SELF- AND WEAKLY- SUPERVISED DEEP LEARNING METHODS WITH APPLICATIONS IN BIOMETRIC AND BIOMEDICAL DATA

by
MEHMET CAN YAVUZ

Submitted to the Graduate School of Engineering and Natural Sciences
in partial fulfilment of
the requirements for the degree of Doctor of Philosophy

Sabancı University
July 2023

SELF- AND WEAKLY- SUPERVISED DEEP LEARNING METHODS WITH APPLICATIONS IN BIOMETRIC AND BIOMEDICAL DATA

Approved by:

Prof. Dr. Ayşe Berrin YANIKOĞLU YEŞİLYURT ....
(Dissertation Supervisor)

Assoc. Dr. Prof. Öznur TAŞTAN ....

Asst. Prof. Dr. Hüseyin ÖZKAN ....

Prof. Dr. Mine Elif KARSLIGİL ....

Asst. Prof. Dr. Yakup GENÇ ....

Date of Approval: July 13, 2023

ABSTRACT

SELF- AND WEAKLY- SUPERVISED DEEP LEARNING METHODS WITH APPLICATIONS IN BIOMETRIC AND BIOMEDICAL DATA

MEHMET CAN YAVUZ

COMPUTER SCIENCE AND ENGINEERING Ph.D DISSERTATION, JULY 2023

Dissertation Supervisor: PROF. DR. AYŞE BERRİN YANIKOĞLU YEŞİLYURT

Keywords: Variational Methods, Weakly-Labeled Data, Self-Supervised, Pseudo-Labeling, Contrastive Learning, Beta-Divergence

This dissertation introduces novel deep learning methodologies for effectively leveraging weakly-labeled biomedical data and uncurated/unlabeled biometric data. The thesis is divided into three major parts. In the first part, we present a classifier that combines 2D and 3D classifiers that are trained with weak supervision using volume-wise labeled CT lung images. The main contribution of the thesis is a new representation learning method, extending the contrastive learning framework with the variational approach. In the second part of the thesis, we present a semi-supervised approach using the variational contrastive design, applied to learning face attributes from web-collected face images. This technique, called VCL-PL, is specifically designed to counter the inherent noise found in the collected images. Through various experimental setups, the method demonstrates an enhancement in accuracy over supervised or state-of-the-art self-supervised methods. The last part of the dissertation develops a robust self-supervised learning model, VCL, that combines variational contrastive learning with beta-divergence. This model exhibits better performance than state-of-the-art models when used with unlabeled, uncurated, and noisy datasets. Through the development of these methodological advancements and the introduction of novel datasets, this dissertation contributes to learning from weakly-labeled data in the medical domain and introduces the variational contrastive learning approach that better handles noisy data and low data regimes, in the biometric domain.

ÖZET

KENDINDEN- VE ZAYIF- DENETIMLI DERIN ÖĞRENME YÖNTEMLERI ILE BIYOMETRI VE BIYOMEDIKAL VERILERDEKI UYGULAMALARI

MEHMET CAN YAVUZ

BILGISAYAR BILIMLERI VE MUHENDISLIGI DOKTORA TEZİ,
TEMMUZ 2023

Tez Danışmanı: PROF. DR. AYŞE BERRİN YANIKOĞLU YEŞİLYURT

Anahtar Kelimeler: Varyasyonel Yöntemler, Zayıf Etiketlenmiş Veri, Kendinden Denetimli, Sahte Etiketleme, Karşıt Öğrenme, Beta-Divergence

Bu tez, zayıf etiketlenmiş biyomedikal verileri ve düzensiz/etiketsiz biyometrik verileri etkili bir şekilde kullanmak için yeni derin öğrenme metodolojilerini tanıtır. Tez üç ana bölüme ayrılmıştır. İlk bölümde, hacim bazında etiketlenmiş CT akciğer görüntüleri kullanılarak zayıf denetimle eğitilmiş 2D ve 3D tekniklerini kullanan iki sınıflandırıcı sunulmaktadır. Tezin ana katkısı, varyasyonel yaklaşımla karşılaştırmalı öğrenme çerçevesini genişleten yeni bir temsil öğrenme yöntemidir. Tezin ikinci bölümünde, web'ten toplanan yüz görüntülerinden yüz özelliklerini öğrenmek için uygulanan varyasyonel karşılaştırmalı tasarımı kullanan bir yarı denetimli yaklaşım sunuyoruz. Bu teknik, VCL-PL olarak adlandırılır ve toplanan görüntülerde bulunan doğal gürültüyü karşılamak için özellikle tasarlanmıştır. Çeşitli deneysel kurulumlar aracılığıyla, yöntem denetimli veya güncel öz denetimli yöntemler üzerinde bir doğruluk artışı gösterir. Tezin son bölümünde, varyasyonel karşılaştırmalı öğrenme ile beta-diverjans formülizasyonunu birleştiren dayanıklı bir öz denetimli öğrenme modeli, VCL, geliştirilir. Bu model, etiketsiz, düzensiz/etiketsiz ve gürültülü veri kümeleriyle kullanıldığında güncel modellerden daha iyi performans sergiler. Bu metodolojik ilerlemelerin geliştirilmesi ve yeni veri kümelerinin tanıtılmasıyla, bu tez, tıbbi alanda zayıf etiketlenmiş verilerden öğrenmeye katkıda bulunur ve biyometrik alanda gürültülü verileri ve düşük veri rejimlerini daha iyi ele alan varyasyonel karşılaştırmalı öğrenme yaklaşımını tanıtır.

ACKNOWLEDGEMENTS

I am deeply grateful to my advisor, Prof. Dr. Ayşe Berrin YANIKOĞLU YEŞİLYURT, whose expertise, understanding, and patience, added considerably to my graduate experience. Her mentorship was paramount in providing a well-rounded experience consistent with my long-term career goals. She encouraged me to not only grow as a computational researcher but also as an independent individual. Her advice on research have been invaluable.

I would also like to express my gratitude to the thesis jury members, Assoc. Dr. Öznur TAŞTAN, Asst. Prof. Dr. Hüseyin ÖZKAN, Asst. Prof. Dr. Yakup GENÇ, and Prof. Dr. Mine Elif KARSLIGİL for their insightful comments and constructive criticisms on this work. Their willingness to give their time so generously has been very much appreciated.

I am thankful to the Computer Engineering Department at our institution for providing a stimulating environment for carrying out this research. My sincere thanks also go to the academic and administrative staff for their unwavering assistance and support during this journey.

A special thanks goes to TRUBA (TÜBİTAK ULAKBİM High Performance and Grid Computing Center) for the computational facilities. The numerical calculations reported in this PhD thesis were fully performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA resources).

Finally, I wish to express my love and gratitude to my beloved family for their understanding and endless love, through the duration of my studies. Their encouragement when the times got rough are the reason I was able to persevere and complete my dissertation.

This accomplishment would not have been possible without them. Thank you.

Dedicated to the friends I lost at a young age.
Genç yaşta kaybettiğim arkadaşlarıma adanmıştır.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

...tune the algorithms to observe their subtle symphony...

– a robot proverb

From social networks to medical imaging systems, we are generating and collecting data at an unprecedented pace. While this data is invaluable to building artificial intelligence systems, it is challenging to design machine algorithms that can leverage data that is unlabelled or only weakly-labelled, and also possibly noisy. This is particularly true in the field of biomedical and biometric data where the sheer volume and the complexity of data pose tremendous opportunities if we can overcome the challenges.

Machine learning, especially deep learning methods, has shown immense promise in various tasks such as natural language understanding, speech recognition, and image and video understanding in the last few years. With these significant advances, there has also been a surge in interest in building machine learning models for processing, analyzing, and understanding large-scale biomedical and biometric data. Yet, the majority of these methods rely heavily on large volumes of labeled data for training, which is often challenging and expensive to acquire.

The primary focus of this dissertation is on the effective management and use of weakly-labeled and uncurated data in the context of biomedical and biometric data. It presents a series of methodological contributions aimed at dealing with weakly-labeled or uncurated data, ranging from the detecting COVID-19 from CT images to face attribute recognition.

Chapter 2 presents a system that offers insights into the challenges and potential of weakly labeled data for detecting COVID-19 infection from CT images. It further introduces a novel dataset of volume-wise labeled CT images for COVID-19 presence. A new 3D model using the whole CT volume at once is developed using weakly labelled data, to work in complement with the existing 2D approach. The model is evaluated extensively over three public datasets, paving the way for further research in this area.

Chapter 3 introduces the novel variational contrastive learning approach, used in context of a semi-supervised framework. The approach, called VCL-PL, is designed to leverage web-collected face images to improve the performance of a face attribute classifier. The efficacy of this method is evaluated across multiple experimental setups, demonstrating its capability to deliver significant improvements in accuracy, especially in low data regimes and in general noisy data.

In Chapter 4, we present a robust self-supervised learning model called VCL, that extends the proposed variational contrastive learning approach with beta-divergence in the self-supervised domain. The performance of this model with unlabelled and noisy datasets is rigorously evaluated, affirming its effectiveness and robustness.

In conclusion, this dissertation makes a contribution to the field of machine learning, especially with regard to handling and learning from weakly-labeled biomedical and uncurated biometric data. The methods and techniques developed in this work have broad applications and implications, potentially paving the way for future research in this rapidly evolving field.

CHAPTER 2

COMPARISON AND ENSEMBLE OF 2D AND 3D APPROACHES FOR COVID-19 DETECTION IN CT IMAGES

Chapter 2 presents a comprehensive exploration of the challenges and opportunities inherent to weakly labeled data, particularly its application in the detection of COVID-19 through volume-wise labeled CT images.

Detecting COVID-19 in computed tomography (CT) or radiography images has been proposed as a supplement to the RT-PCR test. In this chapter, we present a deep learning ensemble, called IST-CovNet, that combined novel 2D (slice-based) and 3D (volume-based) approaches to this problem. The 3D approach is developed in the context of this thesis.

The proposed ensemble obtains 90.80% accuracy and 0.95 AUC score overall on the newly collected IST-C dataset in detecting COVID-19 among normal controls and other types of lung pathologies; and 93.69% accuracy and 0.99 AUC score on the publicly available MosMedData dataset that consists of COVID-19 scans and normal controls only. The system also obtains state-of-art results (90.16% accuracy and 0.94 AUC) on the COVID-CT-MD dataset which is only used for testing. The system is deployed at Istanbul University Cerrahpaşa School of Medicine where it is used to automatically screen CT scans of patients while waiting for RT-PCR tests or radiologist evaluation.

This work is published in the Neurocomputing journal, (Ahmed, Yavuz, Şen, Gülşen, Tutar, Korkmazer, Samancı, Şirolu, Hamid, Eryürekli & others (2022)).

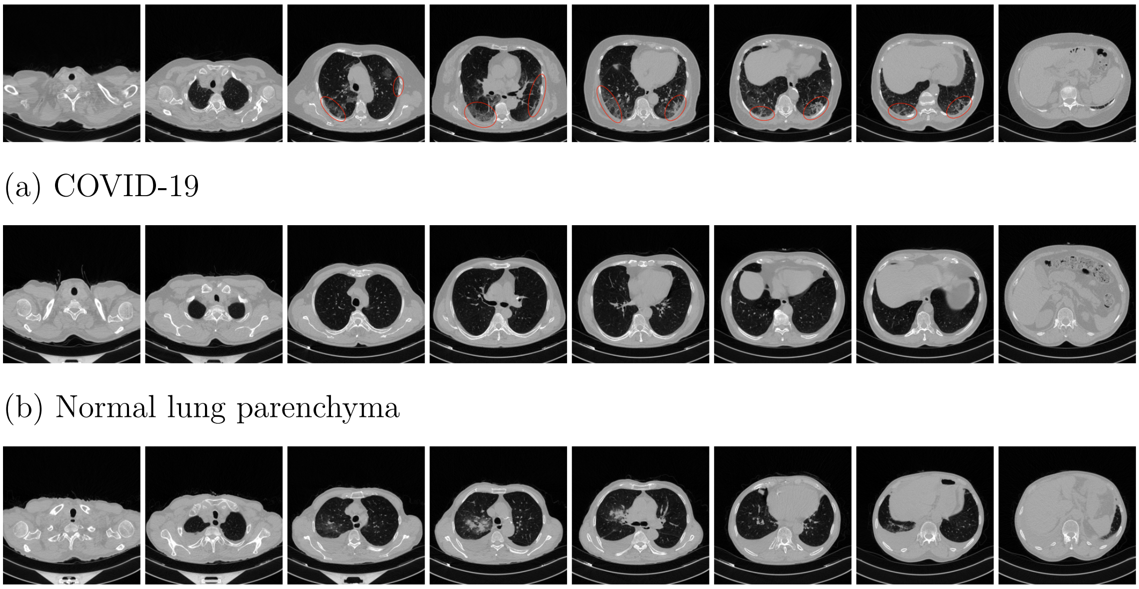
2.1 Introduction

Covid-19 is a highly contagious disease caused by the SARS-CoV-2 virus, which spread rapidly around the world starting early 2020 (Zhu et al. Zhu, Zhang, Wang, Li, Yang, Song, Zhao, Huang, Shi, Lu & others (2020)). The definitive diagnosis of COVID-19 is based on real-time reverse transcriptase polymerase chain reaction (RT-PCR) positivity for the presence of coronavirus Corman, Landt, Kaiser, Molenkamp, Meijer, Chu, Bleicker, Brünink, Schneider, Schmidt & others (2020); Rubin, Ryerson, Haramati, Sverzellati, Kanne, Raoof, Schluger, Volpi, Yim, Martin & others (2020).

Due to the long duration to obtain the RT-PCR results and the prevalence of false negative results Long, Tang, Shi, Li, Deng, Yuan, Hu, Xu, Zhang, Lv & others (2020), the medical community has been in search of alternative or supplementary methods, including screening chest X-ray or Computed Tomography (CT) scans of patients for patterns of pneumonia caused by the COVID-19 infection. This work originated at Istanbul University-Cerrahpaşa Hospital, to automatically analyze CT scans while the patient is still in the tomography room, for successful containment of infected cases.

The chest X-ray consists of a single 2-dimensional, frontal image of the thorax, while a chest CT scan consists of a variable number of 2-dimensional axial slice images. The number of slices in a CT volume vary (typically [200-500]) and the shape and size of lung tissue within the slice vary significantly between slices. Hence, detection of COVID-19 infection in a chest X-ray presents as a typical image classification problem, while the CT scan provides a richer, but also more challenging input.

Detecting COVID-19 in computed tomography or X-ray images has been studied widely since the beginning of the pandemic Chaddad, Hassan & Desrosiers (2021); Hammoudi, Benhabiles, Melkemi, Dornaika, Arganda-Carreras, Collard & Scherpereel (2020); Li, Qin, Xu, Yin, Wang, Kong, Bai, Lu, Fang, Song & others (2020); Liu, Gao, He, Liu & Yin (2020); Narin, Kaya & Pamuk (2020); Wang & Wong (2020); Wang et al. (2020); Xu, Jiang, Ma, Du, Li, Lv, Yu, Ni, Chen, Su & others (2020); Yu, Lu, Guo, Wang & Zhang (2021); Zhang, Zhang, Zhang & Wang (2021). Some of these systems only address the 2-class problem: distinguishing between normal and COVID-19 infected parenchyma (e.g Narin et al. (2020); Yu et al. (2021)), while others aim to detect COVID-19 infection among all possible conditions (normal lung parenchyma and other lung pathologies, including other types of pneumonia). The latter, which is the problem addressed in this work, is a significantly more difficult



(a) COVID-19

(b) Normal lung parenchyma

(c) Others (including Non-COVID-19 pneumonia, tumors and emphysema.)

Figure 2.1 IST-C dataset samples. The ground glass opacities can be observed in the COVID-19 images, marked with the ellipses.

problem as non-COVID-19 pneumonia presents similar patterns to COVID-19.

We developed a deep learning ensemble (IST-CovNet) for detecting COVID-19 infections in high resolution chest CT scans, where we compare and combine *slice-based* and *volume-based* approaches. The slice-based approach takes individual slices as input and outputs the COVID-19 probability for that slice. To obtain the patient-level decision from slice-level predictions, we have evaluated different classifier combination techniques, including simple averaging and Long-Short Term Memory (LSTM) networks. This system is based on transfer learning using the Inception-ResNet-V2 Szegedy, Ioffe, Vanhoucke & Alemi (2017) network that is expended with a novel attention mechanism Dang, Liu, Stehouwer, Liu & Jain (2020).

The volume-based approach is based on the DeCoVNet architecture of Wang et al. Wang et al. (2020) with some modifications to the architecture. In both approaches, we make use of the pretrained U-Net Ronneberger et al. (2015) architecture to focus on the lung regions in the slice images. To combine 2D and 3D systems, we used ensemble averaging, multi-variate regression and Support Vector Machines (SVMs).

A new dataset (IST-C) is collected at Istanbul University-Cerrahpaşa, Cerrahpaşa Faculty of Medicine (IUC), consisting of 712 chest CT scans collected from 645 patients. It includes samples from COVID-19 infected patients, as well as normal lung parenchyma and Non-COVID-19 pneumonia, tumors and emphysema patients. Figure 2.1 shows three samples from the IST-C dataset collected in this work, including a typical COVID-19 involvement pattern termed as *ground glass opacity*, along with normal lung parenchyma and other conditions including non-COVID-19 pneumonia, tumors and emphysema.

The contributions of this work are the following:

- We present a deep neural network ensemble (IST-CovNet) that combines 2D (slice-based) and 3D (volume-based) approaches and achieves state-of-art accuracies on the publicly available MosMedData Morozov et al. (2020) and IST-C datasets collected in this work. The proposed system also obtains close to state-of-art results on the COVID-CT-MD Afshar et al. (2020) dataset which is not used for training, demonstrating the inter-operability of the proposed system.

- Rather than adopting a single approach as done commonly in the COVID-19 AI literature, we compare 2D and 3D approaches, along with relevant preprocessing, attention and combination alternatives on 3 different data sets, and combine the best systems to obtain the final ensemble classifier. Our approaches include novel aspects that contribute to improved performance, such as a new attention model and slice-level combination using LSTMs in the 2D system and an extended novel architecture in the 3D approach.

- We have collected a medium-size dataset consisting of 712 high resolution chest CT scans from 645 people, showing normal lung parenchyma, COVID-19 infections, as well as other pathologies (including non-COVID-19 pneumonia, tumors and emphysema). The IST-C dataset is made public along with our results as benchmark[1].

- The system is deployed at one of the biggest hospitals in Turkey (Istanbul University Cerrahpaşa School of Medicine), to screen for CT scans that show COVID-19 infections for timely containment of infected patients.

2.2 Related Works

Automatic COVID-19 detection research in literature have targeted both chest X-rays Hammoudi et al. (2020); Narin et al. (2020); Wang & Wong (2020) and CT

[1] https://github.com/verimsu/IST-C-dataset

Dataset	Description	Res.	CT Scans	Slices	C19	Nrml	Others
CC-19 Kumar et al. (2020)	CT scans collected from 3 different hospitals and 6 different scanners	High	89	34,006	68	21	0
MosMedData Mozorov et al. (2020)	CT scans with indicated COVID-19 severity level (4 levels)	High	1,110	46,411	856	254	0
BIMCV-COVID19 Iglesia Vaya et al. (2020)	COVID-19 and Normal only	High	2,068	314,056	1,141	927	0
COVID-CT-MD Afshar et al. (2020)	COVID-19, Normal and Other	High	305	45,471	170	77	61
HKBU-HPML-COVID-19 He Wang et al. (2020)	COVID-19, Normal and Other Collected from different hospitals	High	6,878	406,449	2,513	1,927	2,435
IST-C (this work)	COVID-19, Normal, Other CT scans from one hospital	High	712	200,647	336	245	131

Table 2.1 Some of the publicly available COVID-19 CT scan datasets. The first four datasets contain scans of only COVID-19 infected patients and those with normal lung parenchyma. IST-C dataset collected in this work includes non-COVID-19 pneumonia, tumors and emphysema as well.

scans Li et al. (2020); Liu et al. (2020); Wang et al. (2020); Xu et al. (2020) as input and there have been many systems published in peer-reviewed venues or preprint sites since the beginning of the pandemic. There are also systems that aim to leverage the potential of the two biomedical imaging modalities, taking as input both a chest CT and a chest X-ray Chaddad et al. (2021); Chaudhary & Pachori (2021); Zhang et al. (2021).

Comprehensive literature reviews can be found in surveys about artificial intelligence (AI) based approaches to COVID-19 in Islam, Karray, Alhajj & Zeng (2020); Ozsahin, Sekeroglu, Musa, Mustapha & Ozsahin (2020); Shi, Wang, Shi, Wu, Wang, Tang, He, Shi & Shen (2020). Among these surveys, Ozsahin et al. Ozsahin et al. (2020) structure their survey into 3 groups: systems aiming to differentiate between i) COVID-19 versus normal lung parenchyma, ii) COVID-19 versus non-COVID-19 (sometimes called COVID-19 negative) consisting of both normal lung parenchyma and other types of pneumonia, and iii) COVID-19 versus other types of pneumonia. Systems included in this survey report the accuracy and/or the Area Under the Curve (AUC) score related to the Receiver Operating Characteristic (ROC) curve. State-of-art results are above 90% accuracy and 0.95 AUC for the first problem (i)

			Total	**Avg**
	# Patients	**# CT volumes**	**# slices**	**# slices/person**
COVID-19	300	336	92,905	276 ± 83
”Normal”	245	245	67,712	277 ± 67
”Other”	131	131	40,030	306 ± 98
Overall	645	712	200,647	282 ± 82

Table 2.2 Overview of the IST-C dataset: COVID-19 infections are all people diagnosed with the infection; ”Normal” is everyone with no infection whatsoever; ”Other” is all other types, including pneumonia, tumors and emphysema.

and approximately 88% accuracy and 0.90 AUC for the second problem (ii).

AI based COVID-19 detection approaches are two-fold: *2D* or *slice-based* approach, taking a single slice image as input and obtain a score for individual slices Narin et al. (2020), while *3D* or *volume-based* approach, taking the whole volume (sequence of slices) as the input and produce a single score for the patient Hammoudi et al. (2020); Li et al. (2020); Wang et al. (2020); Xu et al. (2020). Note that while a patient may have more than one CT scan, we treat each CT scan as if it belongs to a unique patient and use the terms CT-level and patient-level interchangeably in this work.

