra. The maximal wavelength from all the minimal wavelengths is 3 839.7244 Å. Cutting in a lower wavelength would mean a loss of almost 100 000 spectra. The situation is very similar for maximal wavelengths, where the minimum is 8 914.597 Å. Therefore, the most suitable range of wavelength is to choose these two wavelengths as starting and ending points.

Precisely, the cut will drop 34 487 spectra from our source dataset including 1 949 spectra of QSO. Therefore, the source dataset has 4 816 713 spectra with 627 508 spectra of QSOs that can enter learning of a neural network.

The original sizes of data are unnecessary for experimenting with deep domain adaptation on astronomical spectra. We store each spectrum as a vector of 3 659 single-precision floating-point number (4 bytes). The storage setting gives that the SDSS source dataset has about 70.5 GB and the LAMOST target dataset 132.1 GB. Data of such size usually cannot fit into memory, and access to a disk significantly slows learning on a GPU.

Therefore, we have subsampled the data to the size of ImageNet [51]. We believe that the size of ImageNet is reasonable because ImageNet is the dataset that enables the superiority of deep neural network in computer vision. ImageNet has 1 million training examples, 50 thousand validation examples and 100 thousand testing examples. Accordingly, we randomly subsampled of source and target datasets obtaining training sets of size 1 million and validation sets of size 50 thousand. At the same time, the rest of the data would

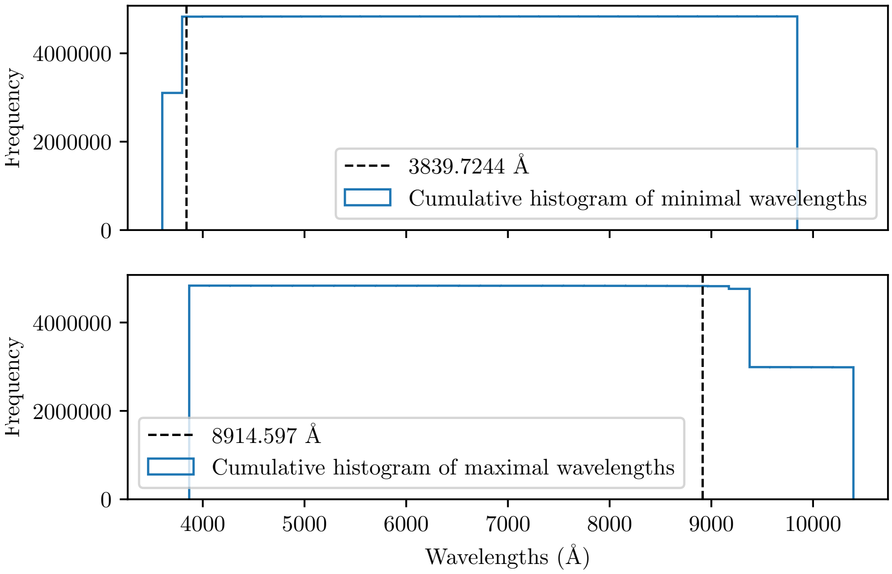


Figure 4.2: The selected wavelength range will inevitably cause a loss of some SDSS DR14 spectra. This figure shows cumulative histograms of the number of spectra and its dependence on minimal and maximal wavelengths. We see that both cuts are before the big drop is the count of spectra.

serve as testing sets. We summarise sizes of datasets with the corresponding number of QSOs in Table 4.1. Table 4.1 shows a significant class imbalance in the LAMOST DR5, where QSOs are very rare (less than 0.4%).

The last step of data preparation is min-max scaling of each spectrum into the $[-1;1]$ range:

$$\mathbf{x}_i = 2\frac{\mathbf{x}_i - \min(\mathbf{x}_i)}{\max(\mathbf{x}_i) - \min(\mathbf{x}_i)} - 1, \tag{4.1}$$

where $\mathbf{x}_i \in \mathbb{R}^{3659}$ is a spectrum as defined in Chapter 3 and functions $\min(\cdot)$ and $\max(\cdot)$ returns the smallest and the largest element of a given vector, respectively.

There are two benefits of the min-max scaling. Firstly, the data will be in a suitable range for the learning of neural networks which will stabilise learning. Secondly, the scaling will remove intensity properties of spectra, leaving us only with the spectrum shape which we want to use for identification of QSOs.

Name	Number of QSO spectra	Total spectra
SDSS DR14	629 457 (12.98%)	4 851 200
usable SDSS DR14	627 508 (13.03%)	4 816 713
SDSS training set	130 904 (13.09%)	1 000 000
SDSS validation set	6 552 (13.10%)	50 000
LAMOST DR5 v3	31 755 (0.35%)	9 026 365
LAMOST training set	3 517 (0.35%)	1 000 000
LAMOST validation set	190 (0.38%)	50 000

Table 4.1: Summary table of sizes of the source and target dataset together with train and validation splits. The validation splits serve for models comparison and hyperparameter optimisation. The second row shows the number of spectra after the cut into a unified range of wavelengths. The table shows the imbalance of the LAMOST target data that contains only a tiny amount of identified QSOs.

4.2 Dimensionality Reduction

In this section, we investigate the structure of joint data space of source and target datasets with three dimensionality reduction methods: *principal component analysis*, *t-Distributed Stochastic Neighbor Embedding* and *Uniform Manifold Approximation and Projection* (UMAP). We would like to get an idea of how well are the source and target data mixed, if there are some separate clusters or the data are rather continuos.

To avoid visualisation overwhelmed with data points, we sampled 2 500 spectra from the source training set and 2 500 spectra from the target training set. The spectra are min-max scaled, not standardised because the relation between features is meaningful, and we do not want to suppress the relationship.

Principal component analysis (PCA) is a simple linear machine learning algorithm that is used either for visualisation or feature extraction by dimensionality reduction. PCA learns a representation whose features are uncorrelated with each other and selects features with the largest variation. [50] We show the visualisation obtained with PCA in Figure 4.3. The plot shows that source data tent to concentrate in the middle while the target data are on the edges. However, no regions are containing only the source or target data. Moreover, in the middle extending to the right, there is a kind of line component.

t-Distributed Stochastic Neighbor Embedding (t-SNE) [52, 53] is a popular method for visualisation of high-dimensional data. The t-SNE method is nonlinear, iterative and performs different transformations of different regions. A tunable hyperparameter of t-SNE is *perplexity* which is a guess about the num-

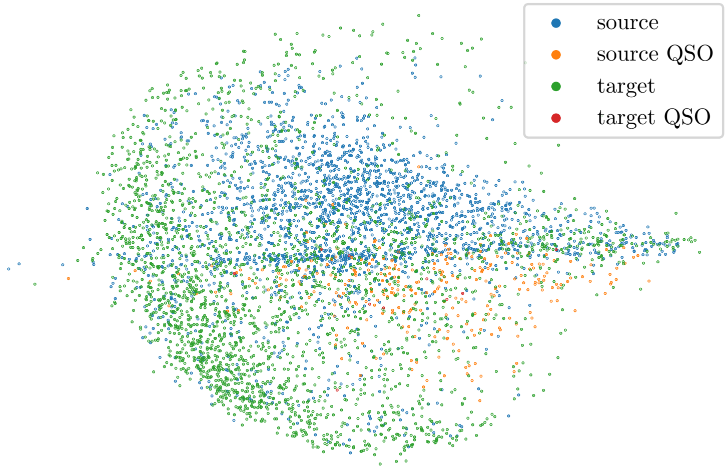


Figure 4.3: The first two principal components of 2 500 source and 2 500 target data points. The projection shows that source data concentrate more in the middle while the target data seem to cluster on the edges.

ber of neighbours of a data point. Typically, the optimal value is between 5 and 50.

We reduce dimensionality for perplexities from $\{5, 10, 30, 50, 100\}$. The best result was for the value 50, and the result is shown in Figure 4.4. The t-SNE embedding has a similar structure as Figure 4.3 of PCA. There are mostly source data in the centre and target data around it. However, the separation between the centre and edges seems to be larger. Still, there is the line component extending downward this time.

t-SNE is often used in the papers presenting a deep domain adaptation methods to show how feature extracted from a higher layer in an adapted network are better mix when a domain adaptation method is employed. But, when a network is not adapted the source and target data can be easily visually separated in a t-SNE visualisation.

Uniform Manifold Approximation and Projection (UMAP) [54] is a non-linear dimensionality reduction algorithm based on manifold learning and ideas from topological data analysis. It achieves visualisations similar to t-SNE, but it is significantly faster. Visualisation with UMAP is displayed in Figure 4.5 and has a very similar structure to t-SNE embedding in Figure 4.4.

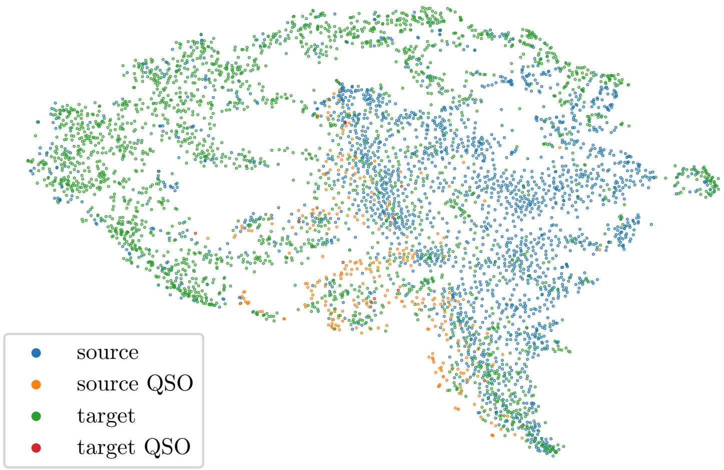


Figure 4.4: Embedding of t-SNE of the same data as in the reduction with PCA shows the very similar result as PCA. However, there is more notable central sort of line component extending downward.

4.3 Baseline: Results without Deep Domain Adaptation

Now, we are ready for training of neural networks. However, before we dive into deep domain adaptation, we will train a classical convolutional network which will serve as a baseline to which we can compare results of networks augmented for deep domain adaptation.

As the baseline, we choose LeNet-5 [55] convolutional neural network, which was initially used to recognise handwritten digits of MNIST [55]. We have chosen the architecture of LeNet-5 because it is the simplest model used in the DANN paper [20]. Thus, we will not need to create our architecture for the DANN experiment. However, the network is designed for processing of two-dimensional images while a spectrum is a one-dimensional image. Therefore, we have to substitute the two-dimensional convolutions with one-dimensional convolutions. Moreover, we increased the kernel size and stride of pooling layers from 2 to 16 so that the output of the convolutional layers is reasonably big. If we left the original pooling layers, the input to the first fully connected layer would be of size 43 872 in comparison to the original input size 768 for the MNIST dataset. The kernel size and stride of 16 will reduce the input size of our network to 672. Figure 4.6a displays the final

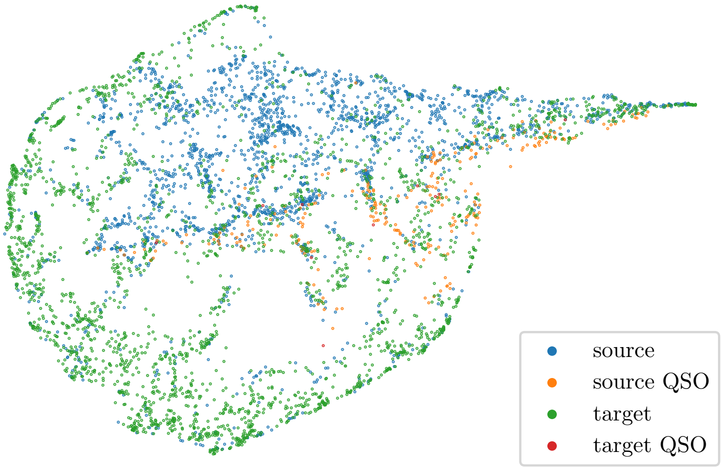


Figure 4.5: UMAP projection to two-dimensions confirms the previous visualisation with PCA and t-SNE. The source data tend to concentrate in the middle while the target data are mostly out of the centre, and there is the line component extending away from the centre.

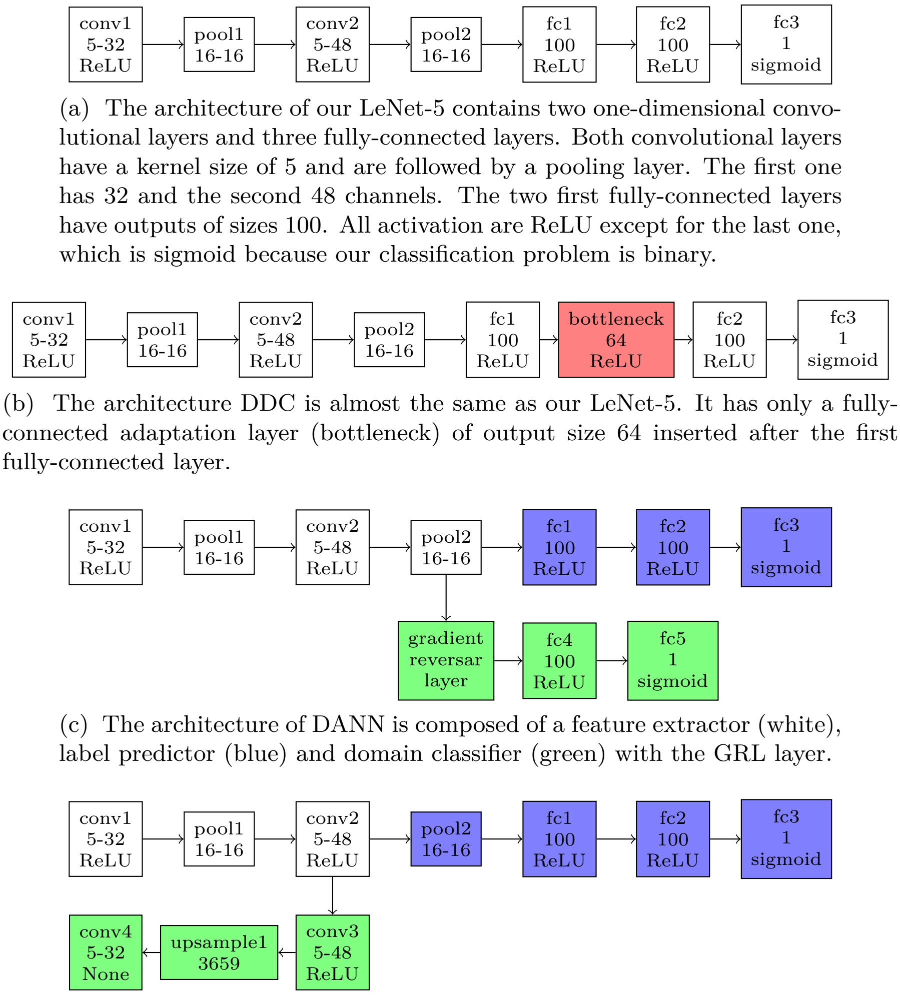
architecture, which we implemented in PyTorch [56] like all other models.

