Designing Efficient Deep Learning Models for Computer-Aided Medical Diagnosis

Hasib Zunair

A Thesis
in
The Concordia Institute
for
Information Systems Engineering

Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Applied Science (Quality Systems Engineering) at
Concordia University
Montreal, QC, Canada

July 2021

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: Hasib Zunair

Entitled: Designing Efficient Deep Learning Models for Computer-Aided Medical Diagnosis.

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (**Quality Systems Engineering**)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____________________________________ Chair
Dr. N. Bouguila

_____________________________________ Examiner
Dr. A. Mohammadi

_____________________________________ Examiner
Dr. N. Bouguila

_____________________________________ Supervisor
Dr. A. Ben Hamza

Approved by _____________________________________
Dr. A. Ben Hamza, Director
Concordia Institute for Information Systems Engineering

____
Dr. M. Debbabi, Dean
Faculty of Engineering and Computer Science

Date _____________________________________

Abstract

Designing Efficient Deep Learning Models for Computer-Aided Medical Diagnosis

Hasib Zunair

Traditional clinician diagnosis, which requires intensive manual effort from experienced medical doctors and radiologists, is notoriously time-consuming, costly and at times error prone. To alleviate these issues, computer-aided diagnosis systems are often used to improve accuracy in early detection, diagnosis, treatment plan and an outcome prediction. While these systems are making strides, significant challenges remain due the scarcity of publicly available data, high annotation cost, and suboptimal performance in detecting rare targets. In this thesis, we design efficient deep learning models for computer-aided medical diagnosis. The contributions are two-fold: First, we introduce an over-sampling method for learning the inter-class mapping between under-represented class samples and over-represented samples in a bid to generate under-represented class samples using unpaired image-to-image translation. These synthetic images are then used as additional training data in the task of detecting abnormalities (i.e. melanoma, COVID-19). Our method achieves improved performance on a standard skin lesion classification benchmark. We show through feature visualization that our approach leads to context based lesion assessment that can reach an expert dermatologist level. Additional experiments also demonstrate the effectiveness of our model in COVID-19 detection from chest radiography images. The synthetic images not only improve performance of various deep learning architectures when used as additional training data under heavy imbalance conditions, but also detect the target class with high confidence.

Second, we present a simple, yet effective end-to-end depthwise encoder-decoder fully convolutional network architecture, dubbed Sharp U-Net, for binary and multi-class biomedical image segmentation. Instead of applying a plain skip connection such as U-Net, a depthwise convolution of the encoder feature map with a sharpening kernel filter is employed prior to merging the encoder and decoder features, thereby producing a sharpened intermediate feature map of the same size as the encoder map. Using this sharpening filter layer, we are able to not only fuse semantically less dissimilar features, but also smooth out artifacts throughout the network layers during the early stages of training. Our extensive experiments on six datasets show that the proposed Sharp U-Net model consistently outperforms or matches the recent state-of-the-art baselines in both binary and multi-class segmentation tasks, while adding no extra learnable parameters.

Ackowledgements

I would like to thank my dear advisor Prof. Abdessamad Ben Hamza. It was my great pleasure and honor to work under his supervision. His insightful guidance, expertise, and wisdom have been truly instrumental to my Masters career. Through all the ups and downs in my Masters career, he has always been supporting me. I would have never got to where I am today without his support, encouragement, and continuous help.

I am also grateful to my family and friends for their constant support, especially during this COVID-19 pandemic, which has immensely helped me to not lose my sanity and be successful in my journey towards a Masters program.

Table of Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

In this chapter, we present the motivation behind this work, followed by the problem statement, objectives of the study, literature review, an overview of deep learning based generative adversarial networks, encoder-decoder networks, and thesis contributions.