In slice-based models, output scores of slices are often combined by averaging, to obtain the patient-level scores and decisions. Among volume-based approaches, most systems use adaptive-pooling operation for combining slice level features Li et al. (2020); Wang et al. (2020), while others use a more implicit combination using Recurrent Neural Networks (RNN) Hammoudi et al. (2020). An advantage of 2D models is the direct interpretability while the 3D models is potentially more powerful as they leverage end-to-end optimization rather than a 2-stage process of obtaining patient-level scores after slice-level scores.

In the remainder of this section, we focus on a subset of the literature due to space limitations, reporting systems that analyze CT scans (not X-rays), address the problem of separating COVID-19 samples from all non-COVID-19 samples (not just normal lung parenchyma), and appear on peer-reviewed venues. While we include performance results reported in the referenced works, it should be kept in mind that most of the results cannot be directly compared, as the test datasets or experimental settings vary between systems.

Li et. al Li et al. (2020) developed a model called COVNet, that is based on the Resnet Szegedy et al. (2017) backbone. The varying number of CT slices are input into parallel branches that use shared weights and the deep features extracted from each are combined by a max-pooling operation. They report 0.96 AUC score on the

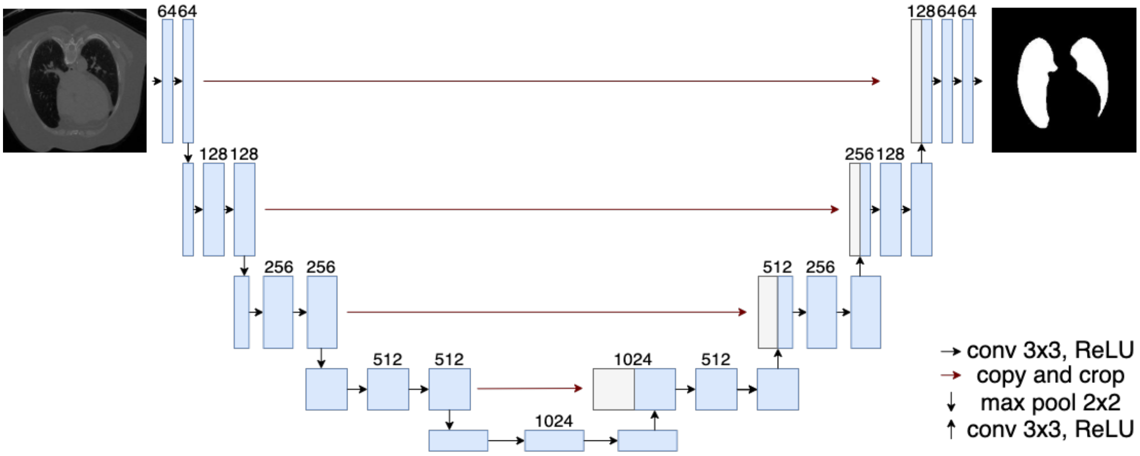


Figure 2.2 Segmentation network U-NetRonneberger et al. (2015): input is a slice image and the output is the corresponding lung mask.

3-class classification problem of distinguishing between normal lung parenchyma, COVID-19 and other lung pathologies.

Wang et. al. Wang et al. (2020) use the pretrained U-Net Ronneberger et al. (2015) architecture to segment lung regions and obtain the lung mask volume. Then, the proposed DeCovNet takes the whole CT volume along with the corresponding lung mask volume as input, and outputs a patient-level probability for COVID-19. The variable number of slices is handled using adaptive maxpool operation. Authors report %0.91 accuracy and a 0.959 AUC score on the 2-class problem of separating COVID-19 positive cases from all others (non-COVID-19, including other pneumonia).

Hammoudi et al. Hammoudi et al. (2020) split a chest X-ray into patches and after obtaining patch-level predictions using deep convolutional networks, they use bidirectional recurrent networks to combine them to predict patient health status.

Liu et. al Liu et al. (2020) fine-tune well-known deep neural networks for the primary task of detecting COVID-19 and the auxiliary task of identifying the different types of COVID-19 patterns (e.g. ground glass opacities, crazy paving appearance, air bronchograms) observed in the slice-image. They report that using the auxiliary task helps with the detection performance, which reaches 89.0% accuracy.

Harmon et al. Harmon, Sanford, Xu, Turkbey, Roth, Xu, Yang, Myronenko, Anderson, Amalou & others (2020) test the performance of a baseline deep neural network approach in a multi-center study. The approach consists of lung segmentation using AH-Net Liu, Xu, Zhou, Pauly, Grbic, Mertelmeier, Wicklein, Jerebko, Cai & Comaniciu (2018) and the classification of segmented 3D lung regions by pretrained DenseNet121 Huang, Liu, Van Der Maaten & Weinberger (2017). On a 1,337-patient test set they report an accuracy of 0.908 and AUC score of 0.949.

Among systems that report on the MosMedData dataset, Jin et al. Jin, Chen, Cao, Xu, Tan, Zhang, Deng, Zheng, Zhou, Shi & others (2020) propose a deep learning slice-based approach employing ResNet-152 He, Zhang, Ren & Sun (2016) architecture. The developed model achieved comparable performance to experienced radiologists with an AUC score of 0.93.

He at al. He, Wang, Chu, Shi, Tang, Liu, Yan, Zhang & Ding (2021) proposed a differentiable neural architecture search framework for 3D chest CT-scans classification with the Gumbel-Softmax technique Jang, Gu & Poole (2016) to improve the searching efficiency. The experimental results show that their automatically searched model outperforms three of the state-of-the-art 3D models achieving an accuracy of 82.29% on MosMedData dataset.

In a critical study, Maguolo and Nanni Maguolo & Nanni (2021) show that some automatic COVID-19 detection systems achieve high accuracies even when the lung region is masked in chest X-rays, indicating that the underlying neural networks are learning patterns in the data that are not correlated to the presence of COVID-19. They also discuss how to construct a fair testing protocol. Our single-channel 3D system that achieves the best results in all datasets inputs CT scans that are *masked* with the lung mask (hence, we can assert that there is no information leakage outside of the lung region). Similarly, our 2D system attends to the lung areas, due to the PCA-based attention module.

In another recent and well-publicized survey, Roberts et al Roberts, Driggs, Thorpe, Gilbey, Yeung, Ursprung, Aviles-Rivero, Etmann, McCague, Beer & others (2021) analyze all COVID-19 AI papers published in the first 9-months period of 2020, in terms of their potential potential biases, according to the criteria indicated in Wolff, Moons, Riley, Whiting, Westwood, Collins, Reitsma, Kleijnen & Mallett (2019). After filtering the 2,212 papers found in an initial search according to relevance and quality, the remaining 62 papers were analyzed in depth. Authors conclude that none of the models identified are of potential clinical use due to methodological flaws and/or underlying biases. While this work points out to some important biases in machine learning systems for Covid-19 detection, it is worth pointing out with this categorization, any system evaluated on a public dataset is directly categorized as having a high risk of participant bias (since the participants cannot be verified) and all deep learning approaches are categorized as having high risk of "predictor bias" (since deep features are deemed as abstract and unknown imaging features). In our work, we evaluate the proposed system on two large public datasets (one not used in training at all) and one private dataset collected in the scope of this work, to address participation and outcome biases. We also report cross-validation results for our final system, to eliminate analysis bias. The results obtained on the unseen data Afshar et al. (2020) are state-of-art (in AUC) and also close to the results obtained on the other two datasets, attesting to the generality of the system.

2.3 IST-C Dataset

While there are many works on automatic detection of COVID-19 infection on X-ray or CT images, there were only a handful publicly accessible COVID-19 CT scan datasets at the time of the preparation of this manuscript, shown in Table 2.1. Three of these datasets, CC-19 Kumar, Khan, Zhang, Wang, Abuidris, Amin & Kumar (2020), MosMedData dataset Morozov et al. (2020) and BIMCV-COVID19 de la Iglesia Vayá, Saborit, Montell, Pertusa, Bustos, Cazorla, Galant, Barber, Orozco-Beltrán, García-García, Caparrós, González & Salinas (2020) only contain COVID-19 and normal lung parenchyma. On the other hand, in MosMedData, the COVID-19 samples are also labelled with the severity of the infection in 4 levels (CT-1 to CT-4). In addition to using two large public datasets Afshar et al. (2020); Morozov et al. (2020) in evaluating the system developed in this study, we have also collected a new open-source dataset called IST-C, retrospectively from patients admitted to the Radiology department of Cerrahpaşa Faculty of Medicine from March 2020 to August 2020. The collected dataset consists of 336 chest CT scans that are positive for COVID-19, along with 245 scans showing normal lung parenchyma and 131 scans from Non-COVID-19 pneumonia, tumors and emphysema patients. The COVID-19 scans are selected by expert radiologists from among the patients to whom CT is performed with clinical suspicion of COVID-19 in the emergency department. These two last groups will be called simply as ”Normal” and ”Other” from here on. The detailed statistics of the dataset are shown in Table 2.2.

The collected CT scans in DICOM format consists of 16-bit gray scale images of size 512×512. Each scan is accompanied with a set of personal attributes, such as patient ID, age, gender, location, date, etc. (not used in this work). The average age of the patients is 52 ± 17 years, in which 405 of the patients are male and 274 patients are female.

The annotation of this dataset is at CT scan level: the CT of a patient as a whole is labelled as COVID-19, ”Normal”, or ”Other” by expert radiologists at Istanbul

University-Cerrahpaşa, Cerrahpaşa Faculty of Medicine.

Sample images extracted from COVID-19, ”Normal” and ”Other” classes are shown in Figure 2.1. The anonymized dataset is now shared publicly at http://github.com/suverim.

2.4 Preprocessing

Pixel values of images in the CT dataset are in Hounsfield Unit (HU) which is a radiodensity measurement scale that maps distilled water to 0 and air to -1000. The HU values range between -1024 and 4096, with higher values being obtained from bones and metal implants in the body and lung regions typically ranging in $[-1024, 0]$. Similar to literature, we process chest CT scans such that values higher than $u_{max} = 600$ are mapped to u_{max} and the range $[-1024, u_{max}]$ is normalized to the $[0, 1]$ linearly.

Slice images that are originally (512×512) are resized to match the input size of the respective deep networks, namely 299×299 for slice-based system and 256×256 for the volume-based system. For the 3D approach, we have also reduced the slice count by half, so that the whole CT volume consisting of up to around 500 slice images fits in the GPU memory. We compared two alternatives for this: interpolation of two subsequent slices and skipping every other slide. We and found that the latter results in higher accuracy, even though interpolation is commonly used in many biomedical applications. This reduction is done for only the IST-C dataset where the number of slices per CT scan is high (Table II).

2.5 Lung Segmentation

Lung shapes vary greatly within a chest CT scan, as can be seen in Figure 1. With the aim of focusing on the lung areas, we make use of the pretrained U-Net network to segment lung regions from non-lung areas. Focusing to lung areas is possible by masking the input with the lung mask as done in the 3D system or guiding the attention of the network to the lung areas. This step is found to be quite important in reducing overfitting Gupta, Kaul, Sharma & others (2020), as well as information leakage found in some previous COVID-19 detection systems Maguolo & Nanni (2021).

The U-Net architecture was first proposed by Ronneberger et al Ronneberger et al. (2015) for biomedical image segmentation in general and trained specifically for lungs by Hofmanninger et al. Johannes, Jeanny, Sebastian, Helmut & Georg (2020). Since then has been used in detecting lung regions extensively in the diagnosis of lung health Li et al. (2020); Wang et al. (2020); Xu et al. (2020). The U-Net network, shown in Figure 2.2, is named after the U-shape formed by the encoder branch consisting of convolutional layers and the decoder branch consisting of deconvolution operations. The network also has skip connections in each layer, carrying the output of earlier layers to later layers.

Lung segmentation is applied to individual slices in the CT volume. The output for each slice is the corresponding binary segmentation mask, separating lung areas (including air pockets, tumors and effusions in lung regions) from background or other organs, as shown in Figure 2.3. The segmentation extracts left and right lungs separately, although this information is not used in our model.

Lung segmentation with U-Net is very successful, as reported in Johannes et al. (2020) and also observed in our case. Nonetheless, in order not to miss infected regions, we dilated the masks with a 10-pixel structuring disk. Sample slices from the IST-C dataset and corresponding lung masks obtained by U-Net and the dilated masks are shown in Figure 2.3.

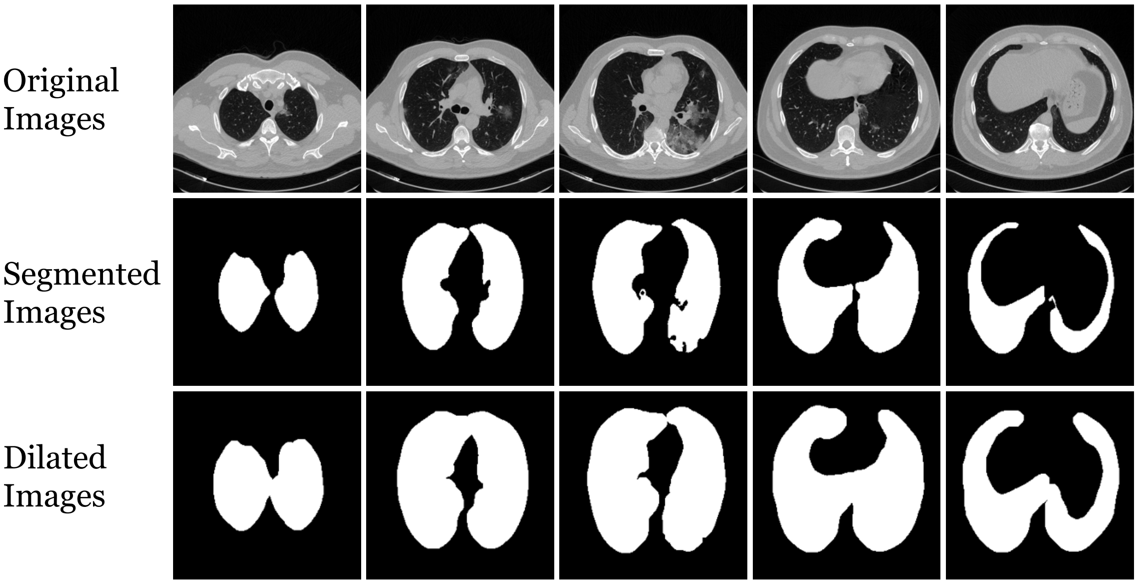


Figure 2.3 Sample slice images along with their segmentation masks as obtained by U-Net and dilated masks.

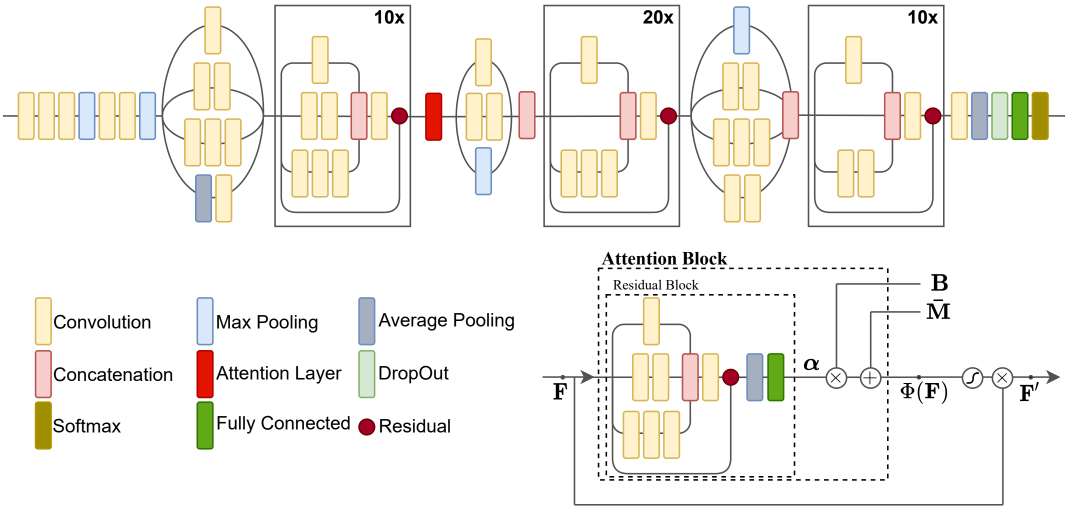


Figure 2.4 The base network and the inserted attention-based layer. Attention layer takes the feature maps $\mathbf{F}$ as an input and estimate the attention map $\Phi(\mathbf{F})$, which is then used to attend to the original features after a sigmoid activation.

2.6 Slice-based Approach

In the 2D approach, CT slices are analyzed independently, before combining them to obtain patient-level predictions. This part of the algorithm is developed by my colleague Dr. Sara Atito Ali during the pandemic research.

2.6.1 Base Model

To construct the base network architecture, we employed Inception-ResNet-V2 architecture Szegedy et al. (2017), one of the top-ranked architectures of the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) 2014 Russakovsky, Deng, Su, Krause, Satheesh, Ma, Huang, Karpathy, Khosla, Bernstein, Berg & Fei-Fei (2015). The network architecture was used successfully in various image classification and object detection tasks Ahmed, Yanikoglu, Zor, Awais & Kittler (2020); Lee, Na & Kim (2019).

Inception-ResNet-V2 network is an advanced convolutional neural network that combines the inception module with ResNet He et al. (2016) to increase the efficiency of the network. The network is 164 layers deep with only 55.9 million parameters. It consists of three main reduction modules with 10, 20, and 10 inception blocks for each module, respectively. The size of the output feature maps of the three reduction modules are 35×35, 17×17, and 8×8, respectively.

Training a large deep learning network from scratch is time consuming and requires a tremendous amount of training data. Therefore, our approach is based on fine-tuning a pre-trained Inception-ResNet-V2 model, that is originally trained on the ImageNet dataset with 1.2 million hand-labeled images of 1,000 different object classes.

2.6.2 Attention Mechanism

To investigate the predictions of the trained base model, we applied Class Activation Mapping (CAM) Zhou, Khosla, Lapedriza, Oliva & Torralba (2016) on some of the images from the validation set. Observing that the attention of the network is not always directed to the area of interest (lung tissues) in misclassified images, we decided to use attention maps and thereby guide the network to the regions that are important to the problem at hand. Attention mechanism has been successfully applied in many computer vision tasks, including fine-grained image recognition Zheng, Fu, Mei & Luo (2017) and face attributes classification Aly & Yanikoglu (2018).

We add an attention map block inserted to the backbone of our base network, as shown in Figure 2.4. The input to the attention layer is a convolutional feature map $\mathbf{F} \in R^{H \times W \times C}$, where H, W, and C are the height, width, and the number of channels, respectively. The output of the attention module is the masked feature map $\mathbf{F}' = \mathbf{F} \odot \sigma(\Phi(\mathbf{F}))$, obtained via element-wise multiplication of the feature maps $\mathbf{F}$ and sigmoid (σ) attenuated attention layer output, $\Phi(\mathbf{F}) \in R^{H \times W}$.

Unlike the standard approach of learning the attention layer fully within the network, the approach used in this work is suggested to be an explainable and modular approach Dang et al. (2020). It makes the assumption that an attention map can be represented using the linear combination of a set of basis vectors, as:

$$\Phi(\mathbf{F}) = \bar{\mathbf{M}} + \mathbf{B} \times \boldsymbol{\alpha}$$

where $\bar{\mathbf{M}} \in R^{H \times W}$ is the average segmentation map; H and W are the height and width of the images; $\mathbf{B} \in R^{H \times W \times n}$ is the matrix of the n basis vectors; and $\boldsymbol{\alpha} \in R^{n \times 1}$ are the coefficients.

The average lung map $\bar{\mathbf{M}}$ and the 12 basis vectors $\mathbf{B}$ are obtained by applying Principal Component Analysis (PCA) to lung masks obtained by U-Net segmentation network. The 12 basis vectors that retain approximately 75% of the variance are shown in Figure 2.5 and U-Net is explained in Section 2.5.

To obtain the attention map coefficients $\boldsymbol{\alpha}$ an additional convolutional block is inserted to the network getting the input from the feature maps $\mathbf{F}$, as shown in Figure 2.4. The convolutional block consists of a separable convolutional layer which is a depth-wise convolution performed independently over each channel of an input, followed by a pointwise convolution, batch normalization, and ReLU activation function. The output of the convolutional block (or attention coefficient block) are the weights α which form the coefficients in the linear basis vector representation.

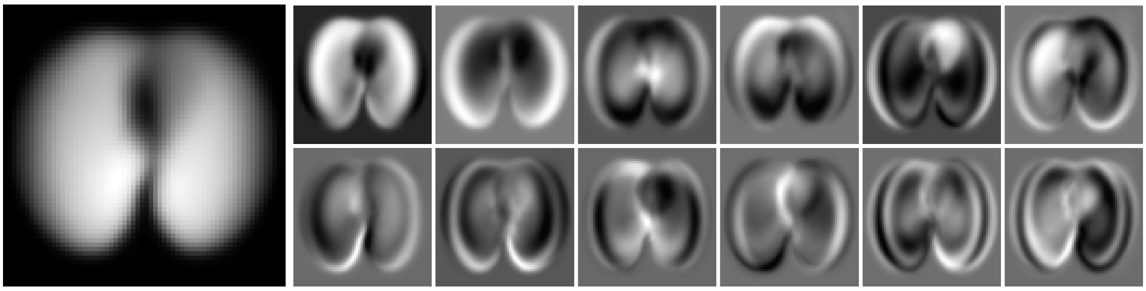


Figure 2.5 (a) Mean mask $\bar{M}$ and (b) The first 12 eigenvectors.
"

2.6.3 Implementation Details

The Inception-ResNet-V2 network used as the base model in the slice-based approach is chosen due to its relatively small size and good performance. The network has an RGB image input size of 299×299. The output layer of the model is replaced with a fully connected layer with 2 hidden units to represent the given classes: COVID-19 vs Non-COVID-19 (including “Normal“ and “Other“ samples). All the layers in the classification network are finetuned and optimized using categorical cross-entropy loss function.

For the attention based model, we added the attention layer after the first reduction block as shown in Figure 2.4. As for the attention loss function, we trained the network in unsupervised manner. Even in the absence of the attention map supervision, we found that the attention module is able to learn the discriminative regions automatically.