Donahue et al. showed in the *Deep Convolutional Activation Feature* (DeCAF) paper [21] that features extracted from a deep CNN can be repurposed to novel tasks if the network was trained on a large fixed set in a fully supervised fashion, which means that it can be used for domain adaptation on its own. Therefore, we can expect that our CNN trained on the SDSS will be able to find features beneficial for domain adaptation. Still, the DeCAF paper shows that a deep CNN cannot remove domain bias completely. Therefore, there is a space for improvement with deep domain adaptation.

We trained our augmented LeNet-5 on batches of size 64 for 20 epochs with the Adam optimiser [57] in its default setting and used the *binary cross entropy loss* defined as:

$$BCE(\theta) = -\frac{1}{M}\sum_{i}^{M}[y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i)], \tag{4.2}$$

where θ are parameters of a model, M is the batch size, $y_i \in \{0, 1\}$ is the true label, and $\hat{y}_i \in [0, 1]$ are the model predictions of the ith example in the batch. Furthermore, we initialised the weights and biases following



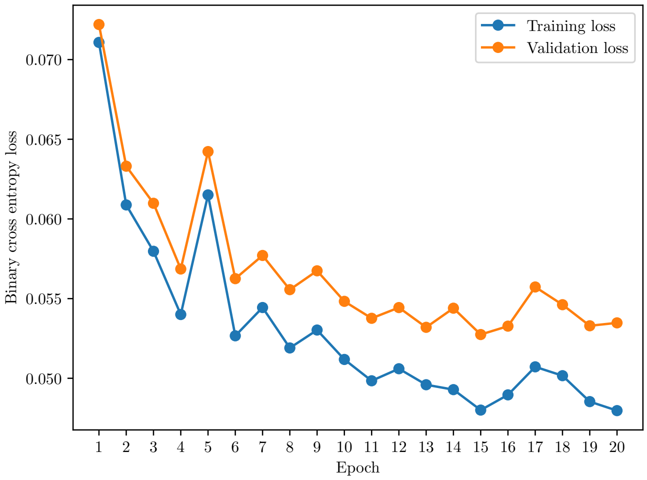
(a) The architecture of our LeNet-5 contains two one-dimensional convolutional layers and three fully-connected layers. Both convolutional layers have a kernel size of 5 and are followed by a pooling layer. The first one has 32 and the second 48 channels. The two first fully-connected layers have outputs of sizes 100. All activation are ReLU except for the last one, which is sigmoid because our classification problem is binary.

(b) The architecture DDC is almost the same as our LeNet-5. It has only a fully-connected adaptation layer (bottleneck) of output size 64 inserted after the first fully-connected layer.

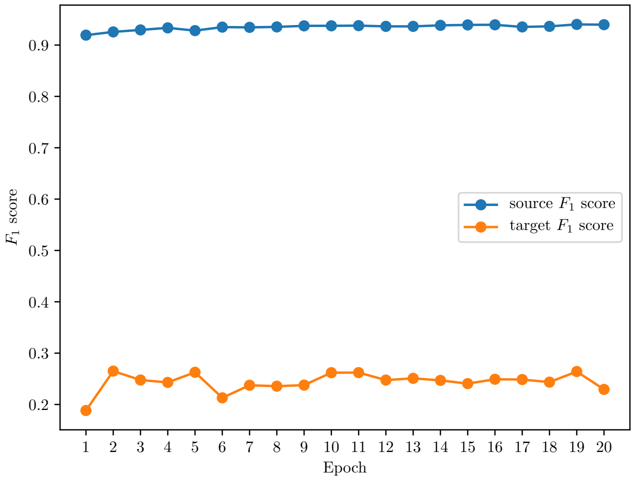
(c) The architecture of DANN is composed of a feature extractor (white), label predictor (blue) and domain classifier (green) with the GRL layer.

(d) The architecture of DRCN contains an encoder (white), decoder (green) with an upsampling layer and classifier (blue).

Figure 4.6: Diagrams of all architectures used in our experiments. Note that the architectures are designed in such way that the classificator is the same for all models (minor exception is the adaptation layer in DDC).



(a) The training and validation losses are similar proving no overfitting.



(b) The source F_1 score is gradually improving while the target F_1 score suffers.

Figure 4.7: Our LeNet-5 is properly learning on the source data. However, it is unable to transfer knowledge to the target domain.

Xavier initialisation [58]. That is weights of our neural network are sampled from uniform distribution $\mathcal{U}$:

$$\mathcal{U}\left(-\frac{\sqrt{6}}{\sqrt{in+out}}, \frac{\sqrt{6}}{\sqrt{in+out}}\right), \tag{4.3}$$

where *in* is the number of input units of a layer and *out* is the number of output units and biases are set to zero.

Before the analysis of results of LeNet-5, we define performance metrics commonly used for imbalanced datasets: *recall* r, *precision* p and F_1 *score* is the harmonic mean between *precision* and *recall*:

$$r = \frac{TP}{(TP + FN)}, \tag{4.4}$$

$$p = \frac{TP}{(TP + FP)}, \tag{4.5}$$

$$F_1 = \frac{2}{r^{-1} + p^{-1}} = 2\frac{rp}{r + p}, \tag{4.6}$$

where r is recall, p is precision, TP is the number of correctly classified QSOs, FN is the number of QSOs incorrectly classified as non-QSOs, and FP is the number of non-QSOs classified as QSOs. When precision and recall are perfect, F_1 score reaches its best value one, and at worst can be zero.

Figure 4.7 displays the training progress of our LeNet-5. We see that the network has converged and is not overfitting because the gap between the training and validation loss is small. The source F_1 score in gradually improving meanwhile, the target F_1 score suffers.

The results[1] of our baseline are quite good on the source domain because the source F_1 score is 0.9397 (the source recall is 95.82% and the source precision is 92.19%). However, the target F_1 score is 0.2294 (the target precision is 13.48% and the target recall 76.84%). That is an inferior result for the target data in the light of source performance. We see that there probably is a considerable domain discrepancy, and, therefore, an opportunity for domain adaptation.

4.4 Experiments with Deep Domain Adaptation

We set the baseline result with classical CNN. Now, we apply four deep domain adaptation methods to the same data to analyse if astronomical spectroscopy can benefit from domain adaptation based on neural networks.

[1] Note that all metrics (F_1 score, precision, recall and confusion matrices) were computed on either the source or target validation sets in our experiments.

Predicted class	Actual class	
	QSO	non-QSO
QSO	6 278	532
non-QSO	274	42 916

(a) Confusion matrix for the source domain.

Predicted class	Actual class	
	QSO	non-QSO
QSO	146	937
non-QSO	44	48 873

(b) Confusion matrix for the target domain.

Table 4.2: Confusion matrices of the baseline model for the source and target validation sets. We see the enormous error on the target domain where the model predicts 937 non-QSOs as QSOs and cannot identify 44 QSOs.

We start with two discrepancy-based approaches, which are DDC and Deep CORAL. Then, we continue with DANN, which is an adversarial-based domain adaptation method and with reconstruction-based DRCN. We conclude with the evaluation and comparison of results.

4.4.1 DDC: Deep Domain Confusion

First deep domain adaptation method is *Deep Domain Confusion* (DDC). DDC reduces the domain discrepancy (maximises domain confusion) by extending classification loss of a neural network with the MMD loss (see Equation 3.4). The MMD loss is enforced on the *adaptation layer* that serves as an information bottleneck for domain confusion. More details on DDC are in Subsection 3.3.1.

To select the size and placement of the adaptation layer, we followed the same procedure as in the DDC paper [22]. Firstly, we took the LeNet-5 trained in previous Section 4.3 and extracted features from the first and second fully-connected layer (the last fully-connected layer has a trivial width) for all validation examples. Then, we computed MMD between the source and target data at each layer with the extracted features. The intuition is to place the adaptation layer after a layer with the smallest MMD because low MMD means more domain invariant features. We measured that the MMD at the first fully-connected layer is 50.70, while MMD at the second fully-connected layer is 53.33. Therefore, we will place the adaptation layer after the first fully-connected layer. Secondly, we have to optimise the width of the adaptation layer. Therefore, we trained the LeNet-5 with the adaptation layer of sizes from $\{4, 8, 16, 32, 64\}$, excluding the low width of 2 and not exceeding the output size of the previous layer. The stepping is the power of two as in the original paper. Figure 4.8 plots the resulting MMD values for different setting and shows that the width 64 is the best.

The final architecture of our DDC network is in Figure 4.6b. It is the baseline CNN with the adaptation layer of width 64 after the first fully-connected layer.

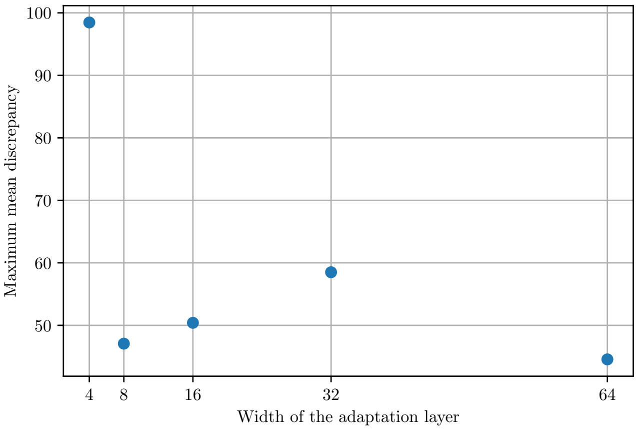


Figure 4.8: We employed the same methodology to optimise hyperparametes as the original paper of DDC. Therefore, we compute the MMD for different sizes of adaptation layers. The scatter plot show that the size of 64 is the best closely followed by the size of 8.

Predicted class	Actual class	
	QSO	non-QSO
QSO	148	1 138
non-QSO	42	48 672

Table 4.3: Confusion matrix of DDC for target data

Furthermore, we set the trade-off parameter λ between the binary cross entropy loss and the MMD loss to 0.25 as in the DDC paper and trained the network in the same way as described in the previous Section 4.3 (Xavier initialisation, Adam optimiser, batch size 64 and 20 epochs).

The training of DDC proceeded similarly to the baseline. Moreover, we see that MMD is also minimalised in the bottom plot of Figure 4.9. On the other side, if we train the network without enforcing MMD loss ($\lambda = 0$), then the MMD is growing, as shown in the top plot of Figure 4.9.

Although, training run as expected, DDC achieved an unfortunate result in comparison to baseline. The F_1 score on the source data is 0.9354 and 0.2005 on the target data that is lower than the baseline in both cases. Precision on target is 11.51%, and the only improvement is the recall of value 77.89%, but the decrease in precision is probably caused by that.

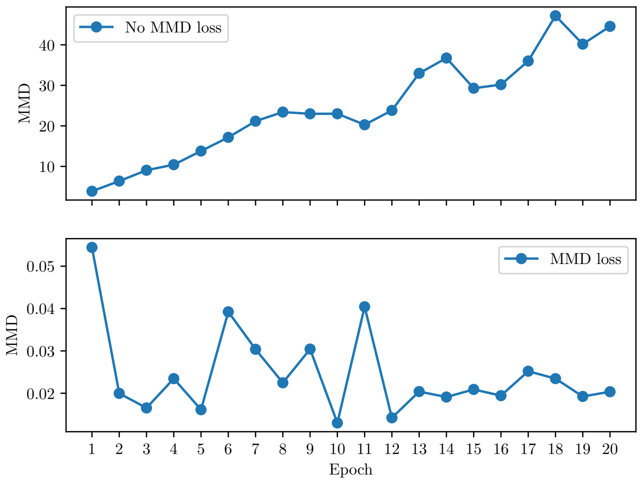


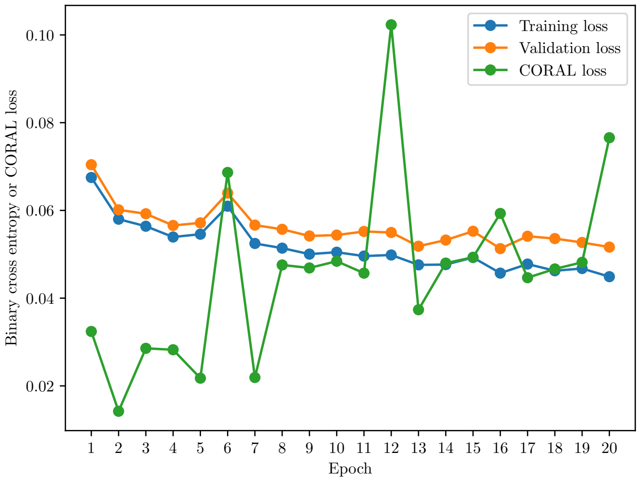
Figure 4.9: This plot illustrates that enforcing the MMD loss has the expected effect. On the other hand, without the MMD loss, the MMD between the source and target domain gradually grows.