1.1 Framework and Motivation

Visual inspection of medical images by dermatologists and radiologist is normally the first step in clinical diagnosis, it is generally followed by dermoscopy or X-ray imaging for further analysis. Even though these imaging techniques provide detailed visual context of regions of interest, it is costly, error prone, and achieves only average sensitivity in detecting certain conditions such as melanoma [1]. This has triggered the need for developing more precise computer-aided diagnostics systems that would assist in early detection of certain conditions from medical images. Despite significant strides in medical image recognition, the recognition process remains a challenging task due to various reasons, including the high degree of visual similarity (i.e. low inter-class variation) between certain conditions (i.e. malignant and benign lesions), making it difficult to distinguish between two during the diagnosis of patients. Also, the contrast variability and boundaries between certain regions owing to image acquisition make automated detection from medical images an intricate task. In addition to the high intra-class variation, texture, shape, size and location in dermoscopic images [2], there are also artifacts such as hair, veins, ruler marks, illumination variation, and color calibration charts that usually cause occlusions and blurriness, further complicating the situation [3].

1.2 Problem Statement

Image classification and segmentation are fundamental problems in medical image analysis.

1.2.1 Image Classification

Image classification is all about labeling images in a dataset and organizing them into a known number of classes so they can be found quickly and efficiently, and the goal is to assign new images to one of these classes. In supervised learning tasks, the available data for classification is usually split into two disjoint subsets: the training set for learning, and the test set for testing. The training and test sets are usually selected by randomly sampling a set of training instances from the available data for learning and using the rest of instances for testing. The performance of a classifier is then assessed by applying it to test data with known target values and comparing the predicted values with the known values.

1.2.2 Image Segmentation

Image segmentation or semantic segmentation refers to the process of classifying each pixel in an image into its semantic class and hence can be regarded as a classification problem at the pixel level. A lung segmentation task, for example, can be thought of as a binary segmentation problem with two semantic classes: lung and background. However, unlike image classification whose goal is to assign an input image to one label from a fixed set of categories or classes, the output in semantic segmentation is an image, typically of the same size as the input image, such that each pixel is classified to a particular class.

For a multi-class segmentation problem consisting of C classes, we denote by $\mathcal{X} = \{(\mathbf{x}_i, y_i) : i = 1, \ldots, N\}$ a set of N samples, where $\mathbf{x}_i$ is the ith training sample and $y_i \in \{1, \ldots, C\}$ is the corresponding true label. The true label of the ith sample can be represented as a one-hot encoding vector $\mathbf{y}_i = (y_{i1}, \ldots, y_{iC})^\intercal$, such that $y_{ic} = 1$ if $i = c$ and 0 otherwise for each class c.

1.3 Objectives

In this thesis, we propose deep learning approaches for imbalanced medical image classification and further fine-grained task such as segmentation.

- For image classification, we generate synthetic images of the under-represented class to balance the training dataset for classifier training. More specifically, we use generative adversarial networks and well as the training data itself for conditional image synthesis to synthesize new images. The objective of image classification is to accurately predict the target class for each image in the dataset.

- For image segmentation, we develop a new encoder-decoder network for more accurate image segmentation for multiple medical image modalities. We build on the popular U-Net architecture by replacing the skip connections with sharpening spatial kernels in an effort to reduce feature mismatch between the encoder and decoder during optimization.

1.4 Literature Review

Deep learning has recently emerged as a very powerful way to hierarchically find abstract patterns using large amounts of training data. The tremendous success of deep neural networks in image classification, for instance, is largely attributed to open source software, inexpensive computing hardware, and the availability of large-scale datasets [4]. Deep learning has proved valuable for various medical image analysis tasks such as classification and segmentation [5–10]. In particular, significant performance gains in melanoma recognition have been achieved by leveraging deep convolutional neural networks in a two-stage framework [11], which uses a fully convolutional residual network for skin lesion segmentation and a very deep residual network for skin lesion classification. However, the issues of low inter-class variation and class imbalance of skin lesion image datasets severely undermine the applicability of deep learning to melanoma detection [11, 12], as they often hinder the model's ability to generalize, leading to over-fitting [13]. In this chapter, we employ conditional image synthesis without paired images to tackle the class imbalance problem by generating synthetic images for the minority class. Built on top of generative adversarial networks (GANs) [14], several image synthesis approaches, both conditional [15] and unconditional [16], have been recently adopted for numerous medical imaging tasks, including melanoma detection [17–19]. Also, approaches that enable the training of diverse models based on distribution matching with both paired and unpaired data were introduced in [20–23]. These approaches include image translation from CT-PET [24], CS-MRI [25], MR-CT [26], XCAT-CT [27] and H&E staining in histopathology [28, 29]. In [30, 31], image synthesis models that synthesize images from noise were developed in an effort to improve melanoma detection. However, Cohen *et al.* [32] showed that the training schemes used in several domain adaptation methods often lead to a high bias and may result in hallucinating features (e.g. adding or removing tumors leading to a semantic change). This is due in large part to the source or target domains consisting of over- or under-represented samples during training (e.g. source domain composed of 50% malignant images and 50% benign; or target domain composed of 20% malignant and 80% benign images).