The implementation is done using the Inception-ResNet-V2 model provided in the

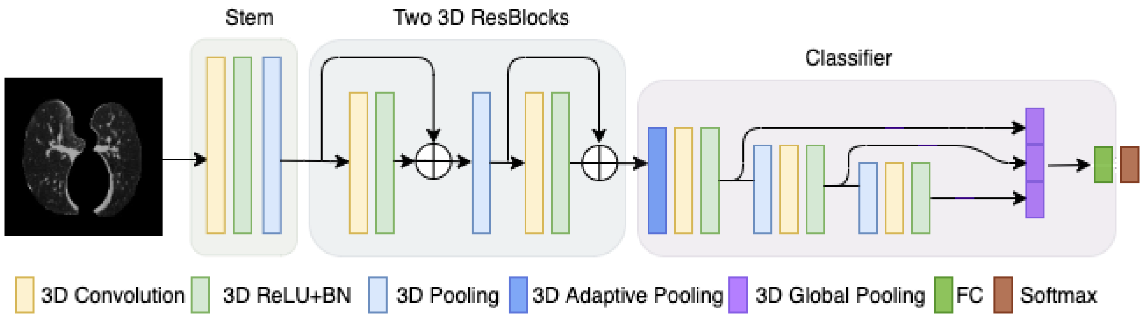


Figure 2.6 Architecture of the classification network which is based on DeCoVNet Wang et al. (2020).

Matlab deep learning toolbox. Several commonly used data augmentation techniques are applied during training, such as rotation [-5 to 5], x and y translation [-5, 5], and x and y scaling [0.9, 1.1].

For all 2D systems, we set the batch size equal to 64 and the initial learning rate as 1e-5 with a total of 50 epochs with Adam optimizer. The training process takes around 100 minutes per epoch for the IST-C dataset and 40 minutes per epoch for the MosMedData dataset using an 8GB Nvidia GeForce RTX 2080 GPU.

2.6.4 Combining Slice-level Predictions

The straightforward approach to obtain patient-level decision is to combine the predictions of the slice based model using simple averaging of slice-level predictions. This is evaluated as the base model, to obtain the patient-level score.

However, simple averaging does not take into account the information about the characteristics of COVID-19 infection, such as the fact that the patterns are often seen in the lower parts of the lungs. To learn this type of information about the slice sequence and to also handle the variable length of the slice sequence, we also used Recurrent Neural Networks (RNNs) as an alternative Rumelhart, Hinton & Williams (1986).

We used Long Short-Term Memory (LSTM) network Hochreiter & Schmidhuber (1997) that is the most powerful type of recurrent network. The input to the network consists of deep features corresponding to each slice in the CT volume. The features are extracted from the last pooling layer of the slice-based CNN model with the attention module, discussed in Section 2.6). The LSTM learns to combine the slice-level features to obtain patient-level predictions.

The LSTM architecture consists of 3 layers: i) a bidirectional LSTM layer with 1024 hidden units and a dropout layer to reduce overfitting; ii) another bidirectional LSTM layer with 512 hidden units; and iii) a fully connected layer with an output size corresponding to the number of classes (2 or 3 in our case). It is important to note that the number of slices in the CT volumes varies substantially which can introduce lots of padding into the training process of the LSTMs and consequently negatively impact the classification accuracy. To overcome this issue, we normalized each CT sequence into 282 slices (i.e. the mean slice count across the IST-C dataset), by either dropping or replicating slices depending on the length of the volume. After normalization, each slice of the CT volume is passed to the trained CNN model for feature extraction. Then, the LSTM model is trained using the sequence of the feature vectors corresponding to the slices.

2.7 Volume-base Approach

The 3D volume-based approach takes as input the whole CT volume and outputs patient-level decision (COVID-19 positive and negative probabilities), based on a single step processing of the input. It uses the lung segmentation volume obtained by U-Net (described in 2.5), followed by a classification network based on DeCoVNet Wang et al. (2020).

The segmentation network (U-Net) takes as input a single slice of the chest CT and outputs a binary mask indicating the lung region. The classification network subsequently takes the CT volume and the corresponding binary mask volume and outputs the patient-level scores.

2.7.1 Classification network

The classification network used in our work is based on DeCoVNet that has been proposed by Wang et al. Wang et al. (2020). We have made some modifications to this network, without significantly changing its architecture. The network consists of three consecutive blocks, (1) Stem (2) ResBlocks (3) Classifier, as shown in Figure 2.6 and detailed in Table 2.3.

The stem block consists of a convolutional layer with a receptive field size 5x7x7 (depth, height, width), as used in well-known networks AlexNet Krizhevsky, Sutskever & Hinton (2012) and Resnet He et al. (2016). The convolutional layer is followed by a batchnorm layer and a pooling layer. We evaluated using both a single channel input, consisting of the slice image with the lung mask applied, as well as the 2-channel input, consisting of the input slice and its lung mask, as in the original network. As we expected, the 2-channel approach led to less efficient training and did not bring accuracy gains.

The second stage of the networks consists of two 3D residual blocks (ResBlocks), with maxpool operation in between to reduce the volume depth by half 64xT/2x64x64. In each block, there are 2 kernels: 3x1x1, 1x3x3 (depth,height,width) with a stride of 1 in each dimension and padding of 1 wherever needed. The output volume is of size 128xT/2x32x32. This block is adopted without any modification.

The third block, called the Progressive classifier, starts with an adaptive maxpool operation that handles the variable number of slices and outputs 128x16 feature maps of size 32×32. It is followed by 3 convolution layers and pooling operations, followed by a fully connected output layer with softmax activation. The main modification in this block is to enrich the feature representation. The original DeCoVNet had a global max pooling layer with $32 \times 1 \times 1 \times 1$ nodes, in the penultimate layer. We extended the Progressive classifier block by adding a new layer of concatenated features obtained using global max pool operation after each of the 3D convolutional layers. More specifically, from a convolutional layer with FxDxHxW output volume, the global max pooling operation outputs a vector of size F. The resulting 192-dimensional $(96+48+48)$ feature vector is fully connected to the output layer (2 nodes with softmax activation), as shown in Figure 2.6. We thus increased the penultimate layer size from 32 to 192. This feature representation was inspired by the work in Ahmed & Yanikoglu (2019), where authors proposed to approximate a deep learning ensemble by replicating the output layer with connections from earlier layers and extending the loss function to include all the loss terms Ahmed & Yanikoglu (2019).

The modified classification network architecture is given in Table 2.3.

	Operation	Output	Penult.
Stem	Conv3d@5x7x7	16xTx64x64	
ResBlocks	ResBlock@3x1x1&1x3x3	64xTx64x64	
	MaxPool3d	64xT/2x64x64	
	ResBlock@3x1x1&1x3x3	128xT/2x64x64	
Progressive	AdaptiveMaxPool3d	128×16×32×32	
Classifier	Conv3d@3×3×3	96×16×32×32	
	GlobalPool3d		96x1x1x1
	———— 2nd Block	————	
	MaxPool3d	96×4×16×16	
	Conv3d@3×3×3	48×4×16×16	
	Dropout3d (p=0.5)	48×4×16×16	
	GlobalPool3d		48×1×1×1
	———— 3rd Block	————	
	MaxPool3d	48×4×16×16	
	Conv3d@3×3×3+ReLU	48×4×16×16	
	GlobalPool3d		48x1x1x1
	FullyConnected	2	

Table 2.3 The 3D-classification network architecture. The residual blocks have two kernels.

2.7.2 Implementation Details

In training the system, the settings are the same and as follows: the loss function is the categorical cross-entropy; the optimizer is the Adam optimizer used with 1e-5 learning rate. Since the graphical card NVidia 2080 can only process a single batch at a time, the batch size is one due to memory constraints. We also used data augmentation exactly same with DeCoVNet: scaling ($1-1.2$), rotation (10 degrees) and translation ($0-10$ pixels).

All 3D systems were run for a fixed number of 200 epochs, observing validation set accuracy at each epoch. The optimal weights were chosen as those giving the highest validation set results. The training process takes around 8 minutes for an epoch of the IST-C dataset and 4 minutes for an epoch of MosMedData dataset, using an 8GB Nvidia GeForce RTX 2080 GPU.

2.8 Combining Multiple Systems

After training the 2D and 3D systems, we combine output of the systems (*patient-level* predictions) to obtain the final prediction. In contrast, please note that Section 2.6.4 discusses the combination of *slice-level* predictions to obtain patient-level predictions for the 2D approach.

The 2D (slice-based) approach is realized with or without the attention mechanism and using different combination mechanisms to obtain the patient-level decision. Similarly, the 3D (volume-based) approach is realized with 1-channel input where the input is masked with the lung mask, or with 2-channel input as in the original DeCovNet Wang et al. (2020).

The combination methods that were evaluated were averaging, multivariate linear regression and Support Vector Machines (SVM). However, we only report ensemble averaging results because multi-variate regression essentially assigned the same weights to the two combined systems and the SVM did not bring noticeable improvements to justify the more complex combination method.

2.9 Experimental Evaluation

We have trained and evaluated the proposed 2D and 3D approaches along with considered submodules, with the IST-C collected in this work (Section 2.3) and the MosMedData dataset Morozov et al. (2020). These results are given in Tables 2.4 and 2.5, respectively. Furthermore, we report results of the above trained models on the COVID-CT-MD dataset Afshar et al. (2020), to evaluate inter-operability performance. These results are given in Table 2.6.

We have done extensive evaluation comparing different preprocessing, segmentation, architecture and ensemble methods. However for the sake of clarity, we report only the most important experiments, using accuracy and AUC scores, in line with the literature. The accuracy values are given together with 95% confidence intervals that are computed using the Wilson score interval method Wilson (1927) for the number of test samples in each dataset.

We split the IST-C database into training/validation/testing data. For "COVID-19" class, 100 volumes are used for testing and the rest are used of the training and the validation. For "Normal" and "Others" classes, 100 and 50 volumes are used for testing, respectively. In total, we assigned 250 volumes for testing and 462 for training and validation. The MosMed dataset was split randomly as train-test, with a 80-20% split, resulting in a 222 test samples. The full COVID-CT-MD dataset was used only for testing.

2.9.1 2D vs 3D

We first compared the effectiveness of 2D and 3D approaches in identifying COVID-19 positive samples in IST-C and MosMedData datasets. Specifically, we evaluated the 2D approach with or without using the attention module and using simple averaging or the LSTM architecture for combining slide-level features/predictions. For the 3D approach, we compared using a single channel as input (the masked CT scan), two-channel (CT scan and segmentation masks separately). Only the best configurations were evaluated for MosMedData due to long training times needed.

For MosMedData, the systems were trained *only* on the training portion of MosMedData to separate COVID-19 positive samples from the Normal class and tested on the MosMedData test portion, with results given in Table 2.5. For IST-C dataset, the systems were pretrained with all of the 1,110-sample MosMedData and finetuned on the IST-C training set.

The state-of-art results from the literature are also included whenever available Jin et al. (2020), He et al. (2021). We have also implemented DeCovNet Wang et al. (2020), that our 3D approach is based on, using the code supplied by the authors[2], following the same training procedure used for our 3D model.

Considering the results given in Tables 2.4 and 2.5, we see that the best 2D and 3D approach have the same accuracy on the IST-C datasets (87.20%), while the 3D system is slightly better for the MosMedData dataset (93.24% vs 91.89%) and slightly better in AUC score in both datasets. However it should be noted that training was faster for the 3D dataset per epoch thanks to the Python environment (vs. Matlab) and the smaller network afforded longer training times (200 vs 50).

The 2D system on the other hand can be said to be more explainable, since it is

[2] https://github.com/sydney0zq/covid-19-detection

Model	**Accuracy (%)**	**AUC**
2D - Base Network + Averaging	80.80 ± 4.88	0.87
2D - Base + Attention + Averaging	85.60 ± 4.35	**0.90**
2D - Base + Attention + LSTM	**87.20 ± 4.14**	0.89
3D - DeCoVNet Wang et al. (2020)	78.00 ± 5.14	0.78
3D - single channel - interpolation	82.80 ± 4.68	0.86
3D - single channel - skipping	**87.20 ± 4.14**	**0.90**
3D - two channels - skipping	81.45 ± 4.82	0.86
Ensemble - Averaging (IST-CovNet)	**90.80 ± 3.58**	**0.95**

Table 2.4 Performance results for the IST-C test set with $n = 250$ samples from 3 classes. The 2D systems are trained with only IST-C and the 3D systems were trained with MosMedData and IST-C training subsets. DeCoVNet results are obtained with author supplied code. Bold figures indicate the best accuracy in slice-based or volume-based approaches.

Model	**Accuracy (%)**	**AUC**
Jin et al. (2D)	-	**0.93**
He et al. (3D)	82.29	-
3D - DeCoVNet Wang et al. (2020)	**82.43**	0.82
2D - Base + Attention + Averaging	90.09 ± 3.93	**0.96**
2D - Base + Attention + LSTM	**91.89 ± 3.59**	0.95
3D - single channel - skipping	**93.24 ± 3.30**	**0.96**
Ensemble - Averaging (IST-CovNet)	**93.69 ± 3.20**	**0.99**

Table 2.5 Performance results for the MosMedData Morozov et al. (2020) test set with $n = 222$ samples from only 2 classes (COVID-19 and Normal). Our approaches are trained using only MosMedData training subset. DeCoVNet results are obtained with author supplied code. Bold figures indicate the best results in the literature and among our two different approaches.

possible to view slice-level decisions to identify where COVID-19 infection patterns are detected by the system; this information can be displayed to the attending physicians in the deployed system.

Model	Accuracy (%)	AUC
COVID-FACT Heidarian et al. (2020)	**91.83**	-
CT-CAPS Heidarian et al. (2020)	89.80	**0.930**
Deep-CT-Net Dialameh et al. (2020)	86.00	0.886
2D - Base + Attention + Averaging	75.41 ± 4.84	**0.838**
2D - Base + Attention + LSTM	**79.34 ± 4.55**	0.819
3D - single channel	**87.87 ± 3.67**	**0.931**
Ensemble Averaging (IST-CovNet)	**90.16 ± 3.35**	**0.942**

Table 2.6 Inter-operability results using the COVID-CT-MD Afshar et al. (2020) dataset with $n = 305$ samples from 3 classes. Our ensemble system was trained using *only* MosMedData and IST-C datasets to measure the inter-operability of the developed system. Bold figures indicate the best results in the literature and among our approaches.

2.9.2 Comparison to the Results in Literature

Our best results obtained on the IST-C dataset is 90.80% accuracy and 0.95 AUC score with ensemble averaging of the best 2D and and best 3D system (Table 2.4).

The results obtained on the MosMedData dataset with only COVID-19 and Normal classes are better as expected (93.69% accuracy and 0.99 AUC), given the relatively simpler problem with two classes (Table 2.5). In comparison to the best results in the literature, our ensemble accuracy (93.69%) is 10% points higher compared to the state-of-art and the AUC score (0.99) is also very high, exceeding the state-of-art.

2.9.3 Evaluating Novel Sub-Modules

Considering the results in Tables 2.4 and 2.5, we see that the attention layer in the 2D approach increases the accuracy significantly (85.60% vs 80.80% in IST-C), in line with other problems where attention brings performance increase in literature.

The use of LSTM to obtain the patient-level predictions from slice-level features brings another 1-1.2% points improvements in accuracy, for both IST-C and MosmedData, compared to averaging the slice-level predictions. The CT sequence size normalization in LSTM training is an important aspect for this improvement. On the other hand, the LSTM achieves lower AUC scores compared to averaging; we expect that this is due to LSTM outputs being close to 0 or 1.

For the 3D approach, we observed that the 2-channel input also used in DeCoV-

Net achieves significantly lower accuracy (81.45% vs 87.20%), probably due to the difficulty in training the first layer weights and the success in obtaining good segmentation masks.

The model trained with the author supplied DeCoVNet Wang et al. (2020) also achieved lower results compared to our extended version (78.00% vs 81.45% for the two-channel system which is basically the same as DeCovNet except for the added skip connections), showing the benefits of extending the network to deal with the rich information present in the CT scan.

Additionally, we found that the interpolation done to halve the large CT volume in the case of the IST-C dataset, leads to significantly lower performance (%87.20 vs %82.80) compared to skipping every other slice, presumably due to the loss of the fine details in the images. This is something to be aware of when dealing with this or similar problems, as interpolation is commonly used in many biomedical applications.

2.9.4 Inter-Operability

To study the inter-operability of systems with respect to different datasets collected from different patient populations and tomography equipment and settings, we tested the accuracy of the systems trained using the MosMedData and IST-C datasets, on the COVID-CT-MD dataset Afshar et al. (2020). As the COVID-CT-MD dataset was not used in training at all, we used the whole dataset for testing. Hence our results are obtained on the whole dataset, while others are obtained on the testing portion of the dataset. COVID-CT-MD dataset comprises 305 CT scans from 3 classes, as indicated in Table 2.2.

The results shown in Table 2.6 accuracy and AUC results (90.16% and 0.9418) are in line with results reported in literature, even though our systems were not trained or finetuned at all for this dataset. In particular, the AUC of the ensemble is highest and accuracy value is slightly behind the best reported results in literature for this dataset Heidarian, Afshar, Enshaei, Naderkhani, Oikonomou, Atashzar, Fard, Samimi, Plataniotis, Mohammadi & Rafiee (2020).

Furthermore, while the results are not directly comparable, our results on COVID-CT-MD dataset show only a slight decrease compared to the IST-C dataset results (90.80% accuracy and 0.95 AUC vs 90.16% accuracy and 0.942 AUC), indicating the generality of the proposed system.

Actual \Predicted	Covid-19	Non-Covid-19
Covid-19	**91**	9
Normal	9	**91**
Other	5	**45**

Table 2.7 Confusion matrix for the IST-C dataset.

2.9.5 Error Analysis

The confusion matrix of the ensemble that obtained 90.8% accuracy on the IST-C dataset (2.4) is given in Table 2.7. The system predicted 9 false negatives (9/100 COVID-19 samples) and 14 false positives (9/100 Normal and 5/50 Other samples) in total. Hence the error rates were almost the same in each group.

An analysis of the errors by expert physicians revealed that the majority (6/9) of the false negatives were due to minimal lung involvement or respiratory motion artifacts. Respiratory motion artifacts were also observed alone or with atelectasis in 4/9 false positives with normal parenchyma.

2.9.6 Prediction Scores Distribution

The system is designed to alert the attending physicians in case of sufficiently high COVID-19 probability. Hence, we also considered the COVID-19 prediction distribution of the ensemble, shown in Figure 2.7. An adjustable threshold (e.g. 0.3-0.4) can be set to alert the attending physician, at the risk of some increased False positives.

At 0.3 threshold, we obtain 95.0% sensitivity (true positive rate) and 80.0% specificity (1-false positive rate) on the IST-C test data set. ROC figures corresponding to IST-C and MosMedData datasets are given in Figure 2.8.

2.9.7 Lung Segmentation Results

Regarding lung segmentation accuracy, Hofmanninger et al. Johannes et al. (2020) report 97-98% Dice similarity scores measuring how much the mask generated by U-Net and ground-truth overlaps, on different test datasets involving multiple lung

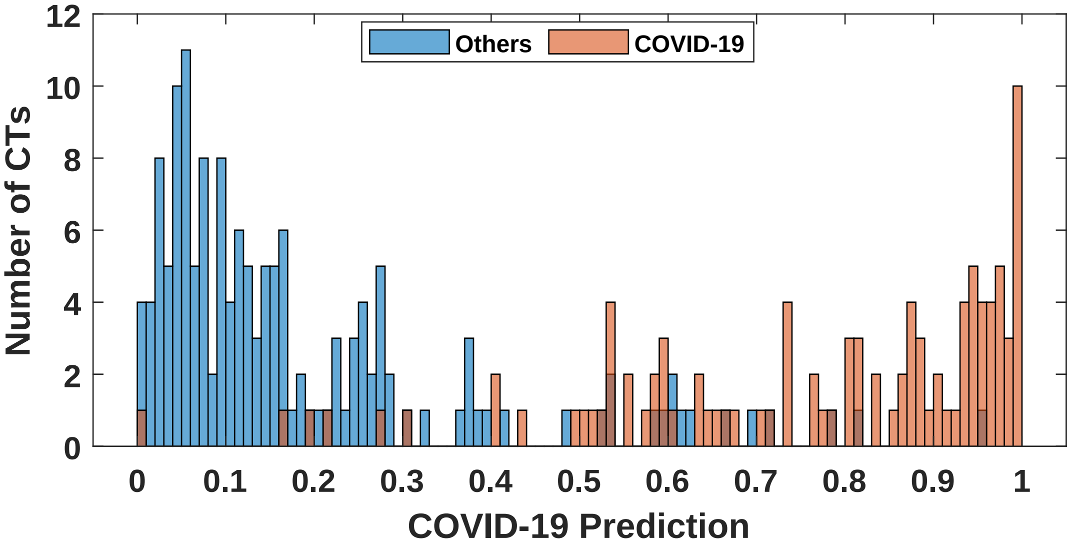


Figure 2.7 COVID-19 predicted probability distribution for the IST-C dataset, using the ensemble.

pathologies. While their tested datasets also included ground glass opacities observed in COVID-19 cases, we evaluated the segmentation network's performance specifically for the COVID-19 detection problem by visually checking the segmentation results of 5 slices from sampled at regular intervals from 1,156 CT scans (all covid patients from IST-C and MosMedData datasets), for a total of 5,783 slice images. We found around 11 serious segmentation errors, corresponding to roughly %0.19, which is in line with Johannes et al. (2020). Samples of these images are given in 2.9, where lung areas that are considered as background and are highlighted by ellipses. Noting that the errors occur only in some of the slices within one CT scan, we conclude that U-Net provides a successful segmentation, suitable for COVID-19 detection.