4.4.2 Deep CORAL: Deep Correlation Alignment

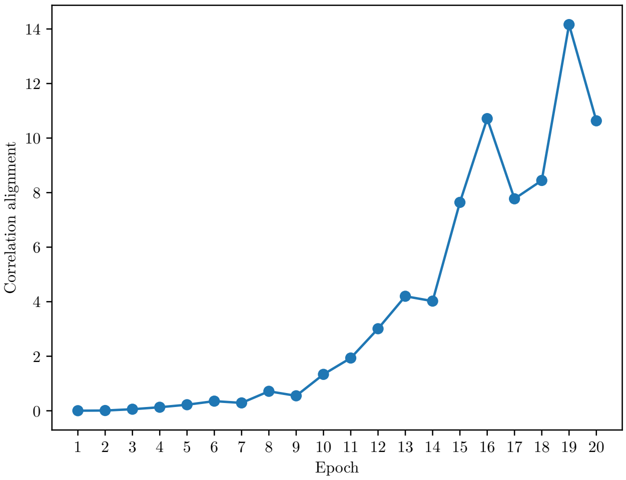
Deep Correlation Alignment (Deep CORAL) is very similar to DDC. DDC aligns means of the source and target distributions with MMD loss while Deep CORAL aligns correlations with CORAL loss (see Equations 3.5 and 3.6). Moreover, Deep CORAL applies the CORAL loss straight to a layer in a network not creating an adaptation layer. We implemented Deep CORAL with inspiration from the original code[2] and followed all the steps described in the corresponding paper [25].

Originally, the architecture underlying Deep CORAL is AlexNet [59]. There the CORAL loss was put on the last layer that has ten output units. However, our neural network has to have one output unit, so applying CORAL loss to it does not make sense. Therefore, we applied the CORAL loss to the second fully-connected layer of our LeNet-5 architecture in Figure 4.6a. Then, we optimised the trade-off between classification and CORAL loss λ from $\{0.5, 0.1, 0.05, 0.01, 0.005, 0.001, 0.0005, 0.0001\}$. The best is $\lambda = 0.0005$, which makes the classification loss and the CORAL loss of similar magnitude as suggested by the paper (see Figure 4.10a). Furthermore, we used the same batch size of value 128 as in original experiments of Deep CORAL. The architecture is initialised in the same way as our baseline and optimised with Adam for 20 epochs.

[2] Available from: `https://github.com/visionlearninggroup/CORAL`



(a) The original paper states that the classification loss and the CORAL loss should be almost the same at the end of the training. We show in this plot that the similarity can be achieved by setting $\lambda = 0.0005$.



(b) The CORAL statistic without enforcing the CORAL loss grows significantly in comparison to the scenario in the plot above when the network is trained with the CORAL loss.

Figure 4.10: Training of Deep Correlation Alignment

Predicted class	Actual class	
	QSO	non-QSO
QSO	147	835
non-QSO	43	48 975

Table 4.4: Confusion matrix of Deep CORAL for target data

Firstly, we trained our Deep CORAL with $\lambda = 0$ to see how the CORAL grows. Figure 4.10b shows the same behaviour of CORAL without enforcing the minimisation of correlation loss as in the original paper. Then, we experimented with the best $\lambda = 0.0005$ and obtained the training progress in Figure 4.10a. However, the results are unsatisfactory, as in the case of DDC.

Source data F_1 score is 0.9396 that is almost the same as baseline. There is a small but insignificant improvement in the target F_1 score, which is of value 0.2509 (the target precision is 14.97% and the target recall is 77.37%). These values are gains, but they are too small, and the model with such small precision is useless for identification of QSOs.

4.4.3 DANN: Domain-Adversarial Neural Network

Domain-Adversarial Neural Network (DANN) is an adversarial-based domain adaptation method. Our DANN architecture is depicted in Figure 4.6c. It consists of a feature extractor, a predictor and a domain classifier that acts adversarially against the feature extractor enforcing domain invariant representation. Further details are in Subsection 3.3.2.

We code and schedule hyperparameters of DANN according to the original implementation[3] and the two papers where DANN was published [20, 60]. Therefore, we implemented the learning rate schedule for SGD with momentum:

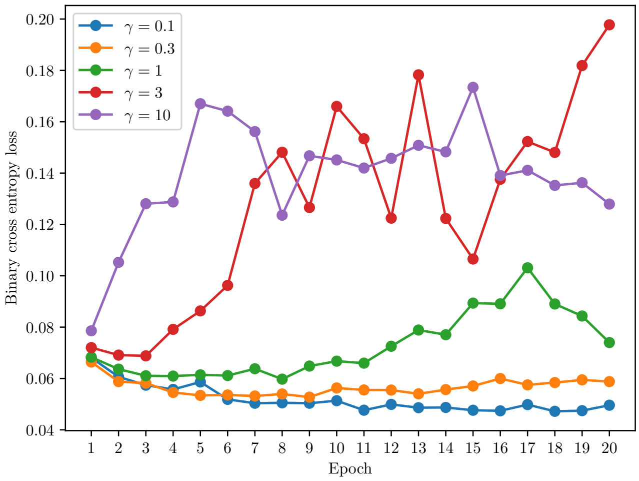
$$\mu_p = \frac{\mu_0}{(1 + \alpha p)^{\beta}}, \tag{4.7}$$

where p is the training progress linearly changing from 0 to 1 in every iteration, initial learning rate $\mu_0 = 0.01$, $\alpha = 10$ and $\beta = 0.75$. Furthermore, we also implemented the domain adaptation parameter λ from Equation 3.17 that starts at 0 and grows to 1 with the schedule:

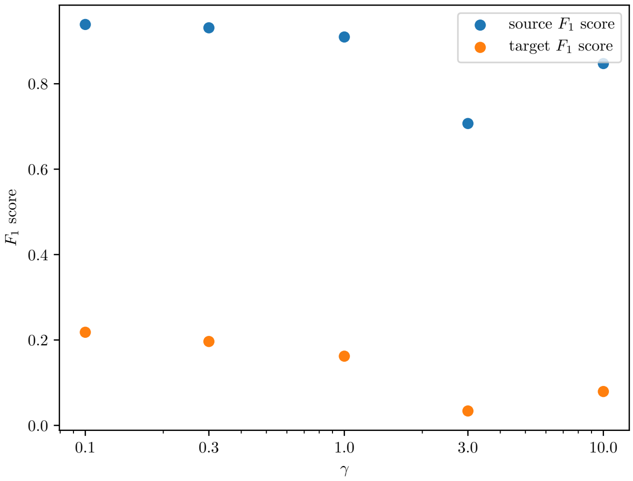
$$\lambda_p = \frac{2}{1 + e^{-\gamma p}} - 1, \tag{4.8}$$

where p is again the training progress and γ that was set to 10 as in the original paper.

[3] Available from: `http://sites.skoltech.ru/compvision/projects/grl/`



(a) For high values of γ DANN diverges.



(b) F_1 score regresses for high values of γ.

Figure 4.11: We could not converge DANN while keeping the F_1 score low even though we did hyperparameter optimisation.

Note that binary cross entropy loss is used for both the classification and domain loss. The optimiser is SGD with the learning rate schedule and momentum 0.9, the network is initialised as our baseline model, and the batch size is 128 where the first half is source domain data and the second half is the target domain data.

Predicted class	Actual class	
	QSO	non-QSO
QSO	142	648
non-QSO	48	49 162

Table 4.5: Confusion matrix of DRCN for target data

Although following the original implementation as closely as possible and doing hyperparameter optimisation of $\gamma \in \{0.1, 0.3, 1, 3, 10\}$ we were not able to get reasonable results with DANN. We infer from Figure 4.11 that if gamma is high, the training will diverge. On the other hand, if gamma is low F_1 score regress.

4.4.4 DRCN: Deep Reconstruction-Classification Network

The last deep domain adaptation model is *Deep Reconstruction-Classification Network* (DRCN) which uses a reconstruction of target data as an auxiliary task. Intuition is that the auxiliary task will enforce the network to capture also the structure of target data space. More detail are provided in Subsection 3.3.3.

We followed the original implementation[4] of DRCN [32]. There is one big difference between the original paper and the official paper that is the order of training loops. The paper states that the network in an epoch should firstly be trained for classification task and then for reconstruction task. The implementation does it the other way around. We choose the working implementation and trained for reconstruction first.

According to the implementation, we used Adam optimiser with batch size 128 for 20 epochs and Xavier initialisation. The trade-off parameter λ was set to 0.5. Figure 4.13 shows that the network is able to learn a good representation that can suppress noise while maintaining important spectral lines.

However, achieving poor results with previous methods, we did not suppose to get a better result now. The source F_1 score is 0.9393 and the target F_1 score 0.2898, which is better than baseline but the development of the F_1 score in Figure 4.12 suggest no significance. Target precision is 17.97% which is the best so far, but target recall 74.74% is the worst.

4.5 Discussion of Experiments

We summarise the results of our baseline and deep domain adaptation methods in Table 4.6. We do not include the performance of DANN because we

[4] Available from: `https://github.com/ghif/drcn`

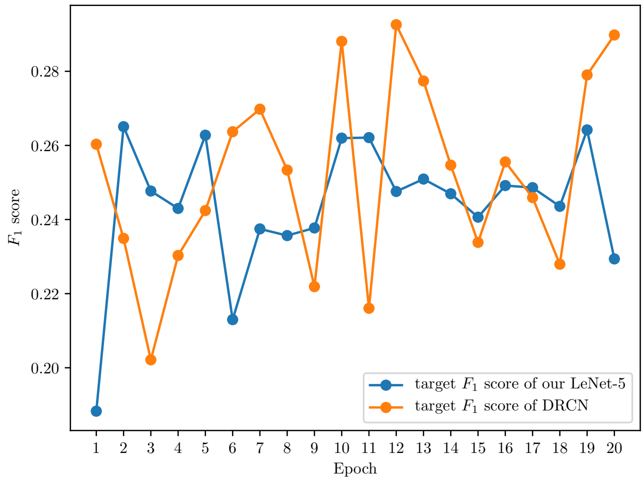


Figure 4.12: Although the final F_1 score of DRCN is higher than the F_1 score of LeNet-5, the full progress shows that the result is not significant.

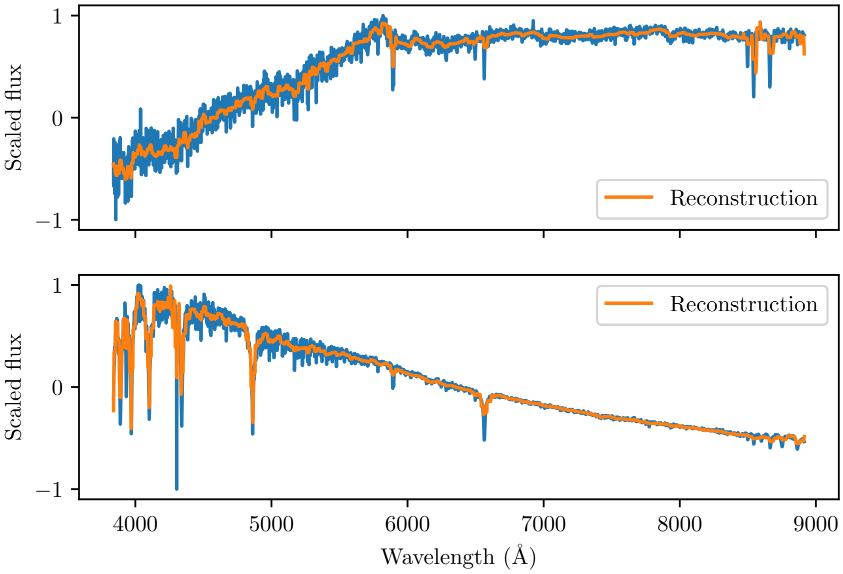


Figure 4.13: The two spectra reconstructed with the convolutional autoencoder in DRCN shows that the training was successful. The autoencoder can reduce noise in the spectra while still keeping track of spectral lines.

Method	Source F_1	Target F_1	Precision (%)	Recall (%)
Baseline	0.9397	0.2294	13.48	76.84
DDC	0.9354	0.2005	11.51	77.89
Deep CORAL	0.9396	0.2509	14.97	77.37
DRCN	0.9393	0.2898	17.97	74.74