1.4.1 Generative Adersarial Networks

The advent of generative adversarial networks (GANs) [14] has accelerated research in generative modeling and distribution learning. With the ability to replicate data distributions and synthesize images with high fidelity, GANs have bridged the gap between supervised learning and image generation. These synthetic images can then be used as input to improve the performance of various deep learning algorithms for downstream tasks, such as image classification and segmentation. GANs have not only been used in natural images' settings, but have also been extensively employed in medical image analysis [33], where labels are usually scarce or almost non-existent.

With the scarcity of annotated medical image datasets, there has been a surge of interest in developing efficient approaches for the generation of synthetic medical images. While several existing generative methods have addressed the translation between multiple imaging modalities CT-PET, CS-MRI, MR-CT, XCAT-CT [24–27] based on distribution matching, other approaches have focused on the scarcity of labeled data in the medical field due in large part to the acquisition, privacy and health safety issues. Conditional and unconditional image synthesis procedures, built on top of these generative models, have been proposed in retinal images [34, 35] and MRI scans [36–38]. These models involve the training of paired data in both source and target domains to synthesize realistic, high-resolution images in order to aid in medical image classification and segmentation tasks.

Image synthesis methodologies have also been proposed in the context of chest X-rays [39]. Our work is significantly different in the sense that we are specifically interested in synthesizing a particular class, whereas in [39] X-rays are generated from surface geometry for landmark detection tasks. While some generative methods only require paired data in the source domain with target domain consisting of unlabeled examples, Cohen *et al.* [32] have demonstrated that the phenomenon of *hallucinating features* (e.g. adding or removing tumors leading to a semantic change) leads to a high bias in these domain adaptation techniques.

1.4.2 Encoder-Decoder Networks

The task of biomedical is to label each pixel of an object of interest in biomedical images, and is often used in clinical applications such as computer-aided diagnosis. Recent variants of the U-Net architecture have focused primarily on improving the performance of U-Net by uniformly scaling the network and/or using pre-trained CNN models on the ImageNet dataset as encoders. Zhou *et al.* [40] propose Wide U-Net, which uniformly scales U-Net by increasing the number of filters in both the encoder and decoder subnetworks of U-Net. Also, Zhou *et al.* [40] introduce UNet++ , which consists of an ensemble of U-Nets with varying depths and decoders that are densely connected at the same resolution through redesigned skip connections. In spite of improved performance, the UNet++ model is quite complex, requires additional learnable parameters, and some of its components are redundant for specific tasks [41]. Kalinin *et al.* [42] employ ImageNet pre-trained encoders to further improve the performance of U-Net in angiodysplasia lesion segmentation from wireless capsule endoscopy videos and semantic segmentation of robotic instruments in surgical videos. Inspired by Inception modules that are used in CNNs to allow for more efficient computation, Ibtehaz and Rahman [43] introduce MultiResUNet, an enhanced U-Net architecture, which uses a chain of convolutional layers with residual connections instead of simply concatenating the feature maps from the encoder path to the decoder path. These residual connections not only reduce the semantic gap between the features of the encoder and decoder, but also make the learning easier, while robustly segmenting images from various modalities at different scales.

In light of the availability and quality of medical datasets, there is also a growing interest in developing deep learning frameworks for learning from noisy labels and detecting small anatomical structures with blurry boundaries [44, 45]. Motivated by the performance drop of U-Net in detecting smaller anatomical structures with blurred noisy boundaries, Valanarasu *et al.* [44] propose Ki-Net, an over-complete architecture, which projects the data onto high dimensions. When used in conjunction with U-Net, this network yields improved segmentation performance, while having fewer number of parameters. In order to address the problem of detecting small structures with blurry boundaries, Mirikharaji *et al.* [45] extend the idea of example reweighting in image classification to pixel-level segmentation by training fully convolutional networks from both a large set of weak labels and a small set of expert labels. The idea is to use meta-learning to focus on pixels that have gradients closer to those of expert labels. Ji *et al.* [46] employ a neural architecture search based method for volumetric medical image segmentation by searching for scale-wise feature aggregation strategies and blockwise operators in the encoder-decoder network in an effort to generate better feature representations.