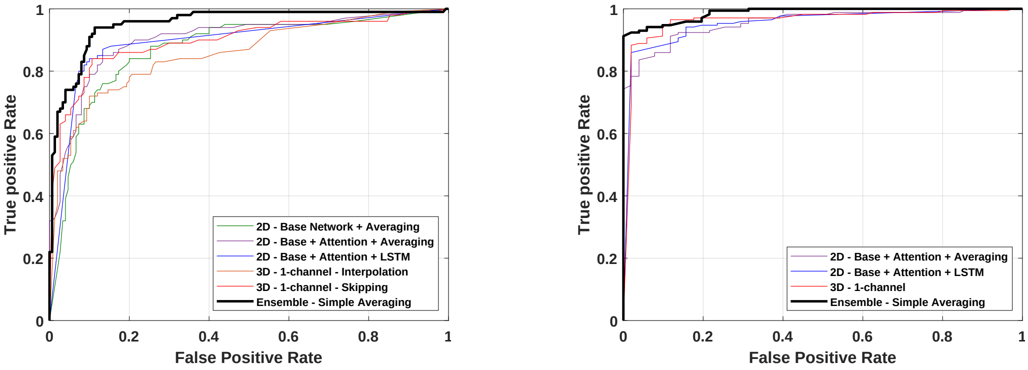


Figure 2.8 ROC curves of the trained models on (a) IST-C dataset and (b) MosMedData dataset.

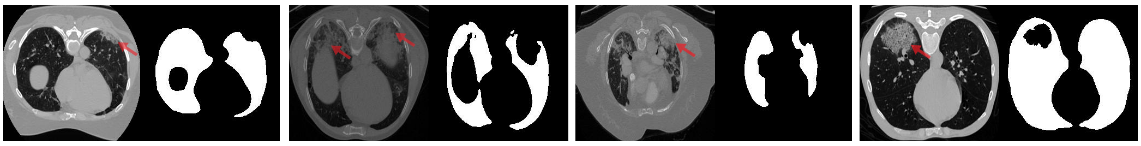


Figure 2.9 Samples of segmentation errors (a) slice image (b) corresponding lung masks. Problematic areas are indicated with red arrows and are often missed lung tissue due to infection or tumors.

2.9.8 Discussion

While our 3D approach is based on DeCoVNet Wang et al. (2020), we were able to outperform its results on both datasets, thanks to the changes made to the model. In particular, using only one input channel leads to more efficient training, especially since the U-Net lung segmentation is very accurate and enriching the network architecture also contributed to higher accuracy.

Similarly, even though the 2D system is based on fine-tuning a pretrained deep network, the use of the novel attention mechanism and LSTMs to combine slice-level features bring significant improvements over the base network and the standard approach of averaging slice predictions. We are aware of only one other work that also combines a deep network with LSTMs, related to COVID-19 predictions: Hammoudi et al. Hammoudi et al. (2020) use bidirectional LSTMs to predict patient health status by combining the predictions made by a deep network for *image-patches* of an X-ray.

Considering the results in Table IV, we see that our contributions improve accuracies by 6.4 and 9.20 percentage points, in 2D and 3D models respectively (%87.20 vs %80.80 and %87.20 vs %78.00). Furthermore, we gain another 3.6 percentage points when we combine the 2D and 3D systems (%90.80 vs %87.20). Hence, while the main contributions of our work are in the network architectures, the ensemble approach also brings significant improvements.

2.10 Conclusion and Future Works

In addition to presenting a state-of-art system, we provide an evaluation of different 2D and 3D approaches on two datasets and discuss the effects of relevant preprocessing, segmentation and classifier combination steps on performance. A third large and public dataset is used to show inter-operability results.

The collected dataset (IST-C) is made public to contribute to the literature as a challenging new dataset that consists of high resolution chest CT scans from a variety of conditions.

This work was motivated to help combat the pandemic and the developed system (IST-CovNet) is deployed and in use at Istanbul University Cerrahpaşa School of Medicine, to flag suspected COVID-19 cases when the patient is still at the tomography room.

For future works, the research could include the integration of 3D classifiers, taking advantage of the wealth of data from radiology departments for pre-training. These 3D classifiers may yield more precise results due to the three-dimensional nature of the lung structures. To maximize the use of unlabeled data, innovative self-supervised learning methods could be employed.

CHAPTER 3

VCL-PL: SEMI-SUPERVISED LEARNING FROM NOISY WEB DATA WITH VARIATIONAL CONTRASTIVE LEARNING

One of the main contributions of this thesis is the introduction of a novel variational supervised contrastive learning approach. In this chapter, the proposed approach is used in leveraging noisy web-collected data in the face attribute classification problem, within the semi-supervised learning framework. In Chapter 4, the same general approach is extended and applied in a self-supervised manner.

Web data suffers from image set noise, due to unrelated images that may be retrieved in response to the query. We propose a semi-supervised pseudo-labeling approach where the embedding space distribution is learned via variational contrastive learning.

For addressing the multi-label face attribute classification problem, we use 40 Gaussian sampling heads for the 40 attributes in the CelebA dataset and apply supervised contrastive learning over a limited amount of labelled data. Soft pseudo-labeling is then used to label the unlabelled data at attribute level, followed by two-stage domain adaptation.

We show that the proposed method using noisy web data brings improvements in accuracy over supervised multi-label face attribute classification in all experimental settings (over 2% points for very low-data settings). We suggest that learning the embedding distribution and the subsequent soft pseudo-labeling according to the nearest neighbors help in overcoming the noise in the unlabeled data.

This work is published in the International Conference in Pattern Recognition (Yavuz & Yanikoglu (2022)).

3.1 Introduction

Unsupervised and semi-supervised learning paradigms are expected to have a great potential for progress in machine learning, as it is possible to collect images, audio or video from nearly limitless data sources on Internet. For example, a web search can be used to collect images to be used in training a visual concept. Unfortunately, the weakly labelled data found on the web in response to the query, often contains large amounts of irrelevant or noisy images. In the domain of face images, a particular Internet search may return images that are unrelated or that only loosely correspond to the query (e.g. images of makeup for "rosy cheek"). In this paper, we propose a semi-supervised learning approach and evaluate its performance on classifying the 40 face attributes depicted in the CelebA dataset, using the internet as the source of the unlabelled data.

Several different approaches are suggested in the literature to leverage unlabelled data. Among these, we can distinguish two broad categories. In the first category, we see unsupervised or self-supervised methods that are used to learn good feature representations. Among these approaches, one group of algorithms including including SimCLR Chen et al. (2020), Context EncodersPathak, Krahenbuhl, Donahue, Darrell & Efros (2016), Self-Augment Reed, Metzger, Srinivas, Darrell & Keutzer (2021), Deeppermnet Santa Cruz, Fernando, Cherian & Gould (2017), Clusterfit Yan, Misra, Gupta, Ghadiyaram & Mahajan (2020), use a *pretext* task to learn features using self-supervision Bojanowski & Joulin (2017); Caron, Bojanowski, Mairal & Joulin (2019); Chen, Liu & Jia (2021); Chen, Kornblith, Swersky, Norouzi & Hinton (2020); Chen, Zhai, Ritter, Lucic & Houlsby (2019); Doersch, Gupta & Efros (2015); Feng, Xu & Tao (2019); Jenni & Favaro (2018); Kim, Cho, Yoo & Kweon (2018); Kolesnikov, Zhai & Beyer (2019); Larsson, Maire & Shakhnarovich (2016,1); Lee, Hwang & Shin (2020); Lee, Huang, Singh & Yang (2017); Minderer, Bachem, Houlsby & Tschannen (2020); Misra & Maaten (2020); Mundhenk, Ho & Chen (2018); Noroozi, Pirsiavash & Favaro (2017); Wu, Xiong, Yu & Lin (2018); Yang, Parikh & Batra (2016); Zhai, Oliver, Kolesnikov & Beyer (2019); Zhang, Isola & Efros (2016). Another group of algorithms, including such as Deep Cluster Caron, Bojanowski, Joulin & Douze (2018), ClusterGANMukherjee, Asnani, Lin & Kannan (2019), SCAN Van Gansbeke, Vandenhende, Georgoulis, Proesmans & Van Gool (2020), aim to learn good feature representations that lead to *good clusters* Bo, Wang, Shi, Zhu, Lu & Cui (2020); Dang, Deng, Yang, Wei & Huang (2021); Figueroa & Rivera (2017); Guo, Zhu, Liu & Yin (2018); Song, Liu, Huang, Wang & Tan (2013); Van Gansbeke et al. (2020); Yang et al. (2016); Yang, Cheung, Li &

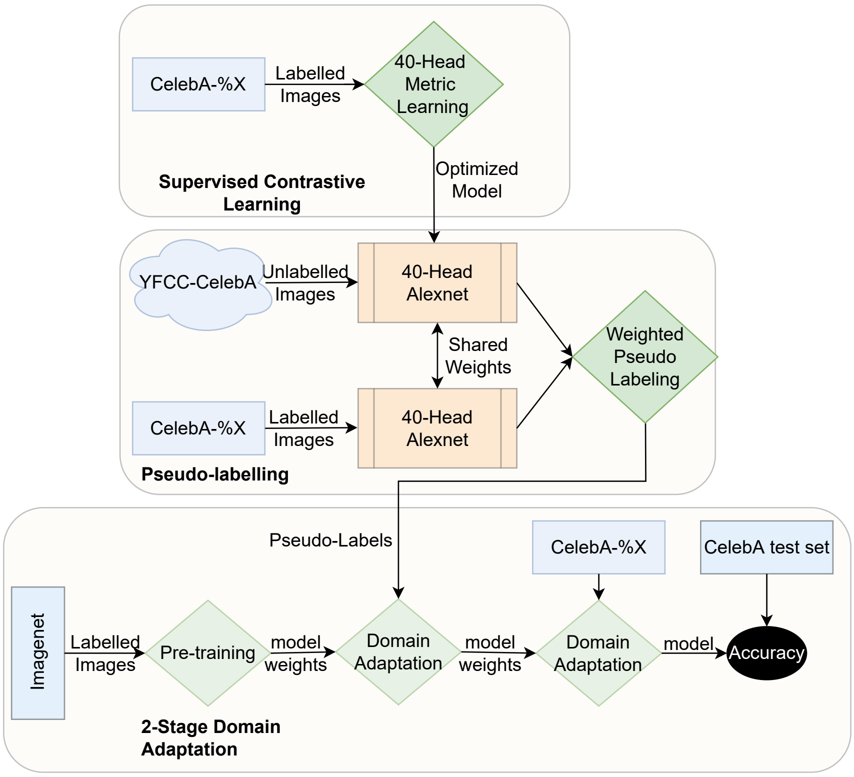


Figure 3.1 Proposed pipeline.

Fang (2019).

In the second category, there are semi-supervised approaches, such as MixMatch Berthelot, Carlini, Goodfellow, Papernot, Oliver & Raffel (2019), FixMatchSohn, Berthelot, Carlini, Zhang, Zhang, Raffel, Cubuk, Kurakin & Li (2020) and FlexMatchZhang, Wang, Hou, Wu, Wang, Okumura & Shinozaki (2020), and others Berthelot et al. (2019); Nassar, Herath, Abbasnejad, Buntine & Haffari (2021); Pham, Xie, Dai & Le (2021); Rizve, Duarte, Rawat & Shah (2020); Xu, Shang, Ye, Qian, Li, Sun, Li & Jin (2021) that use pseudo-labeling or self-labeling, where unlabelled data is assigned *pseudo-labels*. Generative or teacher-student or ensemble models can also be listed among the semi-supervised approaches Dai, Yang, Yang, Cohen & Salakhutdinov (2017); Feng, Kong, Chen, Zhang, Zhu & Chen (2021); Ke, Wang, Yan, Ren & Lau (2019); Li, Xu, Liu, Zhu & Zhang (2021); Miyato, Maeda, Koyama & Ishii (2018); Rasmus, Berglund, Honkala, Valpola & Raiko (2015); Salimans, Goodfellow, Zaremba, Cheung, Radford & Chen (2016); Sellars, Avilés-Rivero & Schönlieb (2021); Tarvainen & Valpola (2017); Wang, Li & Gool (2019); Wang, Kihara, Luo & Qi (2021).

Among the first category, SimCLR Chen et al. (2020) and SimCLRv2 Chen et al. (2020) are the current state-of-the-art self-supervised methods, based on the contrastive learning approach where learning aims to reduce the distance between the embedded representations of the two augmentations by the same image. The effectiveness of these algorithms have been demonstrated on non-noisy benchmark datasets such as Imagenet Deng, Dong, Socher, Li, Li & Fei-Fei (2009), CIFAR10 and CIFAR100 Krizhevsky (2009); however their applicability to multi-label data and noisy web images are yet unaddressed issues. Another successful algorithm is SCAN Van Gansbeke et al. (2020), which aims to form semantic clusters, by using a multi-step learning scheme that starts with the pretext task of SimCLR and continues with novel clustering loss functions.

Our aim in this paper is to increase the accuracy of the existing multi-label face recognition systems by using the visual data collected from the Internet. To this end, we propose a semi-supervised algorithm called *VCL-PL*, consisting of (i) a representation learning step using supervised variational contrastive learning, inspired by variational auto-encoders Kingma & Welling (2013); (ii) a pseudo-labeling step based on the nearest neighbor mining used in Van Gansbeke et al. (2020); and a domain adaptation step where the general deep features learned using ImageNet is adapted to the target domain in two-steps. The algorithm is illustrated in Figure 3.1.

The feature learning component of the proposed method resembles SimCLR Chen et al. (2020), but differs from it by the variational approach that aims to learn the underlying distribution of the latent space. Furthermore, unlike SimCLR, we apply contrastive learning to a fraction of the labelled data and construct a separate embedding space for each attribute in order to address the multi-attribute classification, which would not have been possible with unlabelled data.

The pseudo-labeling component is inspired by the SCAN Van Gansbeke et al. (2020) and SPICE Niu, Shan & Wang (2021) algorithms that use neighborhood mining in the embedding space, but we use a distance weighting and obtain soft pseudo-labels.

For repeatable experiments, we use the YFCC100M dataset as the data collected from the Internet and the YFCC-CelebA subset obtained by filtering YFCC100M with keywords related to the 40 facial attributes present in CelebA Yavuz et al. (2021). Note that YFCC100M is an uncurated dataset with only weak labels and is used without labels in this work.

We demonstrate the effectiveness of the proposed algorithm by using varying amounts of labelled data from the CelebA dataset (%1,%10, or %100) and the YFCC-CelebA dataset as the unlabeled dataset. Our main contributions are learning the embedding space distribution using a variational approach and extending the contrastive learning framework to multi-label problems by using 40 Gaussian heads and a limited amount of labelled data.

Our system also benefits from a weighted nearest neighbor pseudo-labelling, as well as a two-step domain adaptation.

The paper is structured as follows. In Subsection 3.2.1 and 3.2.2, we discuss the backbone network and the Gaussian sampling heads and the supervised metric learning with the variational approach. In Section 3.2.3 and 3.2.4, the pseudo-labeling algorithm and the two-stage domain adaptation are presented, respectively. Last two sections are the Experimental Evaluation and the Conclusion sections.

3.2 Methodology

The proposed algorithm has three consecutive stages and is illustrated in Figure 3.1.

- Supervised Contrastive Metric learning (Section 3.2.1 and 3.2.2). We use the available labelled data (%1 or %10 or %100 of CelebA) and apply contrastive metric learning in a supervised fashion, to learn each of the 40 embedding spaces.
- Nearest neighborhood based weighted pseudo-labeling of the noisy web data (Section 3.2.3).
- Domain adaptation of the backbone network in two stages. We fine-tune the Imagenet pretrained Alexnet network using the pseudo-labeled YFCC-CelebA and then apply a second domain adaptation with the avaliable labelled CelebA subset (Section 3.2.4).

The supervised metric learning and weighted pseudo-labeling is accomplished in the multi-label domain of face attributes (each image has 40 face attribute labels) with Gaussian embeddings.

3.2.1 Feature Extraction and Gaussian Sampling

In this step of the proposed method, we use the labelled set to learn useful embedding distributions, separately for each binary attribute label. The backbone feature

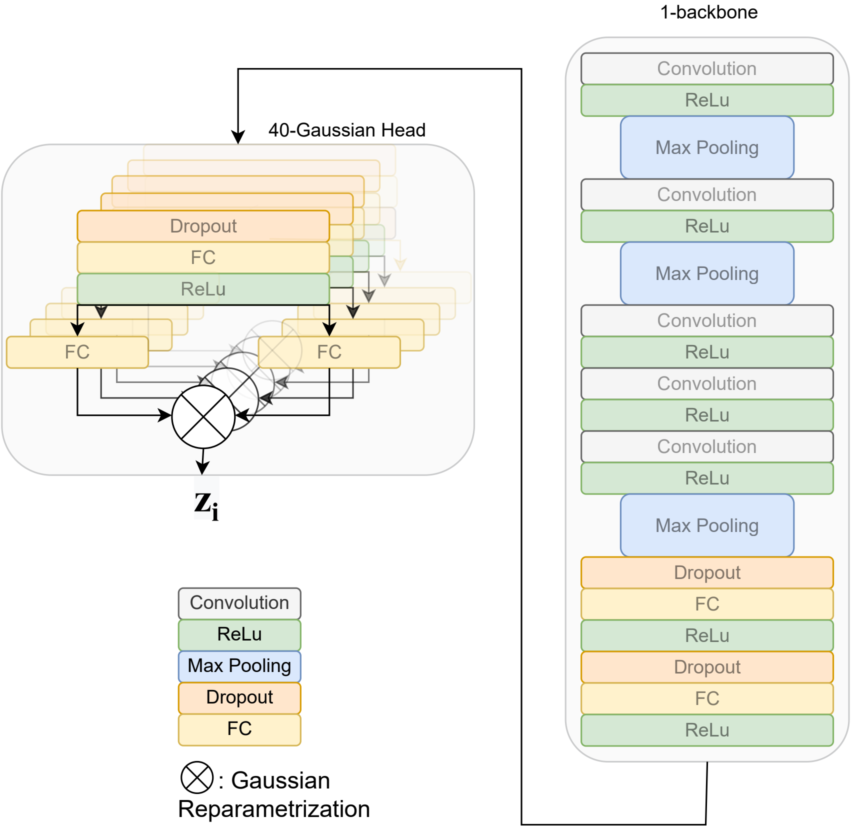


Figure 3.2 Learning embeddings: AlexNet is used as the backbone with 40 Gaussian heads for sampling.

extractor is a standard convolutional network, which is followed by sampling heads, as shown in Figure 3.2.

An input $\mathbf{x}$ undergoes a stochastic transformation t and is then passed through a feature extractor network that extracts the embedding representation f_θ.

The feature extractor is followed by Gaussian sample heads (g_W) that outputs the parameters of the distribution of the learned embedding space. The process is explained in Eq. 1:

$$(\boldsymbol{\mu}, \log \boldsymbol{\sigma}^2) = g_W(f_\theta(t(\mathbf{x}))) \tag{3.1}$$

We then sample from this distribution using the parametrization trick as used in variational autoencoders Kingma & Welling (2013):

$$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma}^2 \odot \boldsymbol{\xi} \tag{3.2}$$

where $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0},,\mathbf{I})$ and $\odot$ denotes element-wise multiplication. The parametrization trick enables the use of backpropagation despite the sampling process.

The network that learns the embeddings consists of two blocks, shown in Figure 3.2. The backbone is the feature extractor network and its the output vector is shared between 40 Gaussian heads. In our implementation the backbone network is Alexnet architecture Krizhevsky (2014) with the dropped FC and softmax layers. The activation vectors obtained from the last layer of Alexnet are 4096 dimensional (for 224x224 pixels input) and shared by 40 Gaussian sampling heads which corresponds to 40 attributes of CelebA.

Each embedding space is modelled by a 128-dimensional multi-variate Gaussian distribution with diagonal covariance matrix and sampled with a Gaussian sampling head that has a non-linear layer followed by a linear layer to get the deterministic values of mean and variance. The output of mean and variance embeddings are the inputs for Gaussian reparametrization. This view corresponds to one branch of contrastive network shown in Figure 3.3.

3.2.2 The Supervised Variational Contrastive Learning

A simple contrastive learning algorithm, based on reducing the distance between augmentations of the same image, is run in each 40 embedding spaces independently, with an objective function consisting of three terms, explained below. The algorithm for the variational contrastive learning is given in Alg. 1 and illustrated in Fig. 3.3.

Large Margin Cosine Loss. The first loss terms is the Large Margin Cosine Loss (LMCL) Wang, Wang, Zhou, Ji, Gong, Zhou, Li & Liu (2018) whose effectiveness has been demonstrated in comparison to softmax Chong & Zak (2004), center loss Wen, Zhang, Li & Qiao (2016), large margin softmax loss Liu, Wen, Yu & Yang (2016), and angular loss Wang, Zhou, Wen, Liu & Lin (2017) in face recognition domain. Given an input x_i with binary label y_i, LMCL is derived starting from the cross-entropy loss, requiring the weights and input to have unit norm and using the large margin formulation. Specifically, given input x_i and the corresponding ground-truth, y_i:

$$L_{xent} = \frac{1}{N}\sum_{i=1}^{N} -\log p_{y_i} = \sum_{i=1}^{N} -\log \frac{e^{f_{i,y_i}}}{\sum_{j=1}^{K} e^{f_{i,j}}} \tag{3.3}$$

where N is the number of training samples, p_{y_i} is the posterior probability of the correct class and $f_{i,j}$ is the output of the jth class for the i-th input sample. Denoting the weight vector of the j-th output node as W_j, we have:

$$f_{i,j} = W_j^T x_i = ||W_j||||x_i|| \cos\theta_{ij} \tag{3.4}$$

Then, using normalized weight and input vectors, the cosine loss is derived first. Finally, LMCL is obtained with the large margin formulation:

$$\mathcal{L}_{LMC} = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{e^{\{s(\cos\theta_{ij}-m)\}}}{e^{\{s(\cos\theta_{iy_i}-m)\}} + \sum_{j\neq y_i} e^{\{s(\cos\theta_{ij})\}}} \tag{3.5}$$

where θ_{ij} is the angle between W_j and x_i; s is a constant and m is the margin parameter.