Table 4.6: Summary table of results of experiments

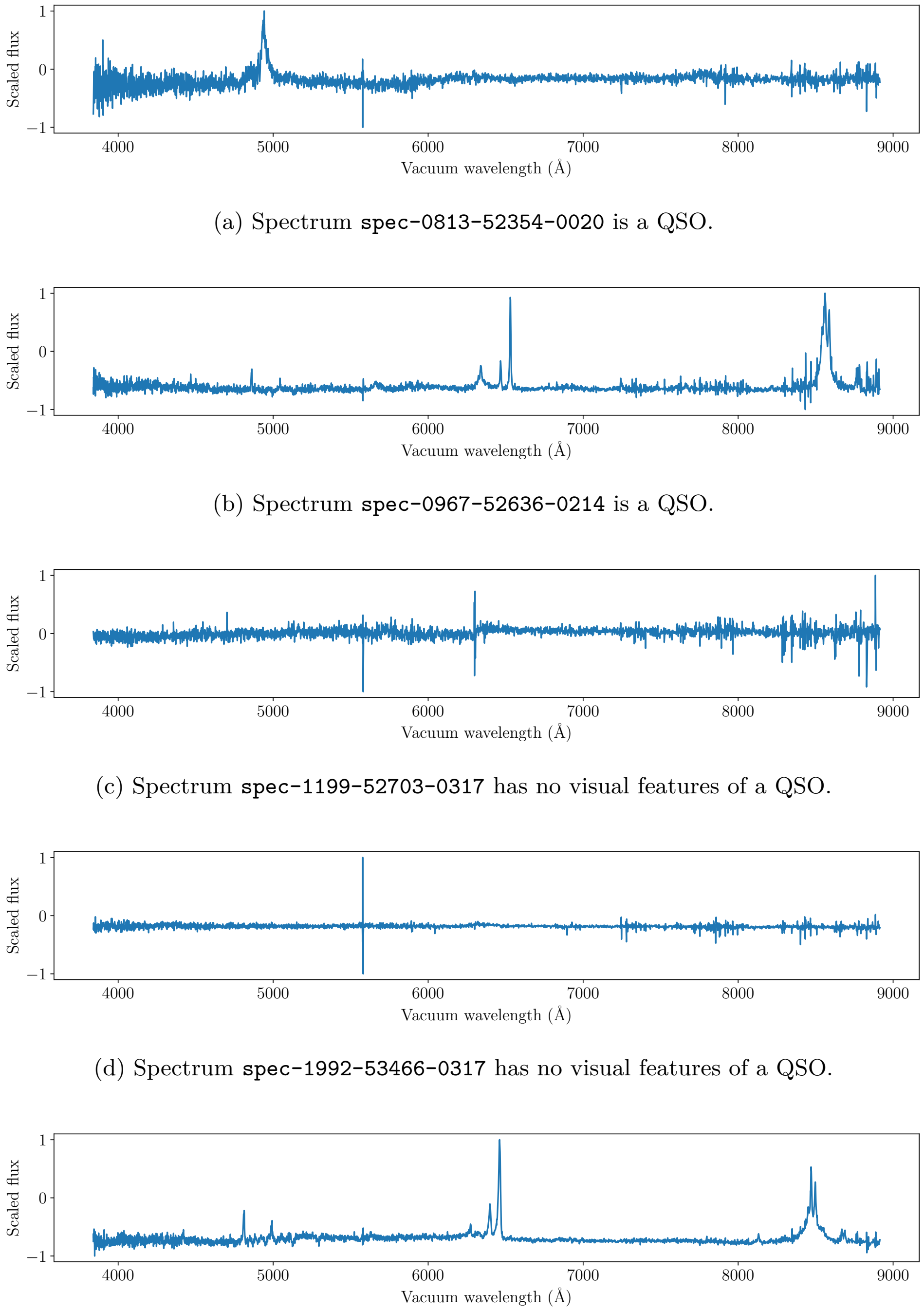
were not able to train it correctly even though we did hyperparameter optimisation. Table 4.6 clearly shows that domain adaptation based on neural networks cannot significantly improve performance in comparison to baseline when applied to astronomical data. We would expect an increase in a metric of at least 5% as in the original paper of the deep domain methods on standard academical datasets.

Domain adaptation did not succeed, although the distributions of the source and target domain are different. We prove the difference in Subsection 2.3.3, where we compared the two surveys. Furthermore, we confirmed the discrepancy with dimensionality reduction techniques. All three PCA, t-SNE and UMAP shows that the data does not occupy a single cluster, but the SDSS spectra concentrate in the centre and LAMOST spectra on the edges (see Figures 4.3, 4.4 and 4.5).

Also, the deep domain adaptation method behaved correctly. DDC kept the MMD between the source and target distributions low, as shown in Figure 4.9. Deep CORAL achieved the same with the CORAL metric for distribution difference (see Figure 4.10), and DRCN learnt the auxiliary reconstruction task that is supposed to support domain adaptation (see Figure 4.13). Moreover, hyperparameter optimisation cannot improve the performance of the deep domain adaptation methods.

All in all, we hypothesise the problem is in the data. Therefore, we visualise the incorrect classification of the baseline (the errors were almost the same for all the methods). In Figure 4.15 and Figure 4.14, we display random spectra from source false positives and source false negatives, respectively. At the same time, we display target false positives and target false negatives in Figure 4.17 and Figure 4.16, respectively. The rest of random spectra is in Appendix A.

We believe the misclassifications are evidence for our conclusion that problem is in our imperfect datasets. The incorrectly classified examples are QSOs not yet identified by a catalogue of the surveys. There are also spectra incorrectly classified as QSOs by the official catalogues. Moreover, there are spectra with artefacts (for example, missing measurements at paricular wavelengths, wrong extraction from CCD chip). However, the original deep domain adaptation methods are trained on well-prepared and clean data. For example, all data in the MNIST [55] or USPS [61] are well-defined digits, and the same applies to the Office dataset [62] commonly used as a domain adaptation benchmark. Such well-formed datasets provide a comfortable environment for basic research. On the other side, they do not resemble the real world or scientific situation. That is a big issue for the application of such method to scientific data. As we have shown, astronomy provides such a volume of data that is impossible to make clean. Therefore, we need either robust machine learning algorithms or automatic procedures that clean data possibly also based on machine learning. However, by cleaning the data, we might lose some interesting objects with strange physical properties. Maybe, the imperfect data problem is the reason why previous applications of domain adaptation in astronomy used active learning (a human expert) after domain adaptation (see Section 3.4).



(a) Spectrum `spec-0813-52354-0020` is a QSO.

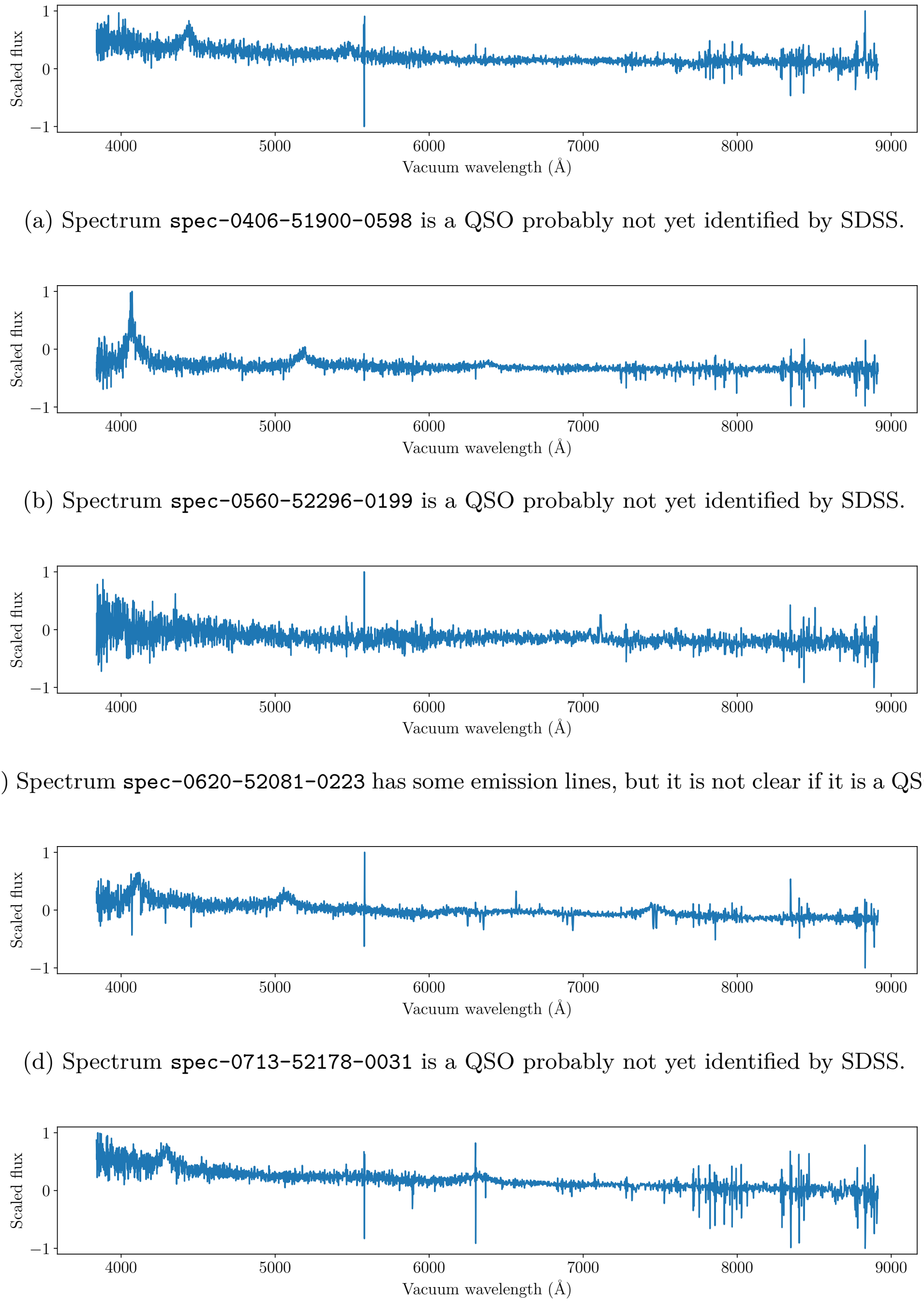
(b) Spectrum `spec-0967-52636-0214` is a QSO.

(c) Spectrum `spec-1199-52703-0317` has no visual features of a QSO.

(d) Spectrum `spec-1992-53466-0317` has no visual features of a QSO.

(e) Spectrum `spec-2656-54484-0409` is a QSO.

Figure 4.14: The first part of sample of source false negatives resembles that the CNN incorrectly classifies some true QSOs. However, there are also spectra not clearly QSOs.



(a) Spectrum `spec-0406-51900-0598` is a QSO probably not yet identified by SDSS.

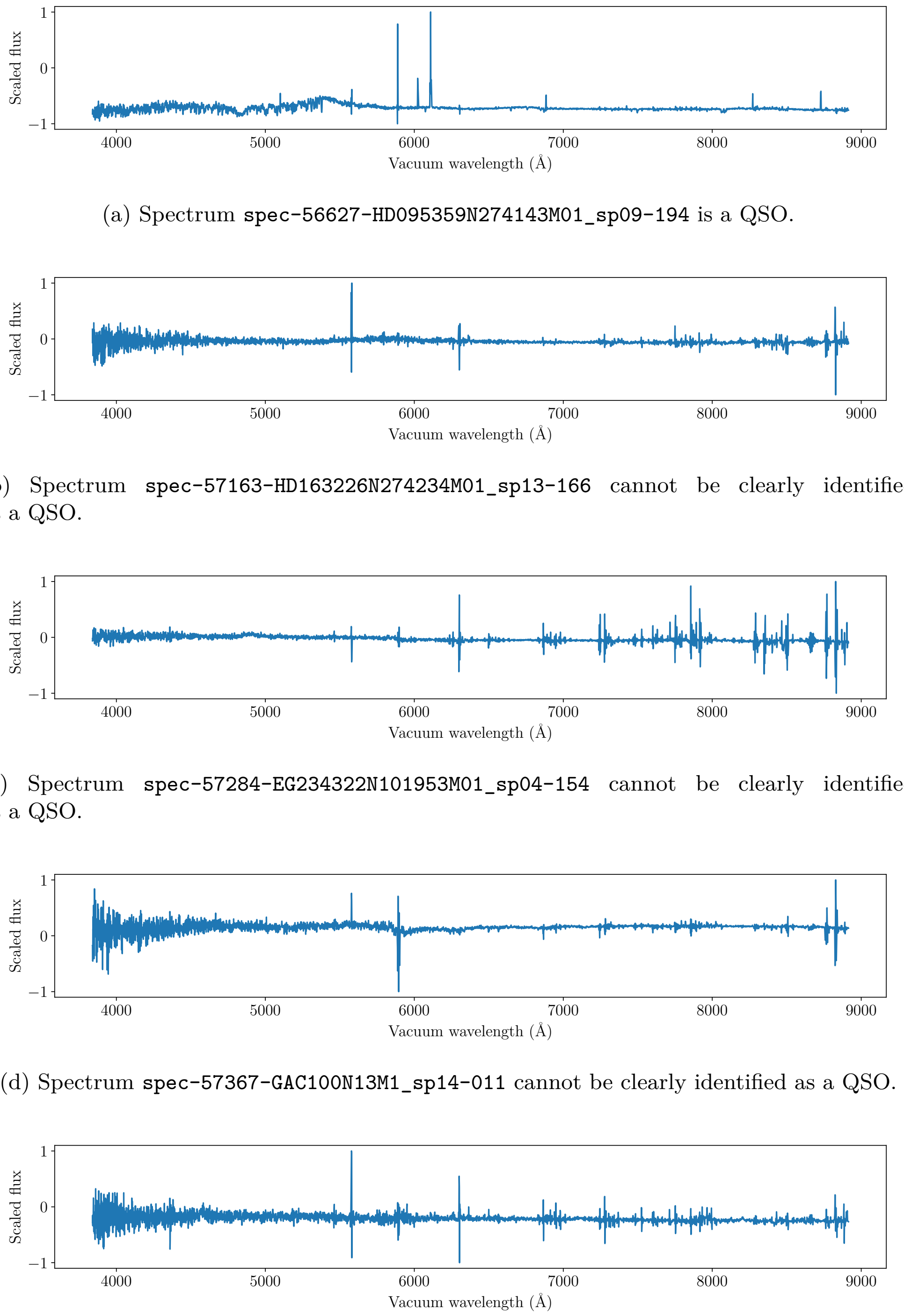
(b) Spectrum `spec-0560-52296-0199` is a QSO probably not yet identified by SDSS.

(c) Spectrum `spec-0620-52081-0223` has some emission lines, but it is not clear if it is a QSO.

(d) Spectrum `spec-0713-52178-0031` is a QSO probably not yet identified by SDSS.

(e) Spectrum `spec-0769-54530-0502` is a QSO probably not yet identified by SDSS.

Figure 4.15: The first of part sample of source false positives shows that they contain a significant amount of QSOs not yet in the catalogues of QSOs of the SDSS.



(a) Spectrum `spec-56627-HD095359N274143M01_sp09-194` is a QSO.