While these variants of the U-Net architecture have shown improved results in biomedical image segmentation, the issue of the large semantic gap between the low- and high-level features of the encoder and decoder subnetworks still remains a daunting task. Our work is significantly different from previous work in the sense that the proposed framework mitigates the problem of feature mismatch between the encoder and decoder subnetworks by replacing skip connections with sharpening spatial filters, resulting in much improved segmentation performance. Moreover, our approach can be applied to any encoder-decoder type network.

1.5 Overview and Contributions

The organization of this thesis is as follows:

- Chapter 1 begins with the motivations and goals for this research, followed by the problem statement, the objective of this study, a literature review with a brief discussion of some algorithms relevant to deep learning in medical image classification and segmentation.

- In Chapter 2, we introduce an integrated deep learning based framework, which couples adversarial training and transfer learning to jointly address inter-class variation and class imbalance for the task of imbalanced medical image classification for multiple modalities [47, 48]. The method synthesizes target class images to adjust the skew in training sets by over-sampling positive cases to mitigate the class imbalance problem, while training classifiers. On a dermatology image analysis benchmark, significant improvements are achieved over several baseline methods for the important task of melanoma detection. The method also enables visual discovery of high activations for the regions surrounding the skin lesion, leading to context based lesion assessment that can reach an expert dermatologist level. In the cases of detecting COVID-19 from chest X-ray images, the method also demonstrates how the data generation procedure can serve as an anonymization tool by achieving comparable detection performance when trained only on synthetic data versus real data in an effort to alleviate privacy concerns.

- In Chapter 3, we propose a novel Sharp U-Net architecture by designing new connections between the encoder and decoder subnetworks using a depthwise convolution of the encoder feature maps with a sharpening spatial filter to address the semantic gap issue between the encoder and decoder features. Sharp U-Net architecture can be scaled for improved performance, outperforming baselines that have three times the number of learnable parameters. We demonstrate through extensive experiments the ability of the proposed model to learn efficient representations for both binary and multi-class segmentation tasks on a variety of medical images from different modalities.

- Chapter 4 presents a summary of the contributions of this thesis, limitations, and outlines several directions for future research in this area of study.

CHAPTER 2

Medical Image Synthesis and Classification

In this chapter, we introduce a a two-stage framework for automatic classification of medical images using adversarial training and transfer learning, specifically for tasks having class imbalance problem. In the first stage, we leverage the inter-class variation of the data distribution for the task of conditional image synthesis by learning the inter-class mapping and synthesizing under-represented class samples (i.e Melanoma/Malignant skin lesion and COVID-19 chest X-ray images) from the over-represented ones (i.e. Benign, Normal and Pneumonia images) using unpaired image-to-image translation. In the second stage, we train a deep convolutional neural network tasked for medical image classification in a supervised fashion using the original training set combined with the newly synthesized under-represented class samples. Experimental results on standard medical image classification benchmarks demonstrate superior performance of the proposed approach compared to existing methods. Motivated by the lack of publicly available datasets of chest radiographs of positive patients with Coronavirus disease 2019 (COVID-19), we make our synthetic dataset consisting of 21,295 synthetic COVID-19 chest X-ray images publicly available to the research community for use.

2.1 Introduction

Melanoma is one of the most aggressive forms of skin cancer [49,50]. It is diagnosed in more than 132,000 people worldwide each year, according to the World Health Organization. Hence, it is essential to detect melanoma early before it spreads to other organs in the body and becomes more difficult to treat.