Distribution Similarity Loss. The second loss term encourages the augmentations of the same image (x_i, x_j) being drawn from similar distributions q and p respectively, by penalizing the divergence between the two using the Kullback-Leibler divergence Odaibo (2019).

$$\begin{aligned} \mathcal{L}_S &= -\frac{1}{N}\sum_{i=1}^{N} D_{KL}(q_1(z_i|x_i)||q_2(\tilde{z_i}|\tilde{x_i})) \\ &= -\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{\sigma_{q_1,i}}{\sigma_{q_2,i}}\right) - \frac{\sigma_{q_1,i}^2 + (\mu_{q_1,i} - \mu_{q_2,i})^2}{2\sigma_{q_2,i}^2} + \frac{1}{2} \end{aligned} \tag{3.6}$$

Distribution Normalizing Loss. This loss encourages the learned distributions to have zero mean and unit variance, as per Odaibo (2019).

$$\mathcal{L}_D = -\frac{1}{N}\sum_{i=1}^{N} D_{KL}(q_\theta(z_i|x_i)||\mathcal{N}(0,1)) \tag{3.7}$$

$$= -\frac{1}{N}\sum_{i=1}^{N} \frac{1}{2}\left[1 + \log(\sigma_{q_i}^2) - \sigma_{q_i}^2 - \mu_{q_i}^2\right] \tag{3.8}$$

(3.9)

Total Loss. The embedding representations are learned for each binary face attribute, using a portion of the labelled dataset. The optimization is done based on the *total* loss:

$$\mathcal{L}_{total} = \frac{1}{40}\sum_{att=1}^{40} \{\mathcal{L}_{LMC} + \mathcal{L}_S + \mathcal{L}_D\}_{att} \tag{3.10}$$

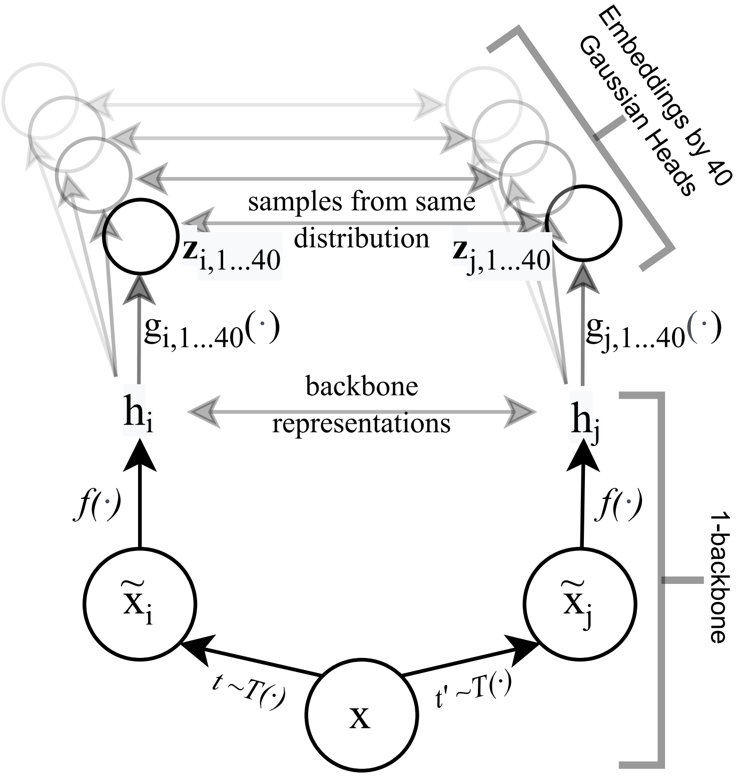


Figure 3.3 Variational Contrastive Learning. Two random augmentations of an image x are input to the network to obtain the backbone representations, followed by the 40-Gaussian sampling heads. The two samples are then compared to reduce the total loss (Eq. 8).Figure inspired by SimCLR Chen et al. (2020)

.

3.2.3 Weighted Pseudo-labeling

We use the k-nearest neighbor (k-NN) algorithm to pseudo-label the elements of the unlabeled YFCC-CelebA dataset, by the labels of their closest neighbor(s) in the CelebA subset.

For an unlabelled image u, we find the k nearest neighbors in the labelled dataset and obtain the confidence-weighted pseudo-label, according to the labels and distance of each neighbor:

$$pseudoLabel(u) = \sum_{i=1}^{k} \frac{\{label_i * e^{-d_i}\}}{k} \tag{3.11}$$

where d_i is the distance from u to nearest neighbor i with label $label_i \in \{-1,+1\}$.

The pseudo-labels are normalized into the $[-1,1]$ range after the pseudo-labelling process. Note that an image can be confident in some labels and less confident in some others. We give the algorithm for $k=1$ in Alg. 2, as it gave the best results.

Algorithm 1 Contrastive learning design for supervised metric learning, as in Chen et al. (2020).

```
input: batch size $N$, networks $f$ and $g$, and augmentation function distribution $T$
for each sampled minibatch $\{\mathbf{x}_k\}_{k=1}^{N}$ do
    for each image $k \in \{1, ..., N\}$ do
        Draw two augmentation functions $t \sim T$, $t' \sim T$
        $\tilde{\mathbf{x}}_{2k-1} = t(\mathbf{x}_k)$ # first augmentation
        $\mathbf{h}_{2k-1} = f(\tilde{\mathbf{x}}_{2k-1})$ # its representation

        for $c \in \{1, ..., 40\}$ do
            $\mathbf{z}_{2k-1,c} = g(\mathbf{h}_{2k-1,c})$ # first sample

        end
        $\tilde{\mathbf{x}}_{2k} = t'(\mathbf{x}_k)$ # second augmentation
        $\mathbf{h}_{2k} = f(\tilde{\mathbf{x}}_{2k})$ # its representation

        for $c \in \{1, ..., 40\}$ do
            $\mathbf{z}_{2k,c} = g(\mathbf{h}_{2k,c})$ # second sample

        end
    end
    Compute loss for each different head using Eq. 3.10
    Update the networks $f$ and $g$ to minimize $\mathcal{L}_{\text{total}}$
end
return: Base and sampler networks $f(\cdot)$ and $g(\cdot)$.
```

3.2.4 Two-Step Domain Adaptation

The domain adaptation is done in two steps. The Imagenet pretrained network is fine-tuned with the pseudo-labelled YFCC-CelebA set first; and then to the labelled CelebA set.

We have found doing the adaptation in two steps brings roughly 1% point in accuracy, compared to a single step adaptation (either directly to CelebA or with both data sets combined.

3.3 Experimental Evaluation

We evaluate the proposed approach on the problem of classifying the 40 facial attributes in the CelebA dataset and compare its performance to : i) standart supervised learning where we fine-tune the ImageNet pretrainet network with the available labelled dataset; ii) DeepCluster Caron et al. (2018); iii) SimCLR Chen et al. (2020); iv) SCAN Van Gansbeke et al. (2020) v) CL-PL (which is the proposed system, only lacking the variational component) vi) VCL-PL (proposed system).

The experiments are run with a portion (%100, %10 or %1) of the CelebA dataset being used as the labelled dataset and YFCC-CelebA dataset as the unlabelled dataset. We used AlexNet Krizhevsky (2014) in all of the experiments, for simplicity.

Datasets As labelled data, we use the CelebA dataset which is resized and cropped into 128 by 128 pixels, along with its ground truth labels. As unlabelled data, we use a subset of the Yahoo Flickr Creative Commons 100 Million Dataset (YFCC100M), which is the largest public multimedia collection that have approximately 99.2 million are photos and 0.8 million are videos. The subset YFCC-CelebA Yavuz et al. (2021) consists of approximately 1 million photos that are found when searching in English for the 40 face features that exist in the CelebA set (“attractive“, “eyeglasses“ etc). In addition to the face attribute words, the word "face" was added in these searches (e.g. "chubby face").

CelebA has 40 face attributes, but it is also possible to express the opposite concept when the attribute is an adjective, which was determined using ConceptNet Speer, Chin & Havasi (2017)). The opposite concept was then obtained from Wiktionary Zesch, Müller & Gurevych (2008) and used in enriching the query (e.g. "wide eyes" along with "narrow eyes"). The search and downloads during this process were done automatically. When multiple queries returned the same photo, repetitions were eliminated.

A total of 392K *non-repetitive* images were obtained using 58 query words obtained by the above process and after eliminating low resolution images. As a last operation, the images were aligned and scaled similar to CelebA. For this, the photos are padded from the edges so as to center the faces and then scaled to obtain 128×128 images.

Image Transformation For the augmentations needed in the contrastive metric learning task, we use Resize, Crop, Horizontal Flip, Grayscale, Color Jitter augmentations and sample a random transformation as a combination of these augmentations within the allowed parameter range. The stochastic data augmentation consists of resizing (scale between [0.2, 1.0]), cropping (128 random crops), grayscale transformation (with probability 0.2), and color jitter (with probability 0.8, brightness in [0.6,1.4], contrast in [0.6,1.4], saturation in [0.6,1.4] and hue in [0.9,1.1]).

Training Details. We use AlexNet for the feature extractor (backbone) part of the network, where the embedded representation is 2304 dimensional (for 128x128

Algorithm 2 Weighted Pseudo-labeling

```
Data: $D_{\text{Labeled}}$, $D_{\text{Unlabeled}}$
Initialization: $\{W,\theta\}$ pretrained $\{W^*,\theta^*\}$ Obtain the representation for each $x_l$ in $D_{\text{Labeled}}$
foreach sample $x_u \in D_{Unlabeled}$ do
    Sample augmentation function $t \sim T$
    $\tilde{\mathbf{x}} = t(x_u)$ # an augmentation
    $\mathbf{h} = f(\tilde{\mathbf{x}})$ # representation
    # Gaussian projections in 40-dimensional embedding space
    for $c \in \{1,...,40\}$ do
        $\mathbf{z}_c = g_c(\mathbf{h})$
        (distance, label) = mine $\mathbf{1\text{-}NN}(\mathbf{z}_c)$ in $D_{\text{Labeled}}$
        $pseudoLabel = \text{label} \cdot e^{-\text{distance}}$
    end
    Normalize the labels into the range $[-1,1]$.
end
return: YFCC-CelebA with soft pseudo-labels.
```

pixel input) and the output z of a sample head is 128-dimensional. For training, the network weights are updated using SGD, with a learning rate of 1e-3, momentum coefficient of 0.9 and weight decay of 1e-5. We run pre-training experiments with 400 epochs with Cosine Annihilation Scheduler and fine-tuning experiments with 100 epochs with early validation stopping. The batch size is 128 for pre-training and 64 for fine-tuning. The code is available at https://github.com/verimsu/VCL-PL.

3.4 Results

Table 3.1 gives the comparison between the proposed VCL-PL algorithm to other well-known approaches, as well as the standart supervised training and the CL-PL approach (proposed system, only without the variational component). Here, supervised learning refers to fine-tuning the ImageNet pretrained AlexNet with the available labelled dataset. DeepCluster Caron et al. (2018), SimCLR Chen et al. (2020), and SCAN Van Gansbeke et al. (2020) are well-known, state-of-art unsupervised algorithms that are implemented with the code provided in their official repositories.

We see that VCL-PL outperforms state-of-the-art self-supervised learning schemes and standard supervised learning, for all settings (100% , 10% , 1% of CelebA), showing the effectiveness of the proposed method.

The improvements over the best approach from the literature (SimCLR) are 0.49, 0.93, 0.61% points, respectively for 100%, 10% and 1% settings. Note that the scale of these improvements is on par with those observed between other state-of-art methods.

It is also worth noting that VCL-PL and CL-PL are the only two systems that can outperform supervised training in 100% CelebA settings. Furthermore, VCL-PL consistently outperforms CL-PL, showing the benefit of using the variational approach. The accuracies of four of the methods are plotted in Figure 3.4 for clarity.

In Table 3.2, we observe the k-NN from the pseudo-labeling stage is best for $k = 1$. In fact, the SCAN algorithm Van Gansbeke et al. (2020) also uses 1-NN approach in its nearest neighbor pseudo-labelling. We further observe that that the algorithm benefits from soft labelling (Eq. 9) as compared to using hard labels.

Table 3.1 Average Accuracy for different algorithms and settings. Bold results indicates the best results while underlined results show the best results from the literature

Method	Imagenet	YFCC	CelebA	**Accuracy**
Supervised	yes	no	100%	90.24%
DeepCluster Caron et al. (2018)	no	yes	100%	89.40%
SimCLR Chen et al. (2020)	yes	yes	100%	89.98%
SCAN Van Gansbeke et al. (2020)	yes	yes	100%	89.11%
CL-PL	yes	yes	100%	90.32%
VCL-PL	yes	yes	100%	**90.47%**
Supervised	yes	no	10%	88.65%
DeepCluster Caron et al. (2018)	no	yes	10%	87.53%
SimCLR Chen et al. (2020)	yes	yes	10%	88.75%
SCAN Van Gansbeke et al. (2020)	yes	yes	10%	88.35%
CL-PL	yes	yes	10%	89.43%
VCL-PL	yes	yes	10%	**89.68%**
Supervised	yes	no	1%	85.90%
DeepCluster Caron et al. (2018)	no	yes	1%	84.12%
SimCLR Chen et al. (2020)	yes	yes	1%	87.51%
SCAN Van Gansbeke et al. (2020)	yes	yes	1%	85.85%
CL-PL	yes	yes	1%	87.69%
VCL-PL	yes	yes	1%	**88.12%**

Table 3.2 Average Accuracy for Different k values and using hard labels, at low-data regime

CelebA 1%	1-NN	3-NN	5-NN	Hard labels
Accuracy	88.12%	88.07%	88.03%	87.43%

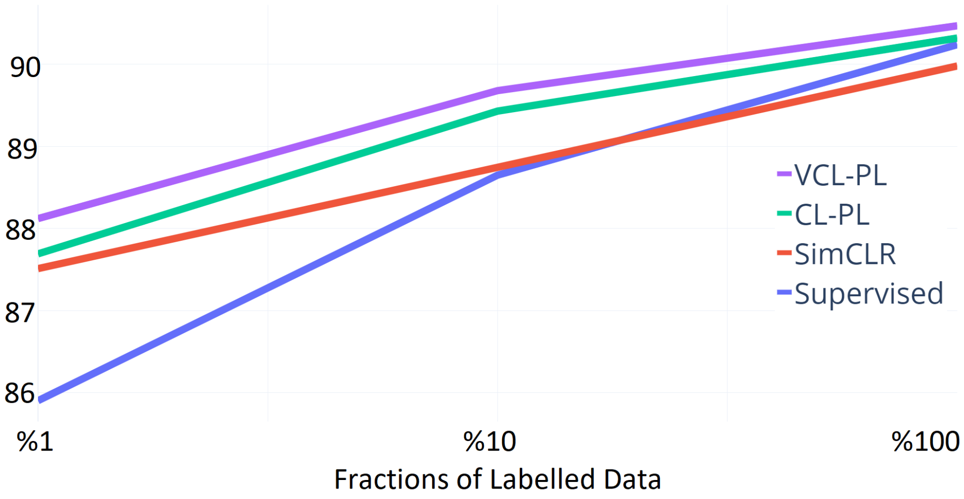


Figure 3.4 Accuracy values for varying fractions of available labelled data.

Table 3.3 Detailed comparison of supervised training and our proposed systems, VCL-PL and CL-PL.

Attributes	Supervised	CL-PL	VCL-PL	Attributes	Supervised	CL-PL	VCL-PL
5 o'Clock Shadow	89.60%	90.56%	91.06%	Male	90.40%	93.54%	93.96%
Arched Eyebrows	76.06%	77.87%	78.58%	Mouth Slightly Open	69.12%	81.15%	82.62%
Attractive	75.59%	77.49%	77.87%	Mustache	96.04%	96.06%	96.07%
Bags Under Eyes	79.13%	81.01%	81.20%	Narrow Eyes	84.92%	84.93%	85.16%
Bald	97.91%	97.83%	97.92%	No Beard	87.36%	90.09%	91.26%
Bangs	91.69%	94.04%	94.37%	Oval Face	70.09%	71.38%	72.12%
Big Lips	67.38%	68.46%	68.93%	Pale Skin	95.67%	95.78%	95.79%
Big Nose	79.52%	80.43%	80.83%	Pointy Nose	70.07%	71.53%	72.33%
Black Hair	81.90%	84.89%	85.84%	Receding Hairline	91.35%	91.56%	91.65%
Blond Hair	93.70%	94.33%	94.53%	Rosy Cheeks	92.74%	92.89%	93.19%
Blurry	95.11%	94.99%	94.94%	Sideburns	95.13%	95.93%	96.30%
Brown Hair	83.69%	84.95%	85.12%	Smiling	76.64%	85.95%	87.46%
Bushy Eyebrows	86.32%	87.27%	87.74%	Straight Hair	77.39%	79.56%	79.98%
Chubby	94.13%	94.09%	94.47%	Wavy Hair	77.49%	79.69%	80.27%
Double Chin	95.38%	95.28%	95.38%	Wearing Earrings	81.26%	83.82%	84.58%
Eyeglasses	95.95%	97.18%	97.31%	Wearing Hat	97.43%	97.93%	97.79%
Goatee	94.88%	95.26%	95.64%	Wearing Lipstick	87.30%	90.45%	90.64%
Gray Hair	96.80%	97.19%	96.99%	Wearing Necklace	85.42%	85.95%	86.30%
Heavy Makeup	84.08%	87.04%	87.66%	Wearing Necktie	95.05%	94.78%	95.20%
High Cheekbones	75.69%	81.74%	82.63%	Young	80.63%	82.58%	83.30%
				Average	85.90%	87.69%	88.12%

We also evaluated other well-known algorithms, namely Semi-supervised Label Propagation Iscen, Tolias, Avrithis & Chum (2019) and MixMatch Berthelot et al. (2019). These methods were observed to underperform compared to supervised learning with the available labelled data. We presume that the main reason for the degradation is that our problem deals with the data noise in the unlabelled set.

3.5 Conclusion and Future Works

We study the problem of improving the performance of existing supervised systems by the use of weakly labelled data collected from the internet. The specific problem addressed in this work is face attribute classification, where we obtained performance improvements over the supervised learning framework (over 2% points for very low-data setting), and two existing baselines (DeepCluster and SimCLR), with the proposed method.

The main contributions are to use a variational approach to learn the underlying distribution of the embedding space and extending the contrastive learning framework to multi-label problems.

As a future work, the expansion of our method to address a broader range of multi-label classification problems is an intriguing prospect. The efficacy of the 40 Gaussian sampling heads approach employed for the CelebA dataset needs to be validated against more complex datasets to determine the generalizability of the method.

CHAPTER 4

VARIATIONAL SELF-SUPERVISED CONTRASTIVE LEARNING USING BETA DIVERGENCE

Chapter 4 presents the development of a robust self-supervised learning model named VCL, which harnesses the combined power of variational contrastive learning and beta-divergence. The chapter offers insights on the self-supervised variational methodologies.

Learning a discriminative semantic space using unlabelled and noisy data remains unaddressed in a multi-label setting. We present a contrastive self-supervised learning model which is robust to data noise, grounded in the domain of variational methods. The model (VCL) utilizes variational contrastive learning with beta-divergence to learn robustly from unlabelled datasets, including uncurated and noisy datasets. We demonstrate the effectiveness of the proposed model through rigorous experiments with multi-label datasets in the face attributes and verification domain. Experiments include linear evaluation and fine-tuning scenarios to show that the model learns effective embedding representations. In all tested scenarios, VCL surpasses the performance of state-of-the-art self-supervised methods, achieving a noteworthy increase in accuracy.

4.1 Introduction

Supervised deep learning approaches require large amounts of labelled data. While transfer learning with pretrained models is commonly used for addressing the labelled data shortage, the recent research focus has been on unsupervised and especially self-supervised methods trained with unlabelled or weakly labelled data that may be collected from the web Cole, Yang, Wilber, Mac Aodha & Belongie (2022); Goyal, Caron, Lefaudeux, Xu, Wang, Pai, Singh, Liptchinsky, Misra, Joulin & others (2021); Tian, Henaff & van den Oord (2021); Zhong, Tang, Chen, Peng & Wang (2022). While it is easy to have access to uncurated data sets collected from the web, these data sets often lack useful labels Li, Wang, Li, Agustsson & Van Gool (2017); Yavuz et al. (2021), making it necessary to develop robust self-supervised learning algorithms to enhance the performance of a visual classifier.

Self-supervised learning utilizes supervisory signals that are generated internally from the data, eliminating the need for external labels. A significant part of self-supervised learning research involves using *pretext tasks* to learn embedding representations that are helpful in downstream tasks. Pretext tasks may involve patch context prediction Mundhenk et al. (2018); solving jigsaw puzzles from the same Noroozi, Vinjimoor, Favaro & Pirsiavash (2018); colorizing images Larsson et al. (2017); Zhang et al. (2016); predicting noise Bojanowski & Joulin (2017); counting Noroozi et al. (2017); inpainting patches Pathak et al. (2016); spotting artifacts Jenni & Favaro (2018); generating images Ren & Lee (2018); predictive coding Van den Oord, Li & Vinyals (2018); or instance discrimination Wu et al. (2018). With any of the above approaches, self-supervised methods seek embedding vectors learned that semantic orientations or clusters.

State-of-the-art self-supervised algorithm literature includes SimCLR Chen et al. (2020), BYOL Grill, Strub, Altché, Tallec, Richemond, Buchatskaya, Doersch, Avila Pires, Guo, Gheshlaghi Azar & others (2020), NNCLR Dwibedi, Aytar, Tompson, Sermanet & Zisserman (2021), VICReg Bardes, Ponce & LeCun (2021), and Barlow Twins Zbontar, Jing, Misra, LeCun & Deny (2021), MoCo Chen, Xie & He (2021) and TiCo Zhu, Moraes, Karakulak, Sobol, Canziani & LeCun (2022) which have been recently proposed for highly curated sets such as Imagenet. A significant number of them uses contrastive learning paradigm endeavors to minimize the distance between embeddings of similar (positive) samples generated from various random transformations or augmentations of the input image, while simultaneously increasing the distance between non-similar (negative) samples. Augmentation applied in contrastive learning can be cropping Bachman, Hjelm & Buchwalter (2019); He, Fan, Wu, Xie & Girshick (2020); Pathak et al. (2016); Srinivas, Laskin & Abbeel (2020); Wu et al. (2018); Ye, Zhang, Yuen & Chang (2019) among others. The variational contrastive representation learning has only recently been studied in semi-supervised fashion Yavuz & Yanikoglu (2022).

When it comes to dealing with noisy data, there is a limited amount of existing research available in academic literature. Some approaches use standard clustering algorithm that can deal with outlier noise in semantic space or learning from noisy label algorithms after self-supervised pre-training Tian et al. (2021); Zheltonozhskii, Baskin, Mendelson, Bronstein & Litany (2022), while others enforce neighbor consistency Iscen, Valmadre, Arnab & Schmid (2022). In addition to addressing noisy data, there is a upcoming body of research on continual learning that inherently tackles the issue of noise Karim, Khalid, Esmaeili & Rahnavard (2022). Furthermore, the literature also explores the use of multi-stage algorithms for managing noisy data effectively Smart & Carneiro (2023). The major weakness of such methodologies relies upon the fact that there might not be a proper label for noisy samples from web-collected sets. Moreover, these efforts mostly depend on initial self-supervised pre-training which we focus.