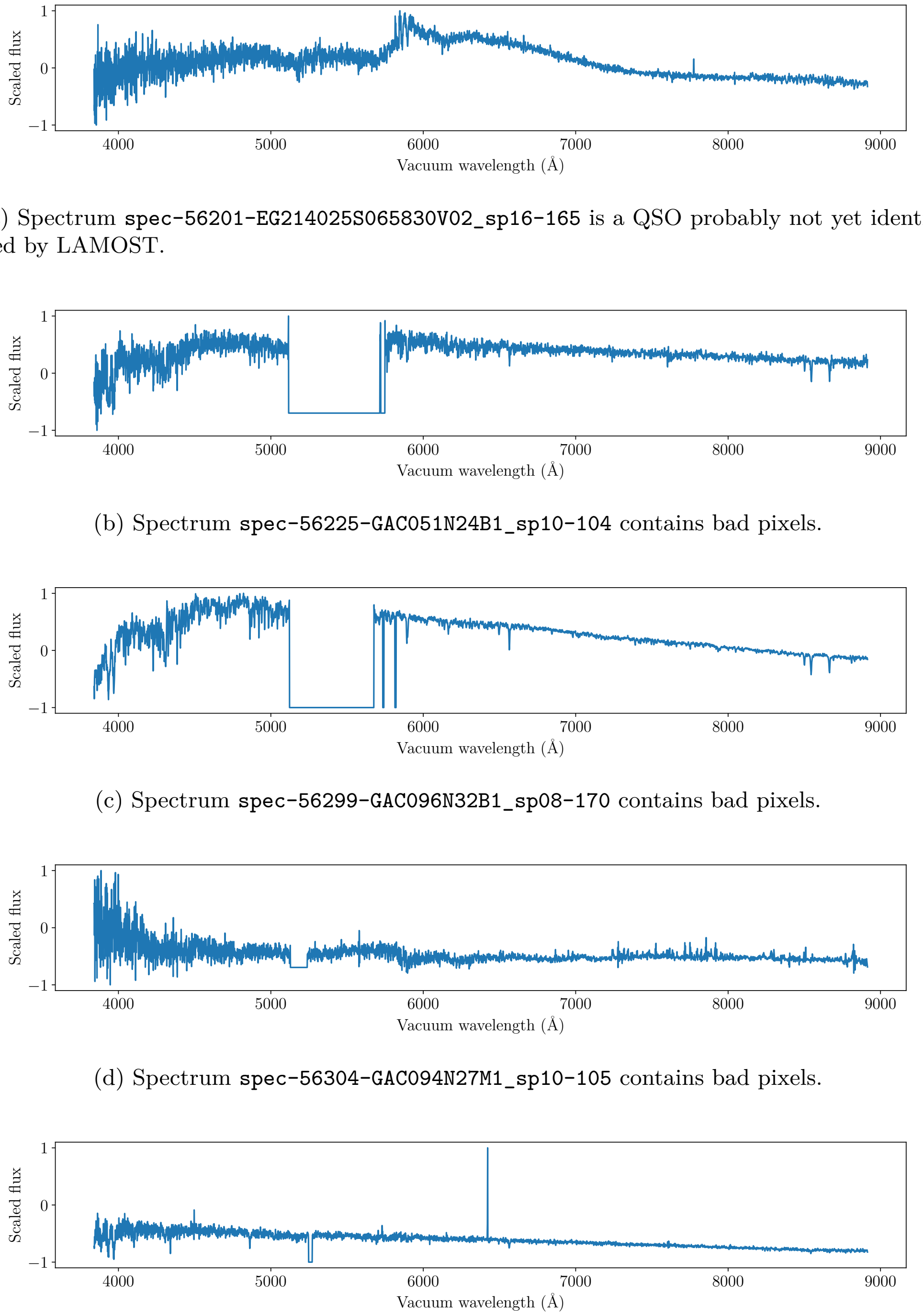
(b) Spectrum `spec-57163-HD163226N274234M01_sp13-166` cannot be clearly identified as a QSO.

(c) Spectrum `spec-57284-EG234322N101953M01_sp04-154` cannot be clearly identified as a QSO.

(d) Spectrum `spec-57367-GAC100N13M1_sp14-011` cannot be clearly identified as a QSO.

(e) Spectrum `spec-57388-EG015238N022953M01_sp11-103` cannot be clearly identified as a QSO.

Figure 4.16: The first part of sample of target false negatives reveals they contain spectra not clearly QSOs from the visual perspective.



(a) Spectrum `spec-56201-EG214025S065830V02_sp16-165` is a QSO probably not yet identified by LAMOST.

(b) Spectrum `spec-56225-GAC051N24B1_sp10-104` contains bad pixels.

(c) Spectrum `spec-56299-GAC096N32B1_sp08-170` contains bad pixels.

(d) Spectrum `spec-56304-GAC094N27M1_sp10-105` contains bad pixels.

(e) Spectrum `spec-56344-GAC088N41V3_sp08-176` contains bad pixels.

Figure 4.17: The first part of sample of target false positives resembles that they contain spectra with bad pixels and a QSO that is not yet identified.

Chapter 5

Conclusion

Our goal was to analyse the impact of domain adaptation based on neural networks in spectroscopic sky surveys. Therefore, we firstly introduced astronomical spectroscopy. Then, we defined quasars because their identification is the task of our domain adaptation setting. To have the domain adaptation setting complete, we introduced the SDSS as the source domain and the LAMOST survey as the target domain. We showed that these two domains are suitable for domain adaptation because of their different instrument and distribution of observations.

Next, we surveyed domain adaptation. Firstly, we put it into the context of the machine and transfer learning, and we formally defined it. Secondly, we distinguished the shallow and deep domain adaptation. The deep domain adaptation category is based on neural networks, thus, further, we focused on it, and we presented its three subcategories: discrepancy-based, adversarial-based and reconstruction-based domain adaptation. We selected four appropriate methods DDC, Deep CORAL, DANN and DRCN. DDC and Deep CORAL are representatives of the discrepancy-based category, DANN of adversarial-based methods and DRCN is based on reconstruction. We explained why we think they are fundamental to experiment with and give the intuition and mathematics behind them.

The thing left was to explore their impact of the quasar identification task from astronomical spectra. Therefore, we briefly introduced data preparation (actually we spend a significant amount of time with data preparation). We followed with dimensionality reduction using PCA, t-SNE and UMAP that demonstrated a visually notable distribution discrepancy between the source and target domain. To have a baseline model to compare the deep domain adaptation methods to, we used the LeNet-5 CNN trained only on the source domain. Its result states that it can be robust classifier on the source domain but is not able to classify well on the target domain as we expected due to distribution difference. We continued with the central experimental part of our thesis. We applied the four deep domain adaptation methods to the data using both the source and target data.

However, the experiments proved that none of DDC, Deep CORAL, DANN and DRCN could achieve a significant improvement on the target data. That happens, even though dimensionality reduction indicate domain discrepancy and, for example, the training of DDC keeps MMD of the two distribution low in comparison when the MMD loss is not enforced. Therefore, we investigated the data in order to analyse what causes the inability of deep domain adaptation methods to improve performance. Our astronomical dataset contains a lot of problematic samples (for example, not identified quasars or bad spectra). On the other hand, the original deep domain adaptation methods are based on well-prepared and clean data.

That is a big issue for the application of such method to scientific data. As we have shown, astronomy provides such a volume of data that is impossible to make clean and labelled. Therefore, we need either robust machine learning algorithms or precise procedures that clean data possibly also based on machine learning. However, by cleaning the data, we might lose some interesting objects with strange physical properties.

5.1 Future Plans

From the research point of view, the result is positive as it opens discovery opportunities because algorithms verified on the well-formed academical dataset like MNIST cannot be directly applied to complex and dirty scientific data. However, we see the future as full of such complex data that we are unable to clean.

Therefore, in future, we plan to make a more detailed statistical analysis of our results and try to make direct consequences for deep domain adaptation methods. The improvements can be either in the form of new robust architectures or novel learning algorithms. Another possibility is to focus on automatic preprocessing algorithms based on machine learning that will be able to filter the data so that they can work with the original deep domain adaptation methods.

Bibliography

[1] Appenzeller, I. *Introduction to Astronomical Spectroscopy.* Cambridge: Cambridge University Press, 2012, ISBN 978-1-107-01579-1.

[2] Trypsteen, M. F. M.; Walker, R. *Spectroscopy for Amateur Astronomers.* Cambridge, United Kigdom: Cambridge University Press, first edition, 2017, ISBN 978-1-107-16618-9.

[3] Cochard, F. *Successfully Starting in Astronomical Spectroscopy: A Practical Guide.* France: EDP Sciences, 2018, ISBN 978-2-7598-2248-5.

[4] Bennett, J.; Donahue, M.; et al. *The Essential Cosmic Perspective.* San Francisco: Addison Wesley, third edition, 2005, ISBN 0-8053-8934-2.

[5] Beckmann, V.; Shrader, C. *Active Galactic Nuclei.* Weinheim, Germany: Wiley-VCH, first edition, 2012, ISBN 978-3-527-31078-1, 382 pp.

[6] York, D. G.; Adelman, J.; et al. The Sloan Digital Sky Survey: Technical Summary. *The Astronomical Journal*, volume 120, no. 3, sep 2000: pp. 1579–1587, doi:10.1086/301513.

[7] Gunn, J. E.; Carr, M.; et al. The Sloan Digital Sky Survey Photometric Camera. *The Astronomical Journal*, volume 116, no. 6, Dec 1998: pp. 3040–3081, doi:10.1086/300645.

[8] Smee, S. A.; Gunn, J. E.; et al. The Multi-Object, Fiber-Fed Spectrographs for the Sloan Digital Sky Survey and the Baryon Oscillation Spectroscopic Survey. *The Astronomical Journal*, volume 146, no. 2, Jul 2013: p. 32, doi:10.1088/0004-6256/146/2/32.

[9] Pâris, Isabelle; Petitjean, Patrick; et al. The Sloan Digital Sky Survey Quasar Catalog: Fourteenth data release. *Astronomy & Astrophysics*, volume 613, 2018, doi:10.1051/0004-6361/201732445.

[10] Cui, X.-Q.; Zhao, Y.-H.; et al. The Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST). *Research in Astronomy and Astrophysics*, volume 12, no. 9, Aug 2012: pp. 1197–1242, doi:10.1088/1674-4527/12/9/003.

[11] Ai, Y. L.; Wu, X.-B.; et al. The Large Sky Area Multi-object Fiber Spectroscopic Telescope Quasar Survey: Quasar Properties from the First Data Release. *The Astronomical Journal*, volume 151, no. 2, Feb 2016, doi:10.3847/0004-6256/151/2/24.

[12] Dong, X. Y.; Wu, X.-B.; et al. The Large Sky Area Multi-Object Fibre Spectroscopic Telescope (LAMOST) Quasar Survey: Quasar Properties from Data Release Two and Three. *The Astrophysical Jounal*, volume 155, no. 5, May 2018, doi:10.3847/1538-3881/aab5ae.

[13] Yao, S.; Wu, X.-B.; et al. The Large Sky Area Multi-object Fiber Spectroscopic Telescope (LAMOST) Quasar Survey: The Fourth and Fifth Data Releases. *The Astrophysical Journal*, volume 240, no. 1, Jan 2019, doi:10.3847/1538-4365/aaef88.

[14] Wang, M.; Deng, W. Deep Visual Domain Adaptation: A Survey. *Neurocomputing*, volume 312, 2018: pp. 135–153, doi:10.1016/j.neucom.2018.05.083.

[15] Csurka, G. *A Comprehensive Survey on Domain Adaptation for Visual Applications*. Cham: Springer International Publishing, 2017, pp. 1–35, doi:10.1007/978-3-319-58347-1_1.

[16] Daume III, H.; Marcu, D. Domain Adaptation for Statistical Classifiers. *Journal of Artificial Intelligence Research*, volume 26, 2006: pp. 101–126.

[17] Torrey, L.; Shavlik, J. Transfer Learning. *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, 2010: pp. 242–264, doi:10.4018/978-1-60566-766-9.ch011.

[18] Pan, S. J.; Yang, Q. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, volume 22, no. 10, Oct 2010: pp. 1345–1359, doi:10.1109/TKDE.2009.191.

[19] Weiss, K.; Khoshgoftaar, T. M.; et al. A Survey of Transfer Learning. *Journal of Big Data*, volume 3, no. 1, May 2016, doi:10.1186/s40537-016-0043-6.

[20] Ganin, Y.; Ustinova, E.; et al. Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research*, volume 17, no. 59, 2016: pp. 1–35. Available from: `http://jmlr.org/papers/v17/15-239.html`

[21] Donahue, J.; Jia, Y.; et al. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In *Proceedings of the 31st International Conference on Machine Learning*, *Proceedings of Machine Learning Research*, volume 32, Bejing, China: PMLR, Jun 2014, pp. 647–655. Available from: `http://proceedings.mlr.press/v32/donahue14.html`

[22] Tzeng, E.; Hoffman, J.; et al. Deep Domain Confusion: Maximizing for Domain Invariance. *CoRR*, volume abs/1412.3474, 2014. Available from: `http://arxiv.org/abs/1412.3474`

[23] Long, M.; Cao, Y.; et al. Learning Transferable Features with Deep Adaptation Networks. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, Lille, France: PMLR, Jul 2015, pp. 97–105. Available from: `http://proceedings.mlr.press/v37/long15.html`

[24] Long, M.; Zhu, H.; et al. Deep Transfer Learning with Joint Adaptation Networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, Sydney, Australia: PMLR, Aug 2017, pp. 2208–2217. Available from: `http://proceedings.mlr.press/v70/long17a.html`

[25] Sun, B.; Saenko, K. Deep CORAL: Correlation Alignment for Deep Domain Adaptation. In *Computer Vision – ECCV 2016 Workshops*, Cham: Springer International Publishing, 2016, pp. 443–450, doi:10.1007/978-3-319-49409-8_35.

[26] Goodfellow, I.; Pouget-Abadie, J.; et al. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27*, Curran Associates, Inc., 2014, pp. 2672–2680. Available from: `http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf`

[27] Tzeng, E.; Hoffman, J.; et al. Adversarial Discriminative Domain Adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017, pp. 2962–2971, doi:10.1109/CVPR.2017.316.

[28] Vincent, P.; Larochelle, H.; et al. Extracting and Composing Robust Features with Denoising Autoencoders. In *Proceedings of the 25th Annual International Conference on Machine Learning*, Omnipress, 2008, pp. 1096–1103.

[29] Masci, J.; Meier, U.; et al. Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction. In *Artificial Neural Networks and Machine Learning – ICANN 2011*, Berlin, Heidelberg: Springer, 2011, pp. 52–59, doi:10.1007/978-3-642-21735-7_7.