While visual inspection of suspicious skin lesions by a dermatologist is normallsy the first step in melanoma diagnosis, it is generally followed by dermoscopy imaging for further analysis. Dermoscopy is a noninvasive imaging procedure that acquires a magnified image of a region of the skin at a very high resolution to clearly identify the spots on the skin [51], and helps identify deeper levels of skin, providing more details of the lesions. Moreover, dermoscopy provides detailed visual context of regions of the skin and has proven to enhance the diagnostic accuracy of a naked eye examination, but it is costly, error prone, and achieves only average sensitivity in detecting melanoma [1]. This has triggered the need for developing more precise computer-aided diagnostics systems that would assist in early detection of melanoma from dermoscopy images. Despite significant strides in skin lesion recognition, melanoma detection remains a challenging task due to various reasons, including the high degree of visual similarity (i.e. low inter-class variation) between malignant and benign lesions, making it difficult to distinguish between melanoma and non-melanoma skin lesions during the diagnosis of patients. Also, the contrast variability and boundaries between skin regions owing to image acquisition make automated detection of melanoma an intricate task. In addition to the high intra-class variation of melanoma's color, texture, shape, size and location in dermoscopic images [2], there are also artifacts such as hair, veins, ruler marks, illumination variation, and color calibration charts that usually cause occlusions and blurriness, further complicating the situation [3].

Classification of skin lesion images is a central topic in medical imaging, having a relatively extensive literature. Some of the early methods for classifying melanoma and non-melanoma skin lesions have focused mostly on low-level computer vision approaches, which involve hand-engineering features based on expert knowledge such as color [2], shape [52] and texture [53, 54]. By leveraging feature selection, approaches that use mid-level computer vision techniques have also been shown to achieve improved detection performance [55]. In addition to ensemble classification based techniques [56], other methods include two-stage approaches, which usually involve segmentation of skin lesions, followed by a classification stage to further improve detection performance [1, 54, 55]. However, hand-crafted features often lead to unsatisfactory results on unseen data due to high intra-class variation and visual similarity, as well as the presence of artifacts in dermoscopy images. Moreover, such features are usually designed for specific tasks and do not generalize across different tasks.

Deep learning has recently emerged as a very powerful way to hierarchically find abstract patterns using large amounts of training data. The tremendous success of deep neural networks in image classification, for instance, is largely attributed to open source software, inexpensive computing hardware, and the availability of large-scale datasets [4]. Deep learning has proved valuable for various medical image analysis tasks such as classification and segmentation [5–10]. In particular, significant performance gains in melanoma recognition have been achieved by leveraging deep convolutional neural networks in a two-stage framework [11], which uses a fully convolutional residual network for skin lesion segmentation and a very deep residual network for skin lesion classification. However, the issues of low inter-class variation and class imbalance of skin lesion image datasets severely undermine the applicability of deep learning to melanoma detection [11, 12], as they often hinder the model's ability to generalize, leading to over-fitting [13]. In this chapter, we employ conditional image synthesis without paired images to tackle the class imbalance problem by generating synthetic images for the minority class. Built on top of generative adversarial networks (GANs) [14], several image synthesis approaches, both conditional [15] and unconditional [16], have been recently adopted for numerous medical imaging tasks, including melanoma detection [17–19]. Also, approaches that enable the training of diverse models based on distribution matching with both paired and unpaired data were introduced in [20–23]. These approaches include image translation from CT-PET [24], CS-MRI [25], MR-CT [26], XCAT-CT [27] and H&E staining in histopathology [28, 29]. In [30, 31], image synthesis models that synthesize images from noise were developed in an effort to improve melanoma detection. However, Cohen *et al.* [32] showed that the training schemes used in several domain adaptation methods often lead to a high bias and may result in hallucinating features (e.g. adding or removing tumors leading to a semantic change). This is due in large part to the source or target domains consisting of over- or under-represented samples during training (e.g. source domain composed of 50% malignant images and 50% benign; or target domain composed of 20% malignant and 80% benign images).

The World Health Organization (WHO) has declared COVID-19, the infectious respiratory disease caused by the novel coronavirus, a global pandemic due to the rapid increase in infections worldwide. This virus has spread across the globe, sending billions of people into lockdown, as many countries rush to implement strict measures in an effort to slow COVID-19 spread and flatten the epidemiological curve. Although most people with COVID-19 have mild to moderate symptoms, the disease can cause severe lung complications such as viral pneumonia, which is frequently diagnosed using chest radiography.