The aim of this research is to enhance the pre-training efficiency of classifiers by employing robust self-supervised techniques. This research aims to broaden the pool of training data by harnessing an almost infinite source: internet media. The challenge here is to achieve improved results despite the data noise.

We propose a robust self-supervised variational contrastive learning framework. The variational approach is first suggested for auto-encoders Kingma & Welling (2013) and recently used in Yavuz & Yanikoglu (2022). It is easy to motivate and interpret the variational method within a contrastive design: any two transforms of an image are sampled from the same distribution. Using this variational method, the beta divergence formulation is realized for Gaussian samples to overcome the data noise obstacle.

The proposed approach of VCL is validated across a range of diverse settings, including face attribute learning with the medium-sized CelebA and YFCC-CelebA datasets and face verification with the LFW dataset. Overall, the proposed approach exhibited promising results across a variety of settings and scenarios.

4.2 Methodology

4.2.1 The Contrastive Learning Framework

We propose a novel self-supervised contrastive learning schema that utilizes a Gaussian sampling head to learn the distribution of the images in the embedding space, as illustrated in Figure 4.1.

During training, two augmentations of the input image are first obtained (see Subsection 4.2.2). The backbone network then extracts fixed-length embedding vectors for the two augmentations, followed by the Gaussian sampling head (see Subsection 4.2.3). The network aims to reduce the distance between the two samples that are obtained from the same input with different augmentations. This task involves applying random rotation augmentations in four directions on the input image and using the orientation pseudo-labels as target distribution.

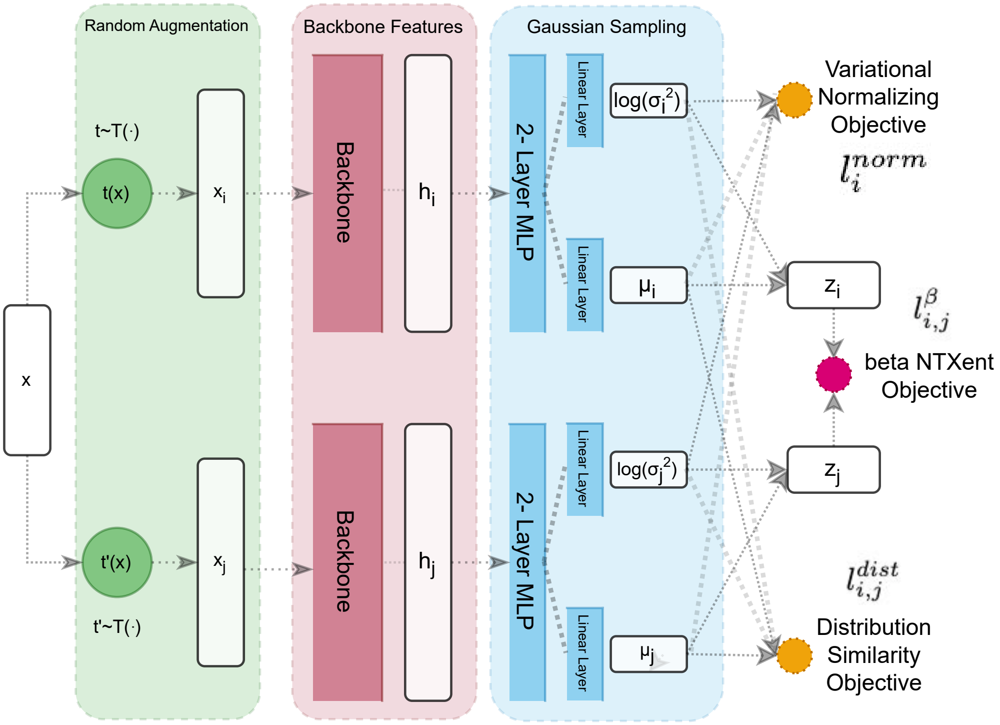


Figure 4.1 Diagram of our proposed model. While the variational objectives encourage two augmentations of the same image to be drawn from the same univarite distribution, the metric objective enhance the robustness to data noise. Light color boxes indicate the same operation or the shared weights.

4.2.2 Augmentations

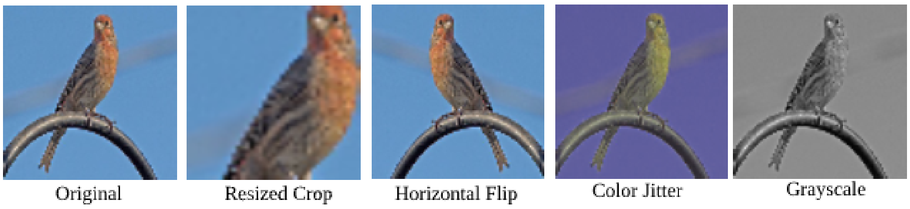


Figure 4.2 Four different augmentations for contrastive design.

Data augmentation is a widely used technique to prevent overfitting in supervised deep learning systems. As illustrated in Figure 4.1, a given an input x, the stochastic function $t(\cdot)$ is used to obtain two random augmentations of the input image, $\widetilde{x_i}$ and $\widetilde{x_j}$. These two correlated images constitute a positive pair, while the other examples in the mini-batch constitute the negative pairs. We follow the augmentation strategies suggested in the original article Chen et al. (2020). Specifically, we use Resize Crop, Horizontal Flip, Grayscale, Color Jitter augmentations (see Figure 4.2).

4.2.3 Feature Extraction and Gaussian Sampling

The proposed method for representation learning uses a backbone model $f(\cdot)$ that extracts the fixed-length embedding vectors for the two augmentations of the input image, indicated as $h_i = f(x_i)$, where $h_i \in R^d$. The Gaussian sampling head (g_z) which is a shallow multi-layer perceptron, then predicts the mean and log-variance of the distributions of the two representations, allowing for the sampling of new representations from the learned distribution. Specifically:

$$(\boldsymbol{\mu}, \log \boldsymbol{\sigma}^2) = g_z(f_\theta(t(\mathbf{x}))) \tag{4.1}$$

A sample $\mathbf{z}$ is then drawn from this distribution, after the so called reparameterization trick, as in Kingma & Welling (2013):

$$\boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{\sigma}^2 \odot \boldsymbol{\xi} \tag{4.2}$$

where $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\odot$ is the element-wise multiplication. The reparameterization allows backpropagation to be used, in spite of the sampling process. The samples obtained from the two augmentations contribute to the loss term given in Eqs. 4.3, 4.4.

Learning the distribution parameters has been previously shown to learn a smooth representation space that enable sampling, in variational auto-encoders Kingma & Welling (2013).

4.2.4 Objective Function

The training process uses a mini-batch of N randomly selected images from the dataset. From each image in the mini-batch, two random augmentations are obtained, resulting in a mini-batch of size $2N$ data points in total. For a positive pair (two augmentations of an image), the rest of the $2(N-1)$ samples within the mini-batch constitute the negative examples. Thus, within a mini-batch of size N, there are N positive and $2(N-1)$ negative pairs.

The training objective is to minimize the loss function shown in Eq. 4.6, containing three terms: i) beta-NT-Xent loss term derived from the beta divergence; ii) Distribution Similarity Loss; and iii) Distribution Normalizing Loss, as explained below.

beta-NT-Xent loss is used as the loss function that considers the similarity between an image embedding sample $z_i \sim p_\theta(Z_i|X_i)$ and another sample $z_j \sim p_\theta(Z_j|X_j)$, compared to the similarity between z_i and the negatives samples $z_{j \neq k}$. The loss for a positive embedding pairs z_i, z_j is defined as the ratio between the similarity of positive and negative samples in the mini-batch:

$$l^\beta_{i,j} = -\log \frac{\exp(\beta dist(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{k \neq i} \exp(\beta dist(z_i, z_k)/\tau)} \tag{4.3}$$

where τ is the temperature parameter, N is the batch size, and $\mathbb{1}_{k \neq i}$ is the indicator function evaluating to 1 if and only if $k \neq i$ and β_{dist} is a similarity metric between two sample embeddings z_i and z_j, such that:

$$\beta dist(z_j, z_i) = -\frac{\beta + 1}{\beta}\left(\frac{1}{(2\pi\sigma^2)^{\beta/2}} exp\Big(-\frac{\beta}{2\sigma^2} d(z_j, z_i)\Big) - 1\right)$$

where σ is the standard deviation of the hypothesized Gaussian distribution which is set to 0.5 as suggested by Akrami, Joshi, Li, Aydöre & Leahy (2022), β is a hyperparameter which is set to X, and d function is based on Euclidean distance between two vectors z_j, z_i in semantic space.

Distribution Similarity Loss. The second loss term encourages that two augmentations $(x_i,\ x_j)$ of the input should be drawn from similar distributions, q_1 and q_2. The loss term penalizes the Jensen–Shannon divergence between the two distributions Odaibo (2019):

Using Gaussian distributions, the Jensen-Shannon divergence reduces to:

$$l_{i,j}^{dist} = -\frac{1}{2}\Big[(\log(\sigma_i) - \log(\sigma_j))\Big] + \frac{1}{4}\left[\frac{(\mu_i - \mu_m)^2 + (\mu_j - \mu_m)^2}{\sigma_m^2}\right] \tag{4.4}$$

where $\log(\sigma_{i/j})$ and $\mu_{i/j}$ are the outputs of the variational encoder. The μ_m and σ_m are the means of the two means and two standard deviations, respectively.

Distribution Normalizing Loss. The third loss term encourages the learned distributions to be standard Gaussian Odaibo (2019), to eliminate degenerate solutions:

$$l_i^{norm} = D_{KL}(q_\theta(z_i|x_i)||\mathcal{N}(0,1)) = -\frac{1}{2}\left[1 + \log(\sigma_i^2) - \sigma_i^2 - \mu_i^2\right] \tag{4.5}$$

where $\log(\sigma)$ and μ are the outputs of the variational encoder and directly computed to minimize.

Overall Objective Function. The optimization is processed based on the total loss, which is comprised of the four loss terms defined in Equations 3-5:

$$\mathcal{L}_{total} = \frac{1}{N}\sum_{i=1}^{N}\left\{l_i^{norm} + \sum_{j=1}^{2N}\mathbf{1}_{i\neq j}(l_{i,j}^{\beta} + l_{i,j}^{dist})\right\} \tag{4.6}$$

where N is the number of images in the mini-batch and $\{2k-1, 2k\}$ notation follows Chen et al. (2020). The reader can refer to Algorithm 3 for a detailed explanation.

We considered three extensions of the Kullback-Leibler divergence that differ in sensitivity and robustness, namely alpha, beta, and gamma divergences. We chose beta divergence which is often preferred for practical algorithms such as robust PCA and clustering Mollah, Sultana, Minami & Eguchi (2010), robust ICA Mollah, Minami & Eguchi (2006), robust NMF Kompass (2007) and robust VAE Akrami et al. (2022), due to its balanced traits.

Algorithm 3 Pseudo-code for Variational Contrastive Learning with Beta Divergence.

```
Result: Optimized parameters of the base encoder $f(\cdot)$
Input: Batch size $N$, temperature parameter $\tau$, beta parameter $\beta$, augmentation
  function $T(\cdot)$, encoder $f(\cdot)$, Gaussian projection head $g(\cdot)$

for each minibatch of $N$ examples $\{x_i\}_{i=1}^{N}$ do
    // Generate augmented views of each example
    Compute $x_i^{'} = T(x_i)$ and $x_i^{''} = T(x_i)$

    // Compute representations of the augmented views
    Compute $h_i^{'} = f(x_i^{'})$ and $h_i^{''} = f(x_i^{''})$

    // Learn the distribution parameters and sample from the learned
       distribution
    Sample $z_i^{'} = g(h_i^{'})$ from $N(\mu_i^{'}, \sigma_i^{'})$
    Sample $z_i^{''} = g(h_i^{''})$ from $N(\mu_i^{''}, \sigma_i^{''})$

    // Compute the beta NT-Xent and variational objectives
    Compute $\mathcal{L}_i = l_i^{norm} + \sum_{j=1}^{2N} \mathbf{1}_{i \neq j}(l_{i,j}^{\beta} + l_{i,j}^{dist})$
    // Back-propagation and parameter update
    Update the parameters of $f(\cdot)$ and $g(\cdot)$ by minimizing $\mathcal{L}$
end
```

4.2.5 Implementation Details

We use the VGG16bn backbone which is often used as a benchmark in the face recognition community Ahmed & Yanikoglu (2021), with an embedding dimension of 4096. For training, we use the AdamW optimizer with 1e-3 learning rate and weight decay of 0.01 and Cosine Annihilation Scheduler. The batch size for all methods is 128. We train the networks with CelebA for 800 epochs (1M iterations) and YFCC-CelebA for 500 epochs (1.5M iterations).

For augmentations, we use the following ranges: resizing (scale between [0.2, 1.0]), cropping (128 random crops), grayscale transformation (with probability 0.2), and color jitter (with probability 0.8, brightness in [0.6,1.4], contrast in [0.6,1.4], saturation in [0.6,1.4] and hue in [0.9,1.1]). We optimized the hyper-parameter beta for different values using grid search and found out the optimal temperature as 0.07

and beta as 0.005 for CelebA and YFCC-CelebA datasets. [1]

4.3 Experimental Evaluation

We evaluate the proposed self-supervised algorithm in two protocols that are widely accepted in the self-supervised learning literature. In both protocols, the self-supervised backbone network is extended by adding a single, fully connected layer that is trained for the corresponding classification task using the learned representations.

(i) In the *linear evaluation protocol* used in the literature, we add a single layer as classification head and train only that layer with the labeled dataset, while the rest of the network is kept unchanged. The aim of this evaluation is to compare the quality of the learned representations to those obtained by fully supervised training.

(ii) In the *low-shot evaluation protocol*, we perform the self-supervised training without using labels and only use a small subset of the training data for fine-tuning all layers of this network. Here, the aim is to demonstrate the representation learning effectiveness of the proposed algorithm for low-data regimes, simulated by using only a small percentage of the data labels. In other words, the inquiry pertains to whether the enhancement of classifier efficiency is attainable through the utilization of web-sourced datasets.

(iii) In tasks of *relative face attributes* and *face verification*, we utilize a paired-image dataset, signifying more pronounced attribute expression and same-person classification respectively. For relative attribute learning, we follow the RankNet Souri, Noury & Adeli (2017) method where the embeddings of two images obtained by the same backbone are subtracted and the final ranking is learned with a logistic regression model. For face verification, we compare the Euclidean distance between embeddings of two images and accept or reject the verification by comparing this distance to a predetermined threshold.

In all settings, we compare our algorithms with state-of-the-art including SimCLR Chen et al. (2020), BYOL Grill et al. (2020), NNCLR Dwibedi et al. (2021), MoCo

[1] We made the code available at https://github.com/verimsu/VCL

Chen et al. (2021), VICReg Bardes et al. (2021), TiCo Zhu et al. (2022), and Barlow Twins Zbontar et al. (2021).

Information about the datasets used in the evaluations is summarized in Table 4.1. CelebA and LFW are well-known public datasets used for face attribute recognition and ranking, where each photograph is labelled in terms of 40 attributes. LFW is similar, but contains images collected from the internet. LFW-10 is a 10-attribute subset of LFW where image pairs are ordered with respect to the given attributes. Finally, YFCC-CelebA is a subset of the web-collected YFCC, where images matching attributes in CelebA or their opposites (if they exist) were selected Yavuz et al. (2021). This dataset is thus weakly labelled and noisy, as shown in Fig. 4.3.

Table 4.1 Summary of the datasets used in the study.

Datasets	Setting	Classes	Train	Validation	Test
Medium-Sized					
CelebA	Multi-labelled	40	162,770	19,867	19,962
YFCC-CelebA	Multi/Weakly labelled	59	392,220	-	-
Small-Scale					
LFW	Pairs	2	-	-	6,000
LFW-10	Pairs	3	500	-	500

4.3.1 Evaluation with Multi-Label CelebA Dataset

The effectiveness of the proposed method for a multi-label problem has been assessed using the CelebA dataset Liu, Luo, Wang & Tang (2015), which is widely recognized in the field of face attribute recognition. While facial attributes have been previously studied using supervised deep learning systems Aly & Yanikoglu (2018); Rozsa, Günther, Rudd & Boult (2016); Sharif Razavian, Azizpour, Sullivan & Carlsson (2014); Song, Tan & Chen (2014); Zhong, Sullivan & Li (2016); Zhu, Luo, Wang & Tang (2014), face attribute recognition has not been well studied in the context of self-supervised learning Sharma, Tapaswi, Sarfraz & Stiefelhagen (2019); Shu, Gu, Yang & Lo (2022); Wiles, Koepke & Zisserman (2018).

In this experiment, we applied self-supervised training on the training portion of the CelebA dataset *without* using the labels. Then the evaluation of the learned representations is carried out in accordance with protocol (i). In other words, we have added a linear layer that takes the learned representations as input and is trained with the whole labeled dataset, while the rest of the network is kept unchanged.

The results are presented in Table 4.2. We first evaluate transfer learning with a VGG16bn network that is trained with the ImageNet dataset and is used as feature extractor prior to the dense layer. While this approach is preferred especially with smaller labelled datasets, it also obtains the lowest accuracy with 86.31%. Self-supervised methods from the literature achive 86.51% to 87.98% accuracies, with best results obtained by SimCLR Kinakh, Taran & Voloshynovskiy (2021) and TiCo Zhu et al. (2022) methods.

The proposed self-supervised method outperforms transfer learning by almost 3% points (86.31 vs 89.23%) and other self-supervised methods by over 1% point (87.98 vs 89.23). For this dataset, VCL and VCL with beta-divergence achieve almost the same performance. On the other hand, using the beta divergence results in clearly better performance when self-supervised learning is done with the noisy YFCC-CelebA dataset, as described in Section 4.3.2

Table 4.2 Comparison of model accuracies using the entire labeled CelebA dataset for linear evaluation. Bold indicates the best results; underline indicates the second best.

Multi-label Logistic Reg. Evaluation	Supervised Pre-training w. Imagenet	Self-supervised Training w. CelebA	Mean Acc. on CelebA Test Set
Transfer Learning			
w. VGG16bn	yes	no	86.31%
Self-supervised methods			
Barlow Twins Zbontar et al. (2021)	no	yes	87.69%
BYOL Grill et al. (2020)	no	yes	86.78%
MoCo Chen et al. (2021)	no	yes	87.92%
NNCLR Dwibedi et al. (2021)	no	yes	86.51%
SimCLR Chen et al. (2020)	no	yes	87.98%
TiCo Zhu et al. (2022)	no	yes	87.98%
VICReg Bardes et al. (2021)	no	yes	87.44%
This work - VCL	no	yes	89.20%
This work - VCL (beta div.)	no	yes	**89.23**%

4.3.2 Evaluation with Web-Collected YFCC-CelebA Dataset

In a bid to further examine the robustness of our model in the face of data noise, we carried out a second experiment. For this, we implemented self-supervised training using the web-collected YFCC-CelebA dataset, described in Yavuz et al. (2021) and previously used in Yavuz & Yanikoglu (2022) within a semi-supervised learning paradigm. This dataset contains many non-face images and wrong labels, as shown in Figure 4.3. This experiment was evaluated following protocol (ii). In other words, networks that are pretrained using the YFCC-CelebA dataset are fine-tuned within a low-data regime using 10% or 1% of the labeled CelebA dataset.

Table 4.3 compares the results of various well-known alternatives from the literature and the proposed model. Transfer learning with a model trained only with supervised learning with the ImageNet dataset (no self-supervised pre-training) achieved the lowest accuracies, as was the case in the previous experiment. Self-supervised and semi-supervised methods from the literature achieved similar results for the 10% labelled data case, while semi-supervised methods achieved better results for the very low-data regime. The proposed VCL model outperformed all other methods, achieving 91.01% and 88.12% accuracies when trained with 10% or 1% labeled data respectively, which corresponds to more than 1% point increase over alternatives.

Even though the two evaluation goals are different, we see that pre-training with the web-collected data and fine-tuning with 10% of labelled CelebA results in higher accuracy (91.01% in Table 4.3), compared to pre-training with CelebA and using 100% of labelled CelebA (89.23% in Table 4.2). Finally, when comparing VCL and VCL with beta divergence, we see that the latter is indeed more effective for this noisy dataset.

Table 4.3 Comparison of model accuracies using the unlabeled YFCC-CelebA dataset for pre-training. The models are then fine-tuned using 10% and 1% of the labeled CelebA dataset. Bold indicates the best results; underline indicates the second best.

Multi-label Model Fine-tuning Evaluation	Supervised Pre-trained w. Imagenet	Self-sup. Training w. YFCC-CelebA	Mean Acc. on CelebA Test Set - CelebA (10%)	Mean Acc. on CelebA Test Set - CelebA (1%)
Transfer Learning	yes	no	89.05%	86.34%
Self-supervised methods				
Barlow Twins	no	yes	89.18%	86.64%
BYOL	no	yes	89.40%	86.98%
MoCo	no	yes	89.83%	86.66%
NNCLR	no	yes	89.95%	86.94%
SimCLR	no	yes	89.89%	86.87%
TiCo	no	yes	89.24%	86.85%
VICReg	no	yes	89.97%	86.99%
Semi-supervised				
CL-PL *(w. Alexnet)*	yes	yes	89.43%	87.69%
VCL-PL *(w. Alexnet)*	yes	yes	89.68%	88.12%
This work - VCL	no	yes	90.15%	87.49%
This work - VCL (beta div.)	no	yes	**91.01%**	**88.12%**

Figure 4.3 Random samples from the YFCC-CelebA dataset Yavuz et al. (2021).

4.3.3 Evaluation of Relative Face Attribute Learning

The evaluation of Relative Face Attributes is often carried out utilizing the LFW-10 dataset, a subset of the Labeled Faces in the Wild (LFW) dataset Huang, Mattar, Berg & Learned-Miller (2008). For evaluation, we compare the results of two supervised baselines and results from various algorithms are pre-trained with CelebA or YFCC-CelebA datasets, after which the whole network is fine-tuned with the training portion of the LFW-10 dataset. The relative attributes task is accomplished using a single-layer logistic regression model on the subtracted embeddings vectors.