[30] Glorot, X.; Bordes, A.; et al. Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach. In *Proceedings of the*

28th International Conference on Machine Learning, New York, NY, USA: ACM, Jun 2011, ISBN 978-1-4503-0619-5, pp. 513–520.

[31] Chen, M.; Xu, Z.; et al. Marginalized Denoising Autoencoders for Domain Adaptation. In *Proceedings of the 29th International Conference on Machine Learning*, New York, NY, USA: Omnipress, Jul 2012, ISBN 978-1-4503-1285-1, pp. 767–774.

[32] Ghifary, M.; Kleijn, W. B.; et al. Deep Reconstruction-Classification Networks for Unsupervised Domain Adaptation. In *Computer Vision – ECCV 2016*, Cham: Springer International Publishing, 2016, ISBN 978-3-319-46493-0, pp. 597–613.

[33] Bousmalis, K.; Trigeorgis, G.; et al. Domain Separation Networks. In *Advances in Neural Information Processing Systems 29*, Curran Associates, Inc., 2016, pp. 343–351. Available from: `http://papers.nips.cc/paper/6255-domain-separation-networks.pdf`

[34] Ho, A. Y. Q.; Ness, M. K.; et al. Label Transfer from APOGEE to LAMOST: Precise Stellar Parameters for 450,000 LAMOST Giants. *The Astrophysical Journal*, volume 836, no. 1, Feb 2017, doi:10.3847/1538-4357/836/1/5.

[35] Ness, M.; Hogg, D. W.; et al. The Cannon: A Data-Driven Approach to Stellar Label Determination. *The Astrophysical Journal*, volume 808, no. 1, Jul 2015, doi:10.1088/0004-637x/808/1/16.

[36] Gupta, K. D.; Pampana, R.; et al. Automated Supernova Ia Classification Using Adaptive Learning Techniques. In *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, Dec 2016, pp. 1–8, doi:10.1109/SSCI.2016.7849951.

[37] Fernando, B.; Habrard, A.; et al. Subspace Alignment for Domain Adaptation. *CoRR*, volume abs/1409.5241, 2014. Available from: `http://arxiv.org/abs/1409.5241`

[38] Gretton, A.; Smola, A.; et al. Covariate Shift by Kernel Mean Matching. In *Dataset shift in machine learning*, The MIT Press, 2009, pp. 131–160, doi:10.7551/mitpress/9780262170055.001.0001.

[39] Settles, B. Active Learning Literature Survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009. Available from: `http://burrsettles.com/pub/settles.activelearning.pdf`

[40] Vilalta, R.; Gupta, K. D.; et al. A General Approach to Domain Adaptation with Applications in Astronomy. *CoRR*, volume abs/1812.08839, 2018. Available from: `http://arxiv.org/abs/1812.08839`

[41] Richards, J. W.; Starr, D. L.; et al. Active Learning to Overcome Sample Selection Bias: Application to Photometric Variable Star Classification. *The Astrophysical Journal*, volume 744, no. 2, Dec 2011, doi:10.1088/0004-637x/744/2/192.

[42] Shimodaira, H. Improving Predictive Inference under Covariate Shift by Weighting the Log-Likelihood Function. *Journal of Statistical Planning and Inference*, volume 90, no. 2, 2000: pp. 227–244, ISSN 0378-3758, doi:10.1016/S0378-3758(00)00115-4.

[43] Heckman, J. J. Sample Selection Bias as a Specification Error. *Econometrica*, volume 47, no. 1, 1979: pp. 153–161. Available from: `http://www.jstor.org/stable/1912352`

[44] Blum, A.; Mitchell, T. Combining Labeled and Unlabeled Data with Co-Training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, 1998, pp. 92–100.

[45] Ackermann, S.; Schawinski, K.; et al. Using Transfer Learning to Detect Galaxy Mergers. *Monthly Notices of the Royal Astronomical Society*, volume 479, no. 1, May 2018: pp. 415–425, doi:10.1093/mnras/sty1398.

[46] Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017, pp. 1800–1807, doi:10.1109/CVPR.2017.195.

[47] Lintott, C.; Schawinski, K.; et al. Galaxy Zoo 1: Data Release of Morphological Classifications for Nearly 900 000 Galaxies. *Monthly Notices of the Royal Astronomical Society*, volume 410, no. 1, Dec 2010: pp. 166–178, doi:10.1111/j.1365-2966.2010.17432.x.

[48] Belhaj, M.; Protopapas, P.; et al. Deep Variational Transfer: Transfer Learning through Semi-supervised Deep Generative Models. *CoRR*, volume abs/1812.03123, 2018. Available from: `http://arxiv.org/abs/1812.03123`

[49] Kingma, D. P.; Welling, M. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations 2014*, Apr 2014. Available from: `http://arxiv.org/abs/1312.6114`

[50] Goodfellow, I.; Bengio, Y.; et al. *Deep Learning*. MIT Press, 2016. Available from: `http://www.deeplearningbook.org`

[51] Russakovsky, O.; Deng, J.; et al. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, volume 115, no. 3, Dec 2015: pp. 211–252, doi:10.1007/s11263-015-0816-y.

[52] van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research*, volume 9, Nov 2008: pp. 2579–2605. Available from: `http://www.jmlr.org/papers/v9/vandermaaten08a.html`

[53] Wattenberg, M.; Viégas, F.; et al. How to Use t-SNE Effectively. *Distill*, 2016, doi:10.23915/distill.00002.

[54] McInnes, L.; Healy, J.; et al. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *CoRR*, volume abs/1802.03426, Feb 2018. Available from: `http://arxiv.org/abs/1802.03426`

[55] Lecun, Y.; Bottou, L.; et al. Gradient-based Learning Applied to Document Recognition. *Proceedings of the IEEE*, volume 86, no. 11, Nov 1998: pp. 2278–2324, doi:10.1109/5.726791.

[56] Paszke, A.; Gross, S.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., 2019, pp. 8024–8035. Available from: `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`

[57] Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. *CoRR*, volume abs/1412.6980, 2014. Available from: `http://arxiv.org/abs/1412.6980`

[58] Glorot, X.; Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, *Proceedings of Machine Learning Research*, volume 9, Sardinia, Italy: PMLR, 13–15 May 2010, pp. 249–256. Available from: `http://proceedings.mlr.press/v9/glorot10a.html`

[59] Krizhevsky, A.; Sutskever, I.; et al. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*, Curran Associates, Inc., 2012, pp. 1097–1105. Available from: `http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf`

[60] Ganin, Y.; Lempitsky, V. Unsupervised Domain Adaptation by Backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*, *Proceedings of Machine Learning Research*, volume 37, Lille, France: PMLR, Jul 2015, pp. 1180–1189. Available from: `http://proceedings.mlr.press/v37/ganin15.html`

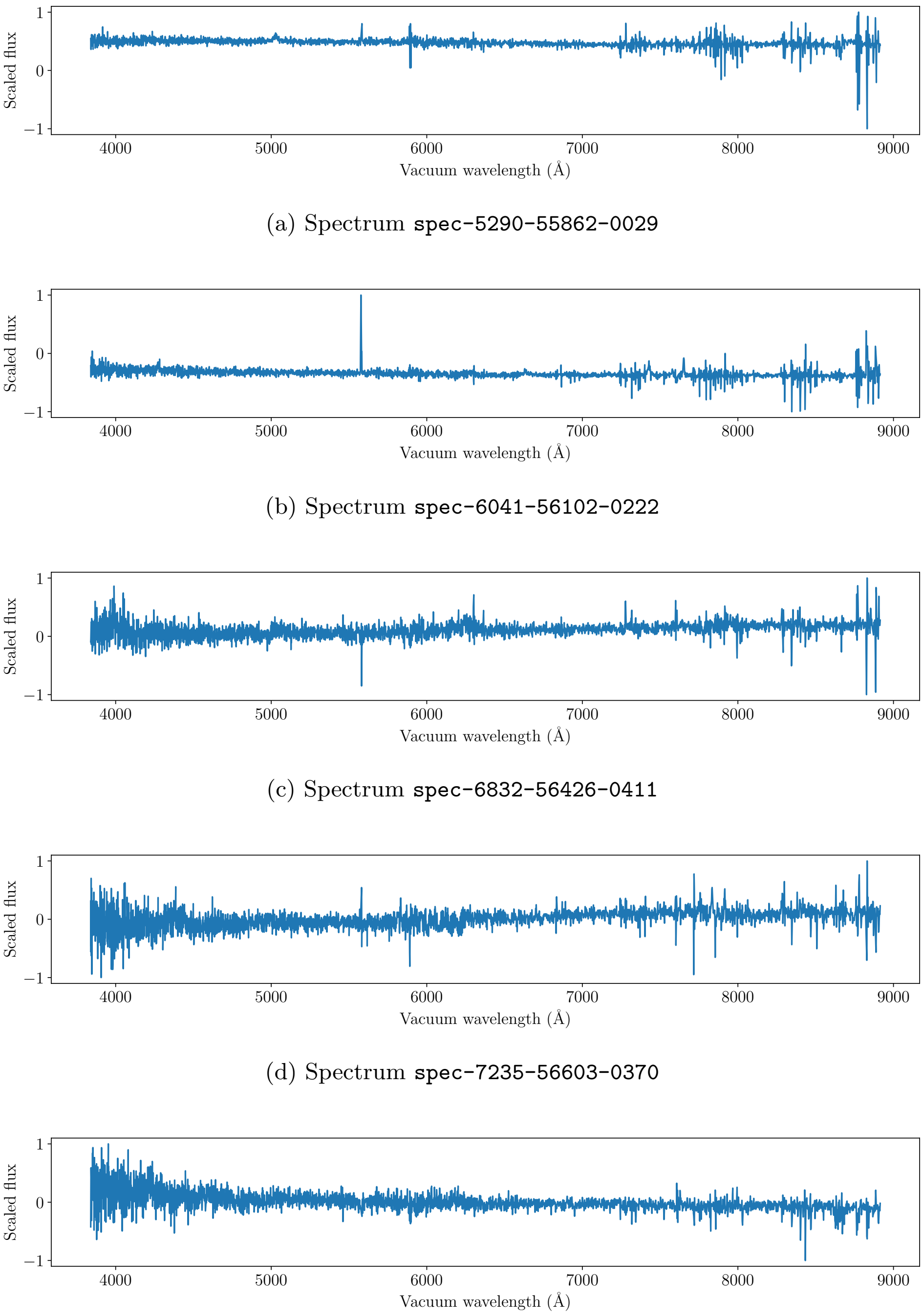
[61] Hull, J. J. A Database for Handwritten Text Recognition Research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 16, no. 5, May 1994: pp. 550–554, doi:10.1109/34.291440.

[62] Saenko, K.; Kulis, B.; et al. Adapting Visual Category Models to New Domains. In *Computer Vision – ECCV 2010*, Berlin, Heidelberg: Springer, 2010, pp. 213–226, doi:10.1007/978-3-642-15561-1_16.

Appendix A

Misclassifications

Random spectra for the source and target false positives and false negatives.