Recent studies have shown that chest radiography images such as chest X-rays (CXR) or computed tomography (CT) scans performed on patients with COVID-19 when they arrive at the emergency room can help doctors determine who is at higher risk of severe illness and intubation [57, 58]. These X-rays and CT scans show small patchy translucent white patches (called ground-glass opacities) in the lungs of COVID-19 patients. A chest X-ray provides a two-dimensional (2D) image, while a CT scan has the ability to form three-dimensional (3D) images of the chest. However, chest CT based screening is more expensive, not always available at small or rural hospitals, and often yields a high false-positive rate. Therefore, the need to develop computational approaches for detecting COVID-19 via chest radiography imaging not only can save healthcare a tremendous amount of time and money, but more importantly, it can save more lives [59]. By leveraging deep learning, several approaches for the detection of COVID-19 cases from chest radiography images have been recently proposed, including tailored convolutional neural network (CNN) architectures [60, 61] and transfer learning based methods [62–65].

While promising, the predictive performance of these deep learning based approaches depends heavily on the availability of large amounts of data. However, there is a significant shortage of chest radiology imaging data for COVID-19 positive patients, due largely to several factors, including the rare nature of the radiological finding, legal, privacy, technical, and data-ownership challenges. Moreover, most of the data are not accessible to the global research community.

In recent years, there have been several efforts to build large-scale annotated datasets for chest X-rays and make them publicly available to the global research community [66–70]. At the time of writing, there exists, however, only one annotated COVID-19 X-ray image dataset [71], which is a curated collection of X-ray images of patients who are positive or suspected of COVID-19 or other viral and bacterial pneumonia. This COVID-19 image data collection has been used as a primary source for positive cases of COVID-19 [60–63], where the detection of COVID-19 is formulated as a classification problem. While the COVID-19 image data collection contains positive examples of COVID-19, the negative examples were acquired from publicly available sources [69] and merged together for data-driven analytics. This fusion of multiple datasets results in predominantly negative examples with only a small percentage of positive ones, giving rise to a class imbalance problem [66–70]. This in turn becomes a challenge of its own, as the merged data becomes highly imbalanced. In the context of a classifier training, the class imbalance problem in the training data distribution yields sub-optimal performance on the minority class (i.e. positive class for COVID-19).

In this chapter, we introduce a two-stage deep neural network based framework for imbalanced medical image classification, to overcome the aforementioned issues. Our approach mitigates the bias problem [32], while improving detection performance and reducing over-fitting. The proposed framework consists of two integrated stages. In the first stage, we generate synthetic images for the minority class (i.e. malignant dermoscopic and COVID-19 infected chest X-ray images) using unpaired image-to-image translation [20] in a bid to balance the training set. These additional images are then used to boost training. In the second stage, we train a deep convolutional neural network classifier by minimizing the focal loss function, which assists the classification model in learning from hard examples, while down-weighting the easy ones. In addition to demonstrating improved melanoma and COVID-19 detection performance through the use of various deep convolutional neural network architectures on the synthetic data to boost training, we show how the proposed data generation and evaluation pipeline can serve as a viable data-driven solution to medical image analysis problems. To aid in COVID-19 CXR machine learning research we make our dataset publicly available, which is currently comprised of 21,295 synthetic images of chest X-rays for COVID-19 positive cases.

The main contributions of this chapter can be summarized as follows:

- We propose an integrated deep learning based framework, which couples adversarial training and transfer learning to jointly address inter-class variation and class imbalance for the task of medical image classification.

- We train a deep convolutional network by iteratively minimizing the focal loss function, which assists the model in learning from hard examples, while down-weighting the easy ones.

- We show experimentally on standard medical image analysis benchmarks significant improvements over several baseline methods for the important task of melanoma and COVID-19 detection.

- We show how our method enables visual discovery of high activations for the regions surrounding the skin lesion, leading to context based lesion assessment that can reach an expert dermatologist level.

- We synthesize chest X-ray images of COVID-19 to adjust the skew in training sets by over-sampling positive cases to mitigate the class imbalance problem, while training classifiers.

- We demonstrate how the data generation procedure can serve as an anonymization tool by achieving comparable detection performance when trained only on synthetic data versus real data in an effort to alleviate privacy concerns.