Considering the results given in Table 4.4, the best results are obtained by the proposed algorithm with 88.15% accuracy when the self-supervised pre-training is done using the CelebA dataset, above the results obtained by well-known self-supervised methods.

When pretrained with the large but noisy YFCC-CelebA dataset, the performance of VCL drops slightly to 87.35%, second best after DRSVM Ahmed & Yanikoglu (2021). It should be noted that RankNet and DRSVM are supervised deep learning methods that are explicitly designed to tackle the relative attribute ranking problem. In contrast, our study pertains to a general-purpose self-supervised method.

When considering the results of VCL and VCL with beta divergence, we see that the beta-divergence is especially useful when the pre-training dataset is noisy.

Table 4.4 Relative attribute classification performances for models pre-trained with unlabeled CelebA and YFCC-CelebA datasets. Bold and underlined indicate the best and second best results, respectively.

Multi-label Relative Face Attributes	Supervised Pre-trained w. Imagenet	Self-sup. Training w. CelebA	Mean Acc. on LFW-10	Self-sup. Training w. YFCC	Mean Acc. on LFW-10
Supervised					
RankNet	yes	no	82.18%	no	82.18%
DRSVM	yes	no	88.12%	no	**88.12**%
Self-supervised					
Barlow Twins	no	yes	87.18%	yes	85.58%
BYOL	no	yes	86.23%	yes	85.02%
MoCo	no	yes	87.39%	yes	85.23%
NNCLR	no	yes	87.12%	yes	83.82%
SimCLR	no	yes	87.77%	yes	85.29%
TiCo	no	yes	87.27%	yes	85.15%
VICReg	no	yes	87.93%	yes	85.47%
This work - VCL	no	yes	87.96%	yes	86.70%
This work - VCL (beta div.)	no	yes	**88.15**%	yes	87.35%

4.3.4 Evaluation of Face Verification

For this evaluation, we employed Labeled Faces in the Wild (LFW) dataset that contains 6,000 pairs of face images Huang et al. (2008). Although the issue of the face verification algorithm has traditionally been addressed within the supervised learning domain Ding & Tao (2015); Parkhi, Vedaldi & Zisserman (2015); Schroff, Kalenichenko & Philbin (2015), the emphasis on evaluating the resultant embedding space has increasingly become salient in the context of self-supervised face verification tasks Liu et al. (2016).

This study, per protocol (iii), aimed to evaluate the quality of the self-supervised models' embedding space through the measurement of pre-training algorithms for the face verification task. The verification task is accomplished by determining a threshold for the Euclidean distance between the embeddings of the image pairs, over the LFW training set; the same threshold is used for the test set.

Considering the results in Table 4.5, the first observation is that the performance of these algorithms fluctuates significantly when trained with CelebA versus YFCC-CelebA datasets. On the other hand, our proposed algorithm achieves very similar results with both pre-training datasets (74.30 and 74.34%), as well as surpassing the results of all other methods with a noteable margin.

Table 4.5 Face verification comparison of different self-supervised models pre-trained with unlabeled CelebA and YFCC-CelebA datasets. Bold indicates the best results; underline indicates the second best.

Face Verification	Mean Acc. on LFW - Self-sup. Training w. CelebA	Mean Acc. on LFW - Self-sup. Training w. YFCC-CelebA
Self-supervised methods		
Barlow Twins Zbontar et al. (2021)	72.80%	68.30%
BYOL Grill et al. (2020)	66.80%	52.70%
MoCo Chen et al. (2021)	68.20%	69.50%
NNCLR Dwibedi et al. (2021)	67.00%	65.30%
SimCLR Chen et al. (2020)	66.90%	67.20%
TiCo Zhu et al. (2022)	70.20%	66.70%
VICReg Bardes et al. (2021)	72.90%	67.70%
This work - VCL	73.50%	73.70%
This work - VCL (beta div.)	**74.30**%	**74.34**%

4.4 Conclusion and Future Works

We proposed an algorithm in the family of contrastive learning framework, using beta-NT-Xent loss term derived from the beta divergence for robustness against outliers in the noisy set. The approach differs from simple contrastive design in its variational approach and use of beta divergence in the self-supervised objective.

Our findings demonstrate that variational methods employing beta-divergence offer a robust alternative for tackling noisy datasets, multi-label settings and low-data regime. The results outperform existing self-supervised models especially for the case where the self-supervised learning is accomplished with a noisy dataset.

We also compared the effect of different objective components on accuracy, as well as temperature and beta value (not included for clarity). The combination of all components yielded the highest accuracies.

Future work will focus on evaluating and refining the proposed VCL model on larger-scale datasets like Imagenet and WebVision, and with different architectures such as ResNet-50 and Vision Transformers.

4.5 Appendix

4.5.1 Different Divergences

The Kullback-Leibler (KL) divergence, a measure of the difference between two probability distributions, can be generalized by using a family of functions known as generalized logarithm functions or α-logarithm,

$$\log_\alpha(x) = \frac{1}{1-\alpha}(x^{1-\alpha} - 1) \tag{4.7}$$

(for $x > 0$) which is a power function of x with power $1-\alpha$. The natural logarithm function is included in this family as a special case, where $\alpha \to 1$.Cichocki & Amari (2010)

By utilizing the concept of generalized logarithm functions, a family of divergences can be derived, which are known as the Alpha, Beta, and Gamma divergences. These divergences are extensions of the Kullback-Leibler (KL) divergence and can be used to measure the dissimilarity between two probability distributions in a more flexible way. Each of these divergences is defined using a different function and parameterization, allowing for different trade-offs between sensitivity and robustness. Especially Beta- and Gama- divergences are robust in respect to outliers for some values of tuning parameters, but Gama- divergence is a "super" robust estimation of some parameters in presence of outlier. Thus, beta divergence is more common choice for practical algorithms in literature, for example, robust PCA and clustering Mollah et al. (2010), robust ICA Mollah et al. (2006), and robust NMF Kompass (2007) and robust VAE Akrami et al. (2022).

4.5.2 Beta-divergence Formulation

We consider a parametric model $p_\theta(X)$ with parameter θ and minimize the KL divergence between the two distributions (i.e. empirical distribution and probability distribution)Akrami et al. (2022):

$$D_{KL}(p(X)||p_\theta(X)) = \int p(X) \log \frac{p(X)}{p_\theta(X)} dx \tag{4.8}$$

where $p(X) = \frac{1}{N}\sum_{i=0}^{N} \delta(X, x^i)$ is the empirical distribution and its approximation to be converged. This is equivalent to minimizing *maximum likelihood estimation*:

$$arg \min_\theta \frac{1}{N} \sum_{i=1}^{N} \ln p_\theta(X)$$

Unfortunately, this formulation is sensitive to outliers because all data points contributes to the error with equal ratio. The density power divergence, or beta- divergence, is a robust alternative to above formulation, proposed in Basu, Harris, Hjort & Jones (1998):

$$D_\beta(g||f) = -\frac{\beta+1}{\beta} \int g(x)(f(x)^\beta - 1)dx + \int f(x)^{1+\beta} dx$$

To motivate the use of the beta-divergence in 4.8, please note that minimizing the beta-divergence with empirical distribution yields:

$$0 = \frac{1}{N} \sum_{i=1}^{N} p_\theta(X)^\beta \frac{\partial}{\partial\theta} \ln p_\theta(X) - \mathbf{E}_{p_\theta(X)} p_\theta(X)^\beta \frac{\partial}{\partial\theta} \ln p_\theta(X)$$

where the second term assures the unbiasedness of the estimator and the first term is likelihood weighted according to the power of the probability density for each data point. Thus, the weights of the outliers will be much less then the major inliers, and the system will be more robust to noise.

Instead of using the empirical distribution, we can reduce the difference between two distributions from the same network:

$$\begin{aligned} &D_\beta(p_\theta(Z_j|X_j)||p_\theta(Z_i|X_i)) = \\ &-\frac{\beta+1}{\beta} \int p_\theta(Z_j|X_j)(p_\theta(Z_i|X_i)^\beta - 1)dX + \int p_\theta(Z_i|X_i)^{\beta+1} dX \end{aligned}$$

where $p_{theta}(Z|X)$ are posterior distributions. This formulation is essential to the formulation of a robust self-supervised objective function in Eqn. 4.4.

4.5.3 Derivation of Variational Objective Functions

The KL divergence between the two Gaussian distributions can be reduced to(i.e. empirical distribution and probability distribution)Yavuz & Yanikoglu (2022):

$$-D_{KL}(q_\theta(z|x_i)||p(z)) = \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{\sigma_q^2 + (\mu_q - \mu_p)^2}{2\sigma_p^2} + \frac{1}{2} \tag{4.9}$$

From this formulation we will derive two variational objective terms in the paper.

- Distribution Normalizing Loss 4.5.3.1
- Distribution Similarity Loss by using Jensen-Shannon Divergence 4.5.3.2

4.5.3.1 Normalizing Objective

We take the $\sigma_p = 1$ and $\mu_p = 0$ for Eq. 4.9:

$$\begin{aligned} D_{KL}(q_\theta(z|x_i)||p(z)) &= -\log(\sigma_q) + \frac{\sigma_q^2 + \mu_q^2}{2} - \frac{1}{2} \\ &= -\frac{1}{2}\log(\sigma_q^2) + \frac{\sigma_q^2 + \mu_q^2}{2} - \frac{1}{2} \\ &= -\frac{1}{2}\Big[1 + \log(\sigma_q^2) - \sigma_q^2 - \mu_q^2\Big] \end{aligned} \tag{4.10}$$

4.5.3.2 Distribution Similarity Objective using Jensen-Shannon Divergence

$$\begin{aligned} \mathcal{L}_{a,k} = JSD(q_1||q_2) &= \frac{1}{2}\Big(D_{KL}(q_1||m) + D_{KL}(q_2||m)\Big) \\ \text{where} \quad m &= \frac{1}{2}(q_1 + q_2) \end{aligned}$$

where the the sum of KL-divergences can be reduced to variational network output:

$$\begin{aligned}
D_{KL}(q_1||m)+D_{KL}(q_2||m) &= -\log\left(\frac{\sigma_{q_1}}{\sigma_m}\right)+\frac{\sigma_{q_1}^2+(\mu_{q_1}-\mu_m)^2}{2\sigma_m^2}-\frac{1}{2}\\
&\quad -\log\left(\frac{\sigma_{q_2}}{\sigma_m}\right)+\frac{\sigma_{q_2}^2+(\mu_{q_2}-\mu_m)^2}{2\sigma_m^2}-\frac{1}{2}\\
&= -(\log(\sigma_{q_1})-\log(\sigma_m))-(\log(\sigma_{q_2})-\log(\sigma_m))\\
&\quad +\frac{\sigma_{q_1}^2+\sigma_{q_2}^2}{2\sigma_m^2}+\frac{(\mu_{q_1}-\mu_m)^2}{2\sigma_m^2}+\frac{(\mu_{q_2}-\mu_m)^2}{2\sigma_m^2}-1\\
&= -(\log(\sigma_{q_1})-\log(\sigma_m))-(\log(\sigma_{q_2})-\log(\sigma_m))\\
&\quad +\frac{(\mu_{q_1}-\mu_m)^2+(\mu_{q_2}-\mu_m)^2}{2\sigma_m^2}
\end{aligned}$$

4.5.4 Ablation Studies for Objective Functions and Hyper-parameters

Table 4.6 Ablation Studies: First table showing the accuracy obtained with and without the objective components. Second table is the hyper-parameter response of the system.

Objectives	beta-NT-Xent	Dist. Sim.	Dist. Norm.	CelebA	YFCC-CelebA
Accuracy	yes	no	no	86.19%	86.11
	yes	yes	no	88.42%	87.92
	yes	no	yes	**88.54%**	**88.01**

	CelebA	YFCC-CelebA		CelebA	YFCC-CelebA
temp. (T)	Acc.	Acc. %1	beta (T=0.07)	Acc.	Acc. %1
0.07	**89.20**	**87.49**	0.001	88.96	88.47
0.1	88.87	87.26	0.005	**89.23**	**88.01**
0.2	88.69	86.79	0.010	89.17	87.77

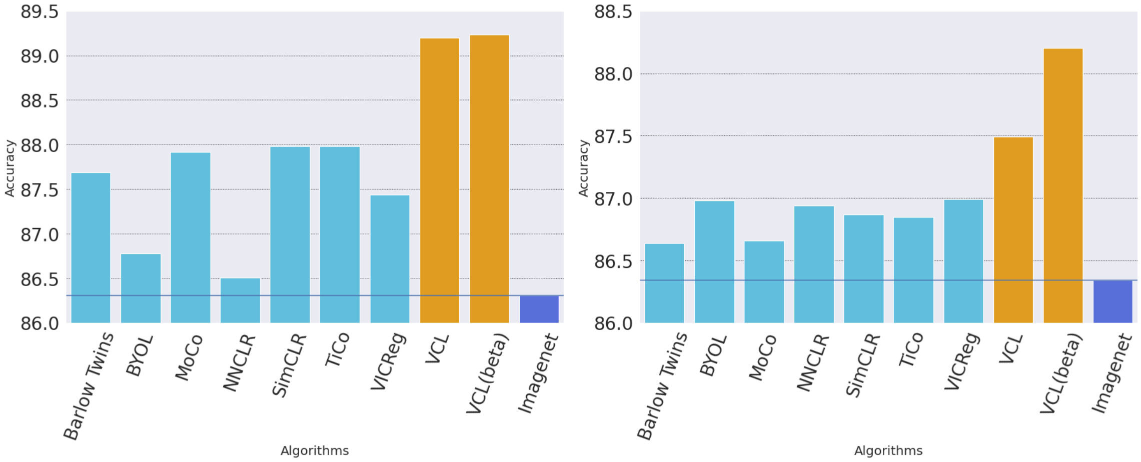


Figure 4.4 (a) CelebA Linear Evaluation (b) YFCC-CelebA Low Shot Ranking. Comparative performance of self-supervised pre-training models in multilabel tasks using the CelebA and YFCC-CelebA datasets. The bar charts differentiate algorithms using color: orange represents VCL algorithms, turquoise represents state-of-the-art models, and blue bars denote supervised baselines.

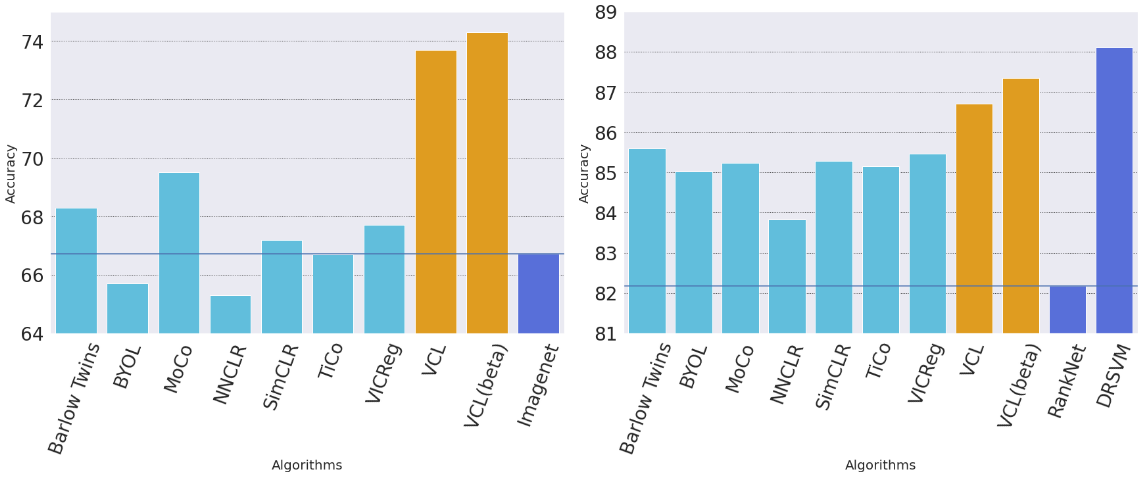


Figure 4.5 (a) Face Verification (b) Relative Attributes Classification. Performance results on the LFW dataset tasks, represented as bar plots: face verification and relative attributes classification. Colored bars distinguish the algorithms - orange for VCL algorithms, turquoise for state-of-the-art models, and blue for supervised baselines.

CHAPTER 5

CONCLUSION

In conclusion, this dissertation presents methodologies in deep learning that effectively utilize weakly-labeled biomedical data and uncurated/unlabeled biometric data. The research introduced two classifiers that employ 2D and 3D techniques under weak supervision, demonstrating their effectiveness with volume-wise labeled CT lung images.

The main contribution of the thesis is a novel representation learning method, which extends the contrastive learning framework through the integration of the variational approach.

Furthermore, a novel semi-supervised pseudo-labeling technique, VCL-PL, was developed to mitigate the inherent noise in web-collected face attribute classifications. This technique demonstrated an improvement in accuracy across various experimental setups, validating its effectiveness in handling noisy data.

The dissertation also introduced a robust self-supervised learning model, VCL, that integrates variational contrastive learning with beta-divergence. This model outperformed state-of-the-art models when applied to unlabeled, uncurated, and noisy datasets, showcasing its robustness and adaptability.

These methodological advancements and the introduction of new datasets contribute to the field of deep learning, particularly in the processing of weakly-labeled biomedical data and the management of noisy biometric data. The research findings have the potential to drive further advancements in the field, paving the way for more accurate and efficient data processing techniques.

5.1 Future Works

This methodology can be evaluated on large-scale datasets such as Imagenet and WebVision using architectures like Resnet-50 and Vision Transformers. These datasets and architectures will allow us to compare our model with state-of-the-art self-supervised methodologies.

Additionally, we aim to explore additional regularization techniques or memory units to stabilize the embedding space for non-contrastive frameworks. This could potentially enhance the performance of our models, particularly in scenarios where contrastive methods may not be the most suitable approach.

By pursuing these avenues, we hope to continue advancing the field of deep learning and further improve the processing of weakly-labeled biomedical data and the handling of noisy biometric data.

BIBLIOGRAPHY

Afshar, P., Heidarian, S., Enshaei, N., Naderkhani, F., Rafiee, M. J., Oikonomou, A., Fard, F. B., Samimi, K., Plataniotis, K. N., & Mohammadi, A. (2020). COVID-CT-MD: COVID-19 computed tomography (CT) scan dataset applicable in machine learning and deep learning. *arXiv 2009.14623*.

Ahmed, S. A. A. & Yanikoglu, B. (2019). Within-network ensemble for face attributes classification. In *International Conference on Image Analysis and Processing*, (pp. 466–476). Springer.

Ahmed, S. A. A. & Yanikoglu, B. (2021). Relative attribute classification with deep-ranksvm. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part II*, (pp. 659–671). Springer.

Ahmed, S. A. A., Yanikoglu, B., Zor, C., Awais, M., & Kittler, J. (2020). Skin lesion diagnosis with imbalanced ECOC ensembles. In *Int. Conf. on Machine Learning, Optimization, and Data Science*. Springer.

Ahmed, S. A. A., Yavuz, M. C., Şen, M. U., Gülşen, F., Tutar, O., Korkmazer, B., Samancı, C., Şirolu, S., Hamid, R., Eryürekli, A. E., et al. (2022). Comparison and ensemble of 2d and 3d approaches for covid-19 detection in ct images. *Neurocomputing*, *488*, 457–469.

Akrami, H., Joshi, A. A., Li, J., Aydöre, S., & Leahy, R. M. (2022). A robust variational autoencoder using beta divergence. *Knowledge-Based Systems*, *238*, 107886.

Aly, S. A. & Yanikoglu, B. (2018). Multi-label networks for face attributes classification. In *2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, (pp. 1–6). IEEE.

Bachman, P., Hjelm, R. D., & Buchwalter, W. (2019). Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, *32*.

Bardes, A., Ponce, J., & LeCun, Y. (2021). Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*.

Basu, A., Harris, I. R., Hjort, N. L., & Jones, M. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, *85*(3), 549–559.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., & Raffel, C. A. (2019). Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, *32*, 5049–5059.

Bo, D., Wang, X., Shi, C., Zhu, M., Lu, E., & Cui, P. (2020). Structural deep clustering network. In *Proc. of The Web Conf. 2020*, (pp. 1400–1410).

Bojanowski, P. & Joulin, A. (2017). Unsupervised learning by predicting noise. In *International Conference on Machine Learning*, (pp. 517–526). PMLR.

Caron, M., Bojanowski, P., Joulin, A., & Douze, M. (2018). Deep clustering for unsupervised learning of visual features. In Ferrari, V., Hebert, M., Sminchisescu, C., & Weiss, Y. (Eds.), *Computer Vision – ECCV 2018*, (pp. 139–156)., Cham. Springer International Publishing.

Caron, M., Bojanowski, P., Mairal, J., & Joulin, A. (2019). Unsupervised pre-training of image features on non-curated data. In *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*, (pp. 2959–2968).

Chaddad, A., Hassan, L., & Desrosiers, C. (2021). Deep CNN models for predicting COVID-19 in CT and X-ray images. *Journal of Medical Imaging*, *8*(S1), 014502.

Chaudhary, P. K. & Pachori, R. B. (2021). FBSED based automatic diagnosis of COVID-19 using X-ray and CT images. *Computers in Biology and Medicine*, *134*, 104454.

Chen, P., Liu, S., & Jia, J. (2021). Jigsaw clustering for unsupervised visual representation learning. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, (pp. 11526–11535).

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, (pp. 1597–1607). PMLR.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., & Hinton, G. E. (2020). Big self-supervised models are strong semi-supervised learners. *Advances in Neural Information Processing Systems*, *33*, 22243–22255.

Chen, T., Zhai, X., Ritter, M., Lucic, M., & Houlsby, N. (2019). Self-supervised gans via auxiliary rotation loss. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, (pp. 12154–12163).

Chen, X., Xie, S., & He, K. (2021). An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (pp. 9640–9649).

Chong, E. K. & Zak, S. H. (2004). *An introduction to optimization.* John Wiley & Sons.

Cichocki, A. & Amari, S.-i. (2010). Families of alpha- beta- and gamma- divergences: Flexible and robust measures of similarities. *Entropy*, *12*(6), 1532–1568.

Cole, E., Yang, X., Wilber, K., Mac Aodha, O., & Belongie, S. (2022). When does contrastive visual representation learning work? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 14755–14764).