(a) Spectrum `spec-5290-55862-0029`

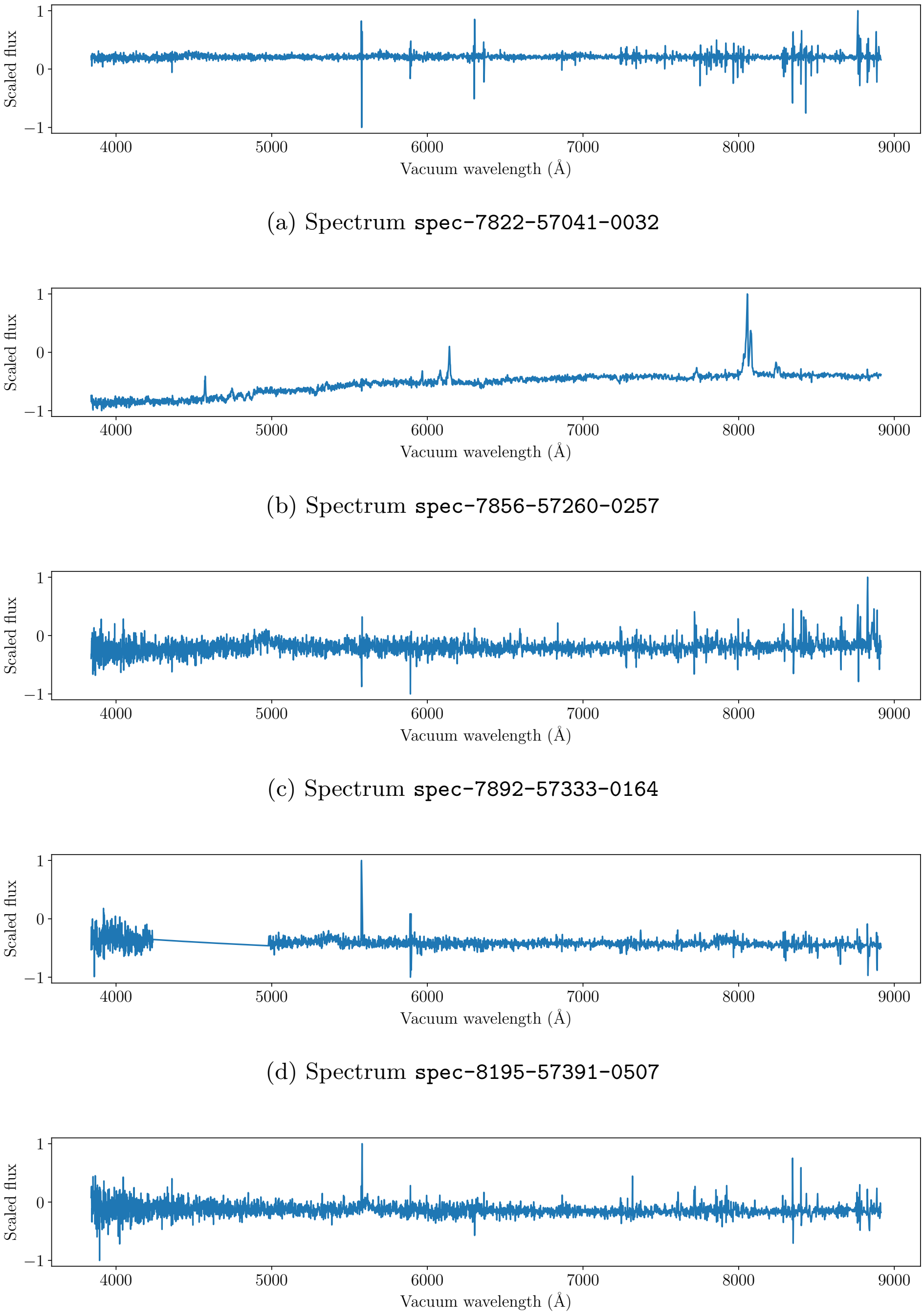
(b) Spectrum `spec-6041-56102-0222`

(c) Spectrum `spec-6832-56426-0411`

(d) Spectrum `spec-7235-56603-0370`

(e) Spectrum `spec-7338-56717-0446`

Figure A.1: The second part of sample of source false negatives



(a) Spectrum `spec-7822-57041-0032`

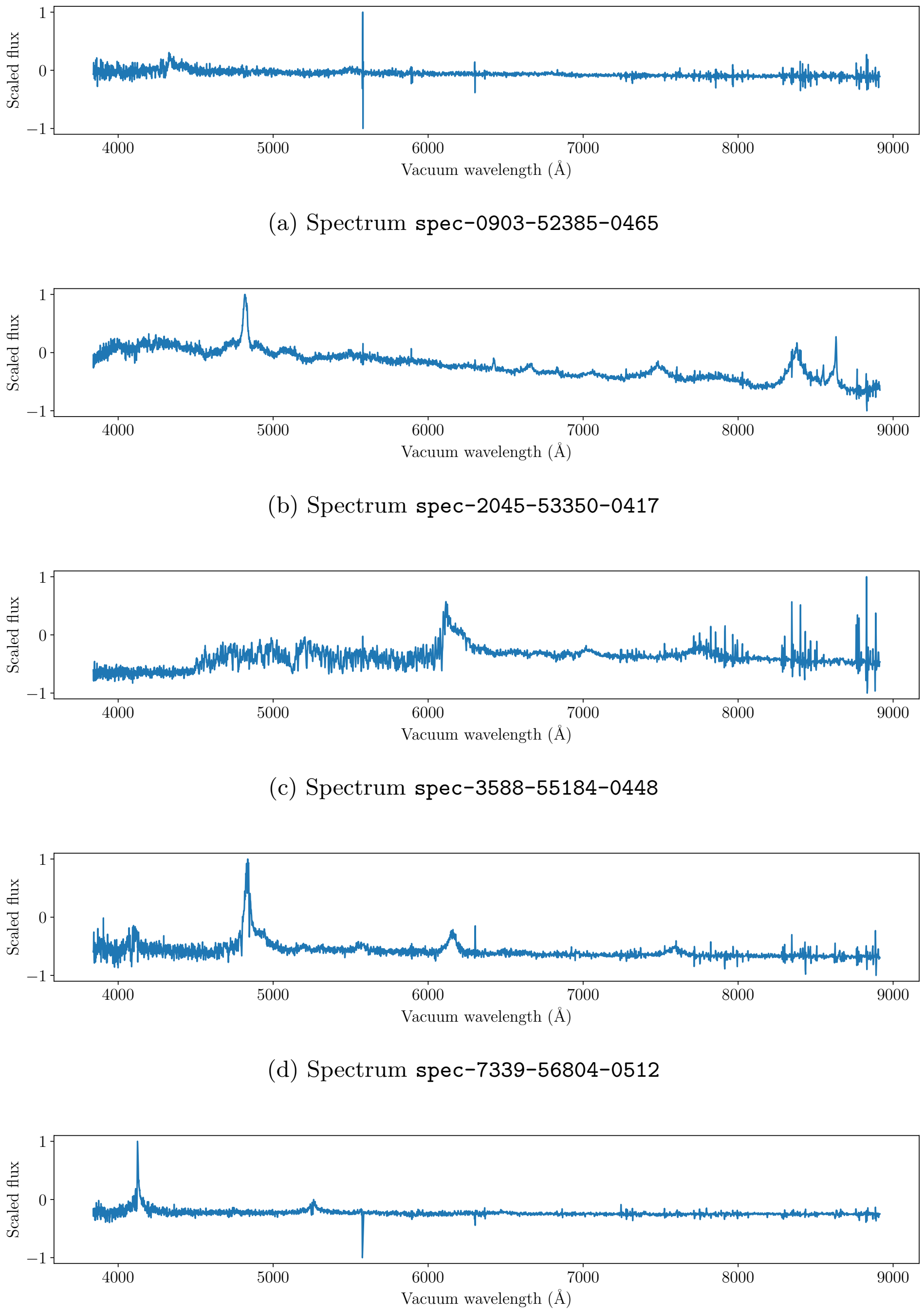
(b) Spectrum `spec-7856-57260-0257`

(c) Spectrum `spec-7892-57333-0164`

(d) Spectrum `spec-8195-57391-0507`

(e) Spectrum `spec-8846-57428-0544`

Figure A.2: The third part of sample of source false negatives



(a) Spectrum spec-0903-52385-0465

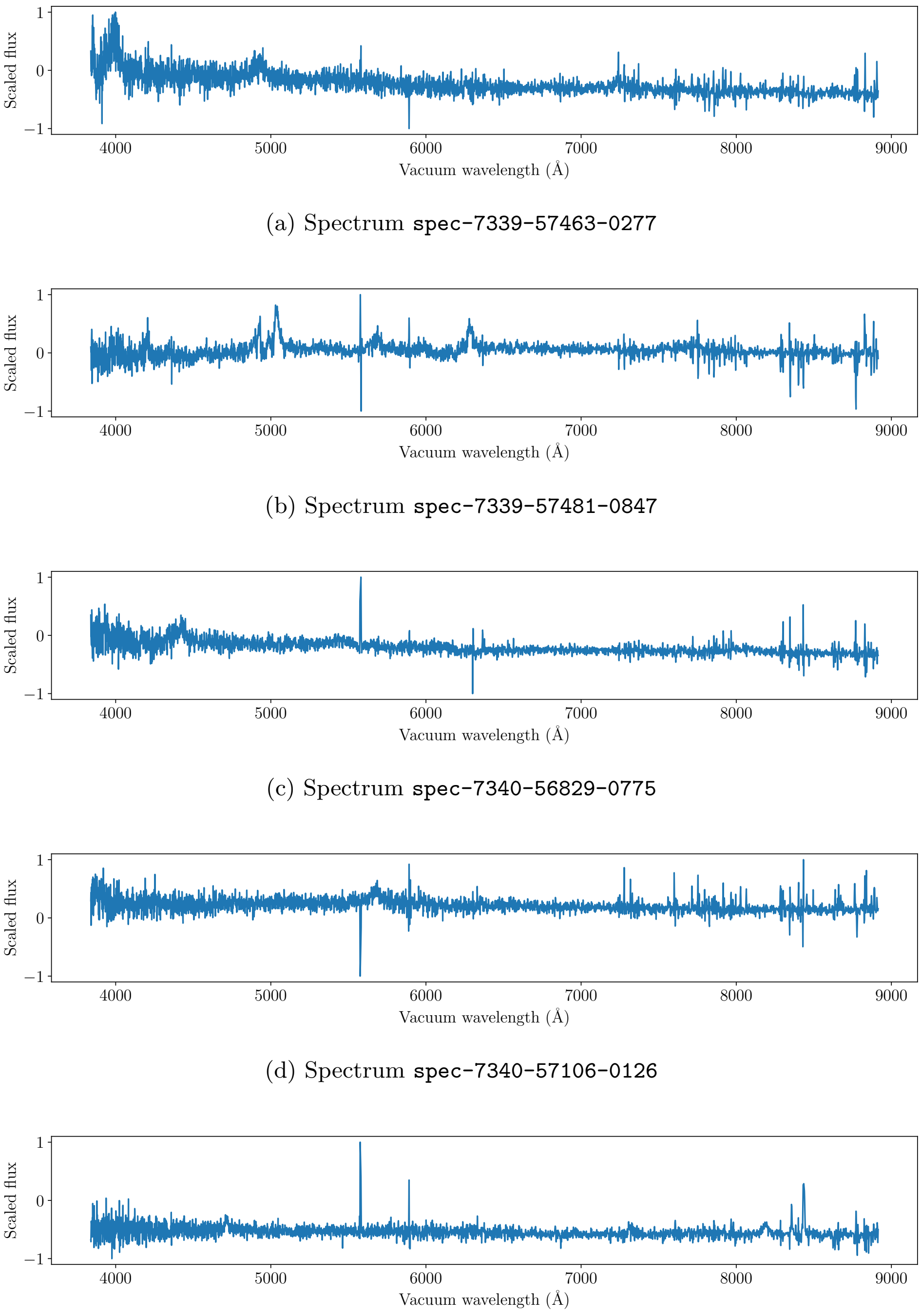
(b) Spectrum spec-2045-53350-0417

(c) Spectrum spec-3588-55184-0448

(d) Spectrum spec-7339-56804-0512

(e) Spectrum spec-7339-56804-0896

Figure A.3: The second part of sample of source false positives



(a) Spectrum `spec-7339-57463-0277`

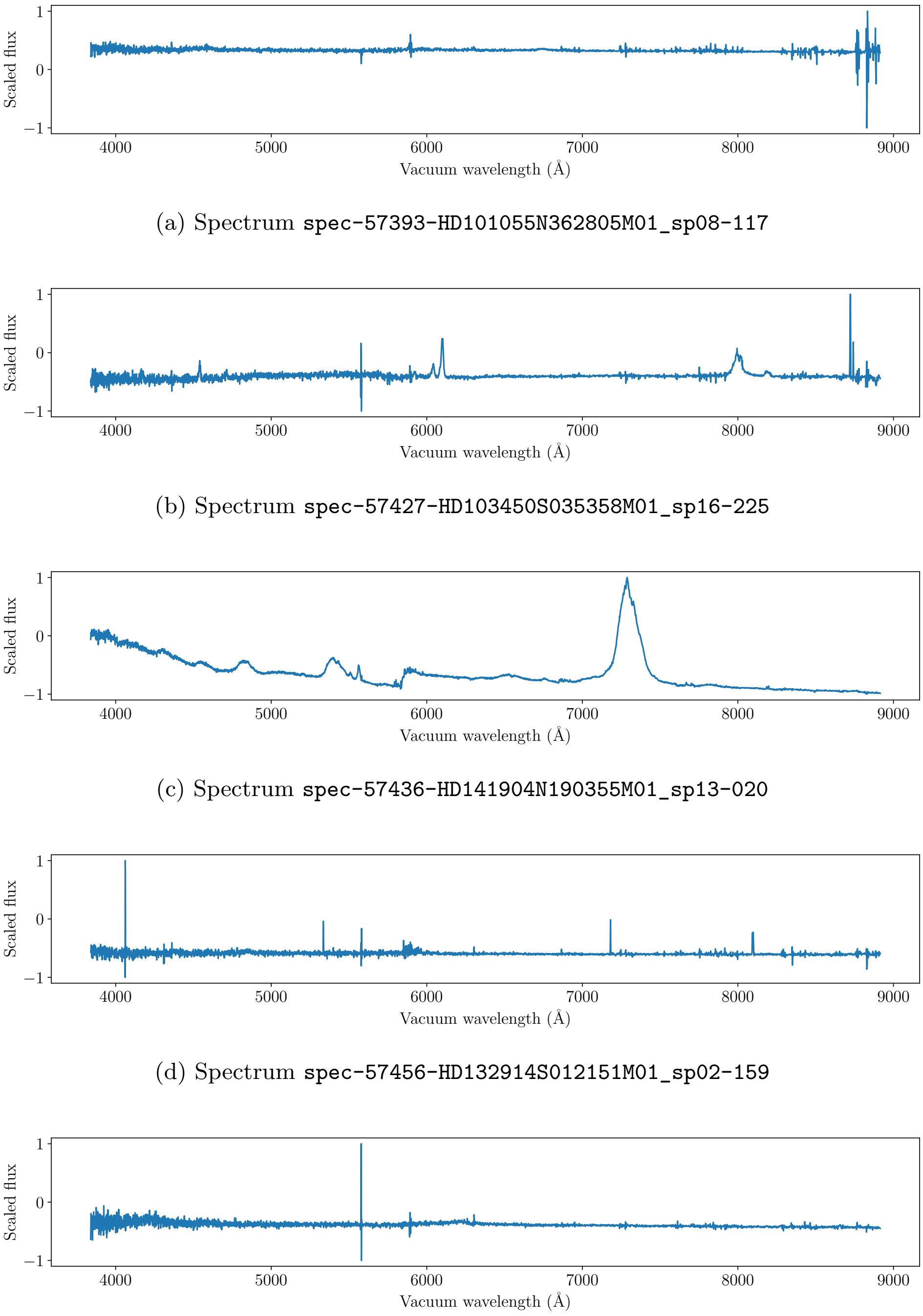
(b) Spectrum `spec-7339-57481-0847`

(c) Spectrum `spec-7340-56829-0775`

(d) Spectrum `spec-7340-57106-0126`

(e) Spectrum `spec-8404-57481-0022`

Figure A.4: The third part of sample of source false positives



(a) Spectrum `spec-57393-HD101055N362805M01_sp08-117`

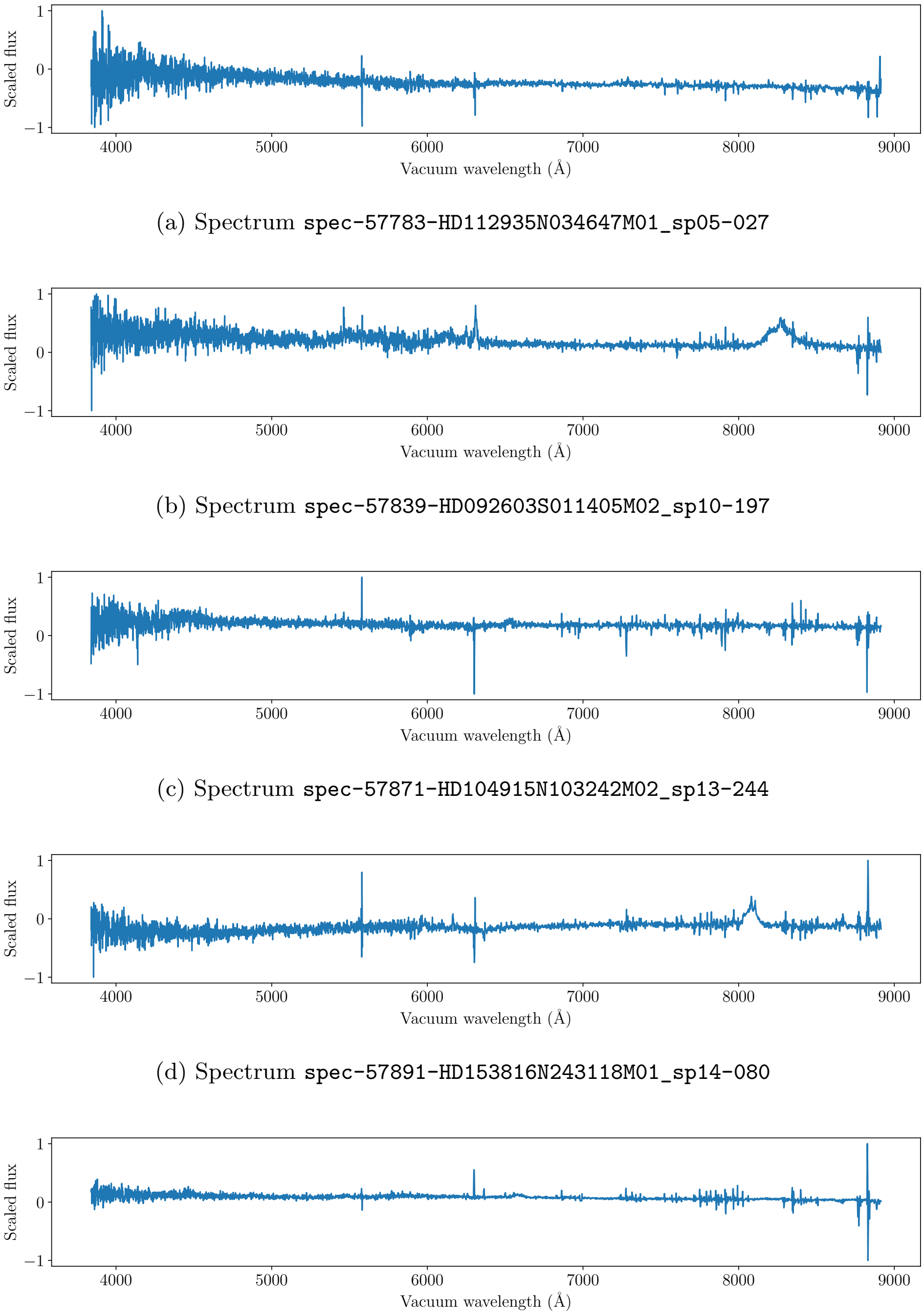
(b) Spectrum `spec-57427-HD103450S035358M01_sp16-225`

(c) Spectrum `spec-57436-HD141904N190355M01_sp13-020`

(d) Spectrum `spec-57456-HD132914S012151M01_sp02-159`

(e) Spectrum `spec-57691-EG024221N200041M01_sp06-047`

Figure A.5: The second part of sample of target false negatives



(a) Spectrum `spec-57783-HD112935N034647M01_sp05-027`

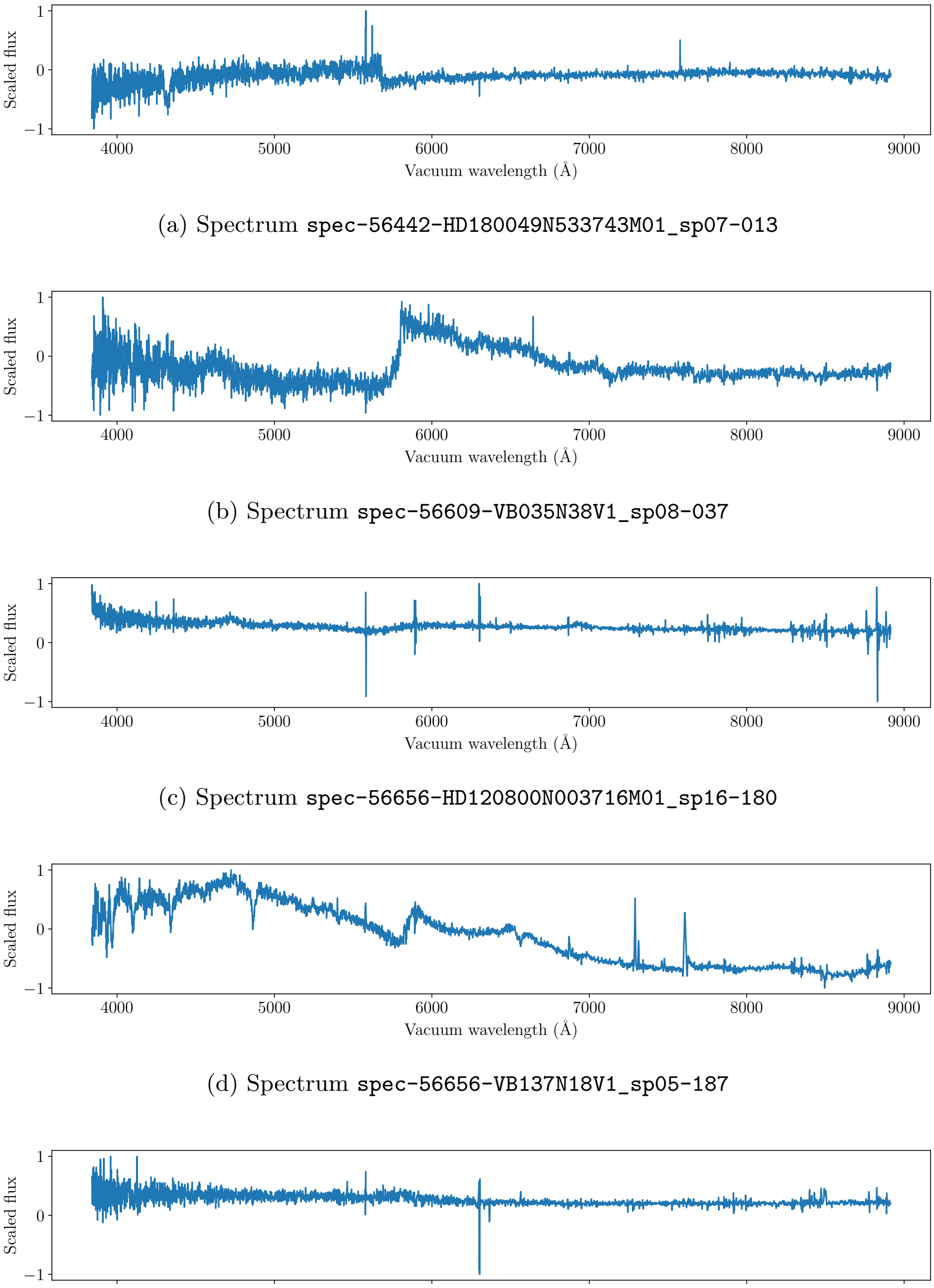
(b) Spectrum `spec-57839-HD092603S011405M02_sp10-197`

(c) Spectrum `spec-57871-HD104915N103242M02_sp13-244`

(d) Spectrum `spec-57891-HD153816N243118M01_sp14-080`

(e) Spectrum `spec-57900-HD154528N083911M01_sp08-226`

Figure A.6: The third part of sample of target false negatives



(a) Spectrum `spec-56442-HD180049N533743M01_sp07-013`

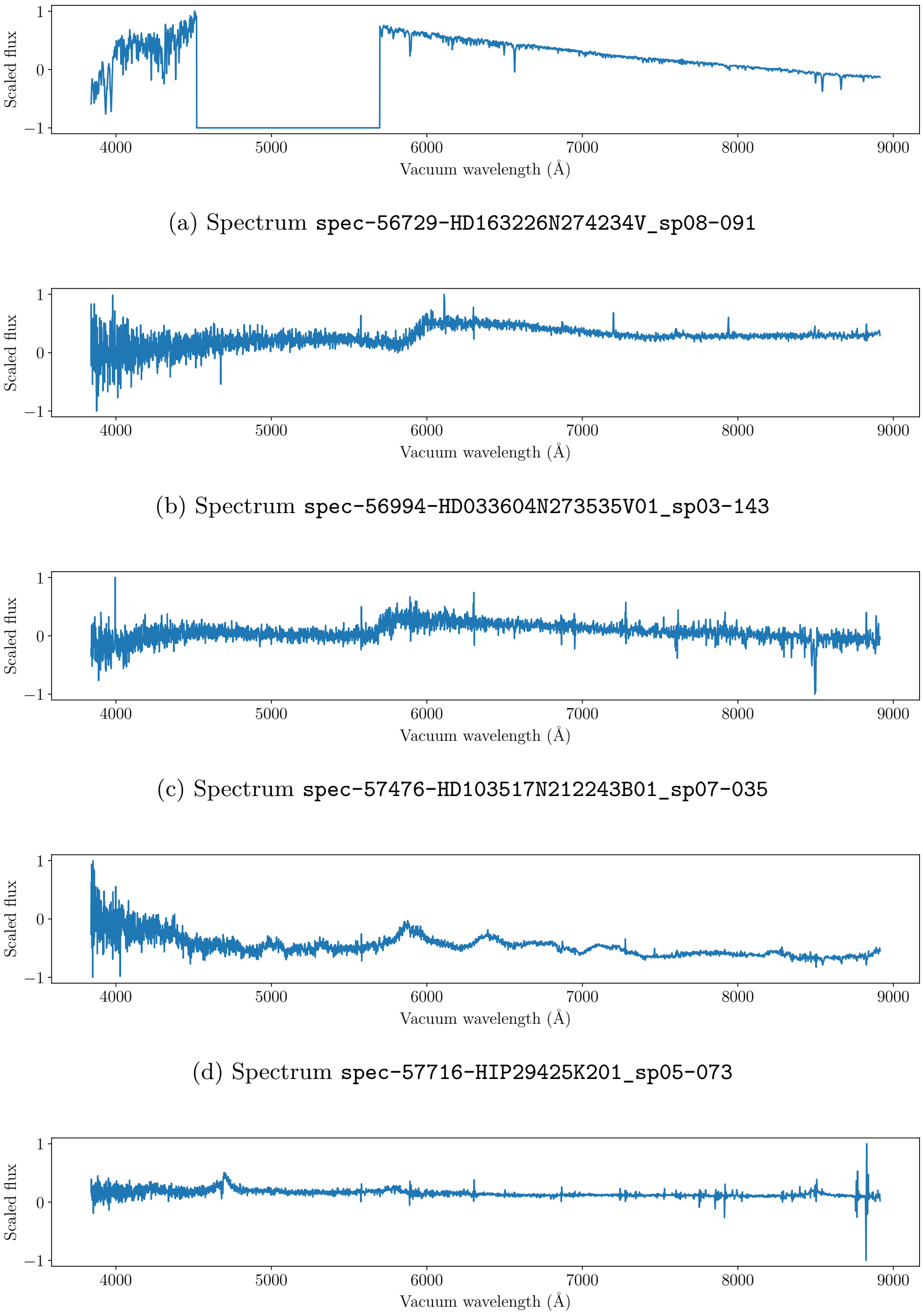
(b) Spectrum `spec-56609-VB035N38V1_sp08-037`

(c) Spectrum `spec-56656-HD120800N003716M01_sp16-180`

(d) Spectrum `spec-56656-VB137N18V1_sp05-187`

(e) Spectrum `spec-56680-GAC057N34B2_sp11-079`

Figure A.7: The second part of sample of target false positives



(a) Spectrum `spec-56729-HD163226N274234V_sp08-091`

(b) Spectrum `spec-56994-HD033604N273535V01_sp03-143`

(c) Spectrum `spec-57476-HD103517N212243B01_sp07-035`

(d) Spectrum `spec-57716-HIP29425K201_sp05-073`

(e) Spectrum `spec-57841-HD081449N430205M02_sp12-236`

Figure A.8: The third part of sample of target false positives

Appendix B

Acronyms

AGN Active galactic nuclei

APOGEE Apache Point Observatory Galactic Evolution Experiment

BOSS Baryon Oscillation Spectroscopic Survey

CDD Charge-coupled device

CORAL Correlation Alignment

CNN Convolutional neural network

DA Domain adaptation

DANN Domain-Adversarial Neural Networks

DDC Deep Domain Confusion

DeCAF Deep Convolutional Activation Feature

DRCN Deep Reconstruction-Classification Network

EM Electromagnetic

FITS Flexible Image Transport System

FWHM Full width at half maximum

GPU Graphics processing unit

GRL Gradient reversal layer

IID Independent and identically distributed

LAMOST Large Sky Area Multi-Object Fiber Spectroscopic Telescope

QSO Quasar, Quasi-stellar object

SDSS Sloan Digital Sky Survey

SGD Stochastic gradient descent

SNR Signal-to-noise ratio

t-SNE t-Distributed Stochastic Neighbor Embedding

UMAP Uniform Manifold Approximation and Projection

Appendix C

Contents of Enclosed CD

```
README.md ............................ file with CD contents description
├─ thesis.pdf ........................ thesis text in PDF format
└─ src ............................... directory of source codes
   ├─ latex .......................... directory of LaTeX source codes of the thesis
   └─ experiments .................... directory of source codes of experiments
```