Corman, V. M., Landt, O., Kaiser, M., Molenkamp, R., Meijer, A., Chu, D. K., Bleicker, T., Brünink, S., Schneider, J., Schmidt, M. L., et al. (2020). Detection of 2019 novel coronavirus (2019-ncov) by real-time RT-PCR. *Eurosurveillance*, *25*(3), 2000045.

Dai, Z., Yang, Z., Yang, F., Cohen, W. W., & Salakhutdinov, R. R. (2017). Good semi-supervised learning that requires a bad gan. *Advances in Neural Information Processing Systems*, *30*.

Dang, H., Liu, F., Stehouwer, J., Liu, X., & Jain, A. K. (2020). On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 5781–5790).

Dang, Z., Deng, C., Yang, X., Wei, K., & Huang, H. (2021). Nearest neighbor matching for deep clustering. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, (pp. 13693–13702).

de la Iglesia Vayá, M., Saborit, J. M., Montell, J. A., Pertusa, A., Bustos, A., Cazorla, M., Galant, J., Barber, X., Orozco-Beltrán, D., García-García, F., Caparrós, M., González, G., & Salinas, J. M. (2020). BIMCV COVID-19+: a large annotated dataset of RX and CT images from COVID-19 patients.

arXiv 2006.01174.

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

Ding, C. & Tao, D. (2015). Robust face recognition via multimodal deep face representation. *IEEE transactions on Multimedia*, *17*(11), 2049–2058.

Doersch, C., Gupta, A., & Efros, A. A. (2015). Unsupervised visual representation learning by context prediction. In *Proc. of the IEEE Int. Conf. on Computer Vision*, (pp. 1422–1430).

Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., & Zisserman, A. (2021). With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (pp. 9588–9597).

Feng, H.-Z., Kong, K., Chen, M., Zhang, T., Zhu, M., & Chen, W. (2021). Shot-vae: Semi-supervised deep generative models with label-aware elbo approximations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, (pp. 7413–7421).

Feng, Z., Xu, C., & Tao, D. (2019). Self-supervised representation learning by rotation feature decoupling. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, (pp. 10364–10374).

Figueroa, J. A. & Rivera, A. R. (2017). Is simple better?: Revisiting simple generative models for unsupervised clustering. In *NIPS Workshop on Bayesian Deep Learning*.

Goyal, P., Caron, M., Lefaudeux, B., Xu, M., Wang, P., Pai, V., Singh, M., Liptchinsky, V., Misra, I., Joulin, A., et al. (2021). Self-supervised pretraining of visual features in the wild. *arXiv preprint arXiv:2103.01988*.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. (2020). Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, *33*, 21271–21284.

Guo, X., Zhu, E., Liu, X., & Yin, J. (2018). Deep embedded clustering with data augmentation. In *Asian Conf. on Machine Learning*, (pp. 550–565). PMLR.

Gupta, N., Kaul, A., Sharma, D., et al. (2020). Deep learning assisted COVID-19 detection using full CT-scans. *TechRxiv 10.36227/techrxiv.13162049.v1*.

Hammoudi, K., Benhabiles, H., Melkemi, M., Dornaika, F., Arganda-Carreras, I., Collard, D., & Scherpereel, A. (2020). Deep learning on chest X-ray images to detect and evaluate pneumonia cases at the era of COVID-19. *arXiv preprint arXiv:2004.03399*.

Harmon, S. A., Sanford, T. H., Xu, S., Turkbey, E. B., Roth, H., Xu, Z., Yang, D., Myronenko, A., Anderson, V., Amalou, A., et al. (2020). Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets. *Nature Communications*, *11*(1), 1–7.

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, (pp. 9729–9738).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition*, (pp. 770–778).

He, X., Wang, S., Chu, X., Shi, S., Tang, J., Liu, X., Yan, C., Zhang, J., & Ding, G. (2021). Automated model design and benchmarking of 3D deep learning models for COVID-19 detection with chest CT scans. *arXiv preprint arXiv:2101.05442*.

Heidarian, S., Afshar, P., Enshaei, N., Naderkhani, F., Oikonomou, A., Atashzar, S. F., Fard, F. B., Samimi, K., Plataniotis, K. N., Mohammadi, A., & Rafiee, M. J. (2020). COVID-FACT: A fully-automated capsule network-based framework for identification of COVID-19 cases from chest CT scans. *arXiv 2010.16041*.

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

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition*, (pp. 4700–4708).

Huang, G. B., Mattar, M., Berg, T., & Learned-Miller, E. (2008). Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*.

Iscen, A., Tolias, G., Avrithis, Y., & Chum, O. (2019). Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 5070–5079).

Iscen, A., Valmadre, J., Arnab, A., & Schmid, C. (2022). Learning with neighbor consistency for noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 4672–4681).

Islam, M. M., Karray, F., Alhajj, R., & Zeng, J. (2020). A review on deep learning techniques for the diagnosis of novel coronavirus (COVID-19).

Jang, E., Gu, S., & Poole, B. (2016). Categorical reparameterization with Gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Jenni, S. & Favaro, P. (2018). Self-supervised feature learning by learning to spot artifacts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 2733–2742).

Jin, C., Chen, W., Cao, Y., Xu, Z., Tan, Z., Zhang, X., Deng, L., Zheng, C., Zhou, J., Shi, H., et al. (2020). Development and evaluation of an artificial intelligence system for COVID-19 diagnosis. *Nature Communications*, *11*(1), 1–14.

Johannes, H., Jeanny, P., Sebastian, R., Helmut, P., & Georg, L. (2020). Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. *European Radiology Experimental*, *4*(1).

Karim, N., Khalid, U., Esmaeili, A., & Rahnavard, N. (2022). Cnll: A semi-supervised approach for continual noisy label learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 3878–3888).

Ke, Z., Wang, D., Yan, Q., Ren, J. S. J., & Lau, R. W. H. (2019). Dual student: Breaking the limits of the teacher in semi-supervised learning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 6727–6735.

Kim, D., Cho, D., Yoo, D., & Kweon, I. S. (2018). Learning image representations by completing damaged jigsaw puzzles. In *2018 IEEE Winter Conf. on Ap-*

plications of Computer Vision (WACV), (pp. 793–802). IEEE.

Kinakh, V., Taran, O., & Voloshynovskiy, S. (2021). Scatsimclr: self-supervised contrastive learning with pretext task regularization for small-scale datasets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (pp. 1098–1106).

Kingma, D. P. & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kolesnikov, A., Zhai, X., & Beyer, L. (2019). Revisiting self-supervised visual representation learning. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, (pp. 1920–1929).

Kompass, R. (2007). A generalized divergence measure for nonnegative matrix factorization. *Neural computation*, *19*(3), 780–791.

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. *Master's thesis, University of Tront.*

Krizhevsky, A. (2014). One weird trick for parallelizing convolutional neural networks.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*, 1097–1105.

Kumar, R., Khan, A. A., Zhang, S., Wang, W., Abuidris, Y., Amin, W., & Kumar, J. (2020). Blockchain-federated-learning and deep learning models for COVID-19 detection using CT imaging. *arXiv preprint arXiv:2007.06537*.

Larsson, G., Maire, M., & Shakhnarovich, G. (2016). Learning representations for automatic colorization. In *European Conf. on Computer Vision*, (pp. 577–593). Springer.

Larsson, G., Maire, M., & Shakhnarovich, G. (2017). Colorization as a proxy task for visual understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 6874–6883).

Lee, H., Hwang, S. J., & Shin, J. (2020). Self-supervised label augmentation via input transformations. In *Int. Conf. on Machine Learning*, (pp. 5714–5724). PMLR.

Lee, H.-Y., Huang, J.-B., Singh, M., & Yang, M.-H. (2017). Unsupervised representation learning by sorting sequences. In *Proc. of the IEEE Int. Conf. on Computer Vision*, (pp. 667–676).

Lee, W., Na, J., & Kim, G. (2019). Multi-task self-supervised object detection via recycling of bounding box annotations. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition*, (pp. 4984–4993).

Li, C., Xu, K., Liu, J., Zhu, J., & Zhang, B. (2021). Triple generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, *PP*.

Li, L., Qin, L., Xu, Z., Yin, Y., Wang, X., Kong, B., Bai, J., Lu, Y., Fang, Z., Song, Q., et al. (2020). Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT. *Radiology*.

Li, W., Wang, L., Li, W., Agustsson, E., & Van Gool, L. (2017). Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*.

Liu, B., Gao, X., He, M., Liu, L., & Yin, G. (2020). A fast online COVID-19 diagnostic system with chest CT scans. In *Proceedings of KDD*.

Liu, S., Xu, D., Zhou, S. K., Pauly, O., Grbic, S., Mertelmeier, T., Wicklein, J., Jerebko, A., Cai, W., & Comaniciu, D. (2018). 3D anisotropic hybrid network: Transferring convolutional features from 2D images to 3D anisotropic volumes. In *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*, (pp. 851–858). Springer.

Liu, W., Wen, Y., Yu, Z., & Yang, M. (2016). Large-margin softmax loss for convolutional neural networks. In *ICML*, volume 2, (pp.7̃).

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, (pp. 3730–3738).

Long, Q.-X., Tang, X.-J., Shi, Q.-L., Li, Q., Deng, H.-J., Yuan, J., Hu, J.-L., Xu, W., Zhang, Y., Lv, F.-J., et al. (2020). Clinical and immunological assessment of asymptomatic SARS-CoV-2 infections. *Nature Medicine*, *26*(8), 1200–1204.

Maguolo, G. & Nanni, L. (2021). A critic evaluation of methods for COVID-19 automatic detection from X-ray images. *Information Fusion*, *76*, 1–7.

Minderer, M., Bachem, O., Houlsby, N., & Tschannen, M. (2020). Automatic shortcut removal for self-supervised representation learning. In *Int. Conf. on Machine Learning*, (pp. 6927–6937). PMLR.

Misra, I. & Maaten, L. v. d. (2020). Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 6707–6717).

Miyato, T., Maeda, S.-i., Koyama, M., & Ishii, S. (2018). Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, *41*(8), 1979–1993.

Mollah, M. N. H., Minami, M., & Eguchi, S. (2006). Exploring latent structure of mixture ica models by the minimum β-divergence method. *Neural Computation*, *18*(1), 166–190.

Mollah, M. N. H., Sultana, N., Minami, M., & Eguchi, S. (2010). Robust extraction of local structures by the minimum β-divergence method. *Neural Networks*, *23*(2), 226–238.

Morozov, S., Andreychenko, A., Pavlov, N., Vladzymyrskyy, A., Ledikhova, N., Gombolevskiy, V., Blokhin, I. A., Gelezhe, P., Gonchar, A., & Chernina, V. Y. (2020). Mosmeddata: Chest CT scans with COVID-19 related findings dataset. *arXiv preprint arXiv:2005.06465*.

Mukherjee, S., Asnani, H., Lin, E., & Kannan, S. (2019). Clustergan: Latent space clustering in generative adversarial networks. In *Proc. of the AAAI Conf. on Artificial Intelligence*, volume 33, (pp. 4610–4617).

Mundhenk, T. N., Ho, D., & Chen, B. Y. (2018). Improvements to context based self-supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 9339–9348).

Narin, A., Kaya, C., & Pamuk, Z. (2020). Automatic detection of coronavirus disease (covid-19) using X-ray images and deep convolutional neural networks. *arXiv preprint arXiv:2003.10849*.

Nassar, I., Herath, S., Abbasnejad, E., Buntine, W. L., & Haffari, G. (2021). All labels are not created equal: Enhancing semi-supervision via label grouping and co-training. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7237–7246.

Niu, C., Shan, H., & Wang, G. (2021). Spice: Semantic pseudo-labeling for image clustering. *arXiv preprint arXiv:2103.09382*.

Noroozi, M., Pirsiavash, H., & Favaro, P. (2017). Representation learning by learning to count. In *Proceedings of the IEEE International Conference on Computer Vision*, (pp. 5898–5906).

Noroozi, M., Vinjimoor, A., Favaro, P., & Pirsiavash, H. (2018). Boosting self-supervised learning via knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 9359–9367).

Odaibo, S. G. (2019). Tutorial: Deriving the standard variational autoencoder (vae) loss function. *ArXiv*, *abs/1907.08956*.

Ozsahin, I., Sekeroglu, B., Musa, M. S., Mustapha, M. T., & Ozsahin, D. U. (2020). Review on diagnosis of COVID-19 from chest CT images using artificial intelligence. *Computational and Mathematical Methods in Medicine*, *9756518*.

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., & Efros, A. A. (2016). Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 2536–2544).

Pham, H., Xie, Q., Dai, Z., & Le, Q. V. (2021). Meta pseudo labels. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11552–11563.

Rasmus, A., Berglund, M., Honkala, M., Valpola, H., & Raiko, T. (2015). Semi-supervised learning with ladder networks. *Advances in Neural Information Processing Systems*, *28*, 3546–3554.

Reed, C. J., Metzger, S., Srinivas, A., Darrell, T., & Keutzer, K. (2021). Selfaugment: Automatic augmentation policies for self-supervised learning. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, (pp. 2674–2683).

Ren, Z. & Lee, Y. J. (2018). Cross-domain self-supervised multi-task feature learning using synthetic imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 762–771).

Rizve, M. N., Duarte, K., Rawat, Y. S., & Shah, M. (2020). In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *International Conference on Learning Representations*.

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A. I., Etmann, C., McCague, C., Beer, L., et al. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence*, *3*(3), 199–217.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Int. Conf. on Medical Image Computing and Computer-assisted Intervention*, (pp. 234–241). Springer.

Rozsa, A., Günther, M., Rudd, E. M., & Boult, T. E. (2016). Are facial attributes adversarially robust? In *2016 23rd International Conference on Pattern Recognition (ICPR)*, (pp. 3121–3127). IEEE.

Rubin, G. D., Ryerson, C. J., Haramati, L. B., Sverzellati, N., Kanne, J. P., Raoof, S., Schluger, N. W., Volpi, A., Yim, J.-J., Martin, I. B., et al. (2020). The role of chest imaging in patient management during the COVID-19 pandemic: A multinational consensus statement from the fleischner society. *Chest*, *158*(1), 106–116.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *Int. Journal of Computer Vision (IJCV)*, *115*(3), 211–252.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training gans. *Advances in neural information processing systems*, *29*, 2234–2242.

Santa Cruz, R., Fernando, B., Cherian, A., & Gould, S. (2017). Deeppermnet: Visual permutation learning. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, (pp. 3949–3957).

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 815–823).

Sellars, P., Avilés-Rivero, A. I., & Schönlieb, C.-B. (2021). Laplacenet: A hybrid energy-neural model for deep semi-supervised classification. *ArXiv*, *abs/2106.04527*.

Sharif Razavian, A., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, (pp. 806–813).

Sharma, V., Tapaswi, M., Sarfraz, M. S., & Stiefelhagen, R. (2019). Self-supervised learning of face representations for video face clustering. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, (pp. 1–8).

Shi, F., Wang, J., Shi, J., Wu, Z., Wang, Q., Tang, Z., He, K., Shi, Y., & Shen, D. (2020). Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for COVID-19. *IEEE Reviews in Biomedical Engineering*, 1–1.

Shu, Y., Gu, X., Yang, G.-Z., & Lo, B. (2022). Revisiting self-supervised contrastive learning for facial expression recognition.

Smart, B. & Carneiro, G. (2023). Bootstrapping the relationship between images and their clean and noisy labels. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, (pp. 5344–5354).

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., Cubuk, E. D., Kurakin, A., & Li, C.-L. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, *33*.

Song, C., Liu, F., Huang, Y., Wang, L., & Tan, T. (2013). Auto-encoder based data clustering. In *Iberoamerican Congress on Pattern Recognition*, (pp. 117–124). Springer.

Song, F., Tan, X., & Chen, S. (2014). Exploiting relationship between attributes for improved face verification. *Computer Vision and Image Understanding*, *122*, 143–154.

Souri, Y., Noury, E., & Adeli, E. (2017). Deep relative attributes. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Tai-*

wan, November 20-24, 2016, Revised Selected Papers, Part V 13, (pp. 118–133). Springer.

Speer, R., Chin, J., & Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

Srinivas, A., Laskin, M., & Abbeel, P. (2020). Curl: Contrastive unsupervised representations for reinforcement learning. *arXiv preprint arXiv:2004.04136*.

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017). Inception-V4, Inception-Resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conf. on Artificial Intelligence*.

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

Tian, Y., Henaff, O. J., & van den Oord, A. (2021). Divide and contrast: Self-supervised learning from uncurated data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (pp. 10063–10074).

Van den Oord, A., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv e-prints*, arXiv–1807.

Van Gansbeke, W., Vandenhende, S., Georgoulis, S., Proesmans, M., & Van Gool, L. (2020). Scan: Learning to classify images without labels. In *European Conf. on Computer Vision*, (pp. 268–285). Springer.

Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., & Liu, W. (2018). Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 5265–5274).

Wang, J., Zhou, F., Wen, S., Liu, X., & Lin, Y. (2017). Deep metric learning with angular loss. In *Proceedings of the IEEE international conference on computer vision*, (pp. 2593–2601).

Wang, L. & Wong, A. (2020). COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. *arXiv preprint arXiv:2003.09871*.

Wang, Q., Li, W., & Gool, L. V. (2019). Semi-supervised learning by augmented distribution alignment. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 1466–1475.

Wang, X., Deng, X., Fu, Q., Zhou, Q., Feng, J., Ma, H., Liu, W., & Zheng, C. (2020). A weakly-supervised framework for COVID-19 classification and lesion localization from chest CT. *IEEE Transactions on Medical Imaging*, *39*(8), 2615–2625.

Wang, X., Kihara, D., Luo, J., & Qi, G.-J. (2021). Enaet: A self-trained framework for semi-supervised and supervised learning with ensemble transformations. *IEEE Transactions on Image Processing*, *30*, 1639–1647.

Wen, Y., Zhang, K., Li, Z., & Qiao, Y. (2016). A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, (pp. 499–515). Springer.

Wiles, O., Koepke, A. S., & Zisserman, A. (2018). Self-supervised learning of a facial attribute embedding from video.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, *22*(158), 209–212.

Wolff, R. F., Moons, K. G., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., & Mallett, S. (2019). Probast: a tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine*, *170*(1), 51–58.

Wu, Z., Xiong, Y., Yu, S. X., & Lin, D. (2018). Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 3733–3742).

Xu, X., Jiang, X., Ma, C., Du, P., Li, X., Lv, S., Yu, L., Ni, Q., Chen, Y., Su, J., et al. (2020). A deep learning system to screen novel coronavirus disease 2019 pneumonia. *Engineering*, *6*(10), 1122–1129.

Xu, Y., Shang, L., Ye, J., Qian, Q., Li, Y.-F., Sun, B., Li, H., & Jin, R. (2021). Dash: Semi-supervised learning with dynamic thresholding. In *ICML*.

Yan, X., Misra, I., Gupta, A., Ghadiyaram, D., & Mahajan, D. (2020). Clusterfit: Improving generalization of visual representations. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, (pp. 6509–6518).

Yang, J., Parikh, D., & Batra, D. (2016). Joint unsupervised learning of deep representations and image clusters. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, (pp. 5147–5156).

Yang, L., Cheung, N.-M., Li, J., & Fang, J. (2019). Deep clustering by gaussian mixture variational autoencoders with graph embedding. In *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*, (pp. 6440–6449).

Yavuz, M. C., Ali Ahmed, S. A., Kısaağa, M. E., Ocak, H., & Yanıkğlu, B. (2021). Yfcc-celeba face attributes datasets. In *2021 29th Signal Processing and Communications Applications Conf. (SIU)*, (pp. 1–4).

Yavuz, M. C. & Yanikoglu, B. (2022). VCL-PL: semi-supervised learning from noisy web data with variational contrastive learning. In *2022 26th International Conference on Pattern Recognition (ICPR)*, (pp. 740–747). IEEE.

Ye, M., Zhang, X., Yuen, P. C., & Chang, S.-F. (2019). Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 6210–6219).

Yu, X., Lu, S., Guo, L., Wang, S. H., & Zhang, Y. D. (2021). Resgnet-c: A graph convolutional neural network for detection of covid-19. *Neurocomputing*, *452*, 592–605.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., & Deny, S. (2021). Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, (pp. 12310–12320). PMLR.

Zesch, T., Müller, C., & Gurevych, I. (2008). Using wiktionary for computing semantic relatedness. In *AAAI*, volume 8, (pp. 861–866).

Zhai, X., Oliver, A., Kolesnikov, A., & Beyer, L. (2019). S4l: Self-supervised semi-supervised learning. In *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*, (pp. 1476–1485).

Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., & Shinozaki, T. (2020). Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In *Neural Information Processing Systems*.

Zhang, R., Isola, P., & Efros, A. A. (2016). Colorful image colorization. In *European conference on computer vision*, (pp. 649–666). Springer.

Zhang, Y.-D., Zhang, Z., Zhang, X., & Wang, S.-H. (2021). MIDCAN: A multiple input deep convolutional attention network for COVID-19 diagnosis based on chest CT and chest X-ray. *Pattern Recognition Letters*, *150*, 8–16.

Zheltonozhskii, E., Baskin, C., Mendelson, A., Bronstein, A. M., & Litany, O. (2022). Contrast to divide: Self-supervised pre-training for learning with noisy labels. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, (pp. 1657–1667).

Zheng, H., Fu, J., Mei, T., & Luo, J. (2017). Learning multi-attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE Int. Conf. on Computer Vision (ICCV)*, (pp. 5209–5217).

Zhong, Y., Sullivan, J., & Li, H. (2016). Face attribute prediction using off-the-shelf cnn features. In *2016 International Conference on Biometrics (ICB)*, (pp. 1–7). IEEE.

Zhong, Y., Tang, H., Chen, J., Peng, J., & Wang, Y.-X. (2022). Is self-supervised learning more robust than supervised learning? *arXiv preprint arXiv:2206.05259*.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, (pp. 2921–2929).

Zhu, J., Moraes, R. M., Karakulak, S., Sobol, V., Canziani, A., & LeCun, Y. (2022). Tico: Transformation invariance and covariance contrast for self-supervised visual representation learning. *arXiv preprint arXiv:2206.10698*.

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., et al. (2020). A novel coronavirus from patients with pneumonia in China, 2019. *New England Journal of Medicine*.

Zhu, Z., Luo, P., Wang, X., & Tang, X. (2014). Multi-view perceptron: a deep model for learning face identity and view representations. *Advances in neural information processing systems*, *27*.