The rest of this chapter is organized as follows. In Section 2.2, we present a generative framework, which couples adversarial training and transfer learning to jointly address inter-class variation and class imbalance. In Section 2.3, we present experimental results to demonstrate improved melanoma and COVID-19 detection performance through the use of various deep convolutional neural network architectures on the generated data.

2.2 Method

In this section, we describe the main components and algorithmic steps of the proposed approach to imbalanced medical image classification. Our image synthesis framework builds on image-to-image translation, which is an increasingly popular machine learning paradigm that has shown great promise in a wide range of applications, including computer graphics, style transfer, satellite imagery, object transfiguration, character animation, and photo enhancement. In an typical image-to-image translation problem, the objective is to learn a mapping that translates an image in one domain to a corresponding image in another domain using approaches that leverage paired or unpaired training samples. The latter is the focus of our work. While paired image-to-image translation methods use pairs of corresponding images in different domains, the paired training samples are, however, not always available. By contrast, the unpaired image-to-image translation problem, in which training samples are readily available, is more common and practical, but it is highly under-constrained and fraught with challenges. In our work, we build upon the idea that there exist no paired training samples showing how an image from one domain can be translated to a corresponding image in another domain. The task is to generate COVID-19 chest X-rays from chest X-ray images to address COVID-19 class imbalance problem. More specifically, our goal is to learn a mapping function between Non-COVID-19 images and COVID-19 in order to generate COVID-19 chest X-rays without paired training samples in an unsupervised fashion.

2.2.1 Conditional Image Synthesis

In order to tackle the challenging issue of low inter-class variation in skin lesion datasets [11, 13], we partition the inter-classes into two domains for conditional image synthesis with the goal to generate malignant lesions from benign lesions. This data generation process for the malignant minority class is performed in an effort to mitigate the class imbalance problem, as it is relatively easy to learn a transformation with given prior knowledge or conditioning for a narrowly defined task [26, 29]. Also, using unconditional image synthesis to generate data of a target distribution from noise often leads to artifacts and may result in training instabilities [72]. In recent years, various methods based on generative adversarial networks (GANs) have been used to tackle the conditional image synthesis problem, but most of them use paired training data for image-to-image translation [33], which requires the generation of a new image that is a controlled modification of a given image. Due to the unavailability of datasets consisting of paired examples for melanoma detection, we use cycle-consistent adversarial networks (CycleGAN), a technique that involves the automatic training of image-to-image translation models without paired examples [20]. These models are trained in an unsupervised fashion using a collection of images from the source and target domains. CycleGAN is a framework for training image-to-image translation models by learning mapping functions between two domains using the GAN model architecture in conjunction with cycle consistency. The idea behind cycle consistency is to ward off the learned mappings between these two domains from contradicting each other.

Given two image domains B and M denoting benign and malignant, respectively, the CycleGAN framework aims to learn to translate images of one type to another using two generators $G_B : B \to M$ and $G_M : M \to B$, and two discriminators D_M and D_B, as illustrated in Figure 2.1. The generator G_B (resp. G_M) translates images from benign to malignant (resp. malignant to benign), while the discriminator D_M (resp. D_B) scores how real an image of M (resp. B) looks. In other words, these discriminator models are used to determine how plausible the generated images are and update the generator models accordingly. The objective function of CycleGAN is defined as

$$\begin{aligned}\mathcal{L}(G_B, G_M, D_M, D_B) = {} & \mathcal{L}_{GAN}(G_B, D_M, B, M) \\ & + \mathcal{L}_{GAN}(G_M, D_B, M, B) \\ & + \lambda\, \mathcal{L}_{cyc}(G_B, G_M),\end{aligned} \tag{2.1}$$

which consists of two adversarial loss functions and a cycle consistency loss function regularized by a hyper-parameter λ that controls the relative importance of these loss functions [20]. The first adversarial loss is given by

$$\begin{aligned}\mathcal{L}_{GAN}(G_B, D_M, B, M) = {} & \mathbb{E}_{m \sim p_{\text{data}}(m)}[\log D_M(m)] \\ & + \mathbb{E}_{b \sim p_{\text{data}}(b)}[\log(1 - D_M(G_B(b)))],\end{aligned} \tag{2.2}$$

where the generator G_B tries to generate images $G_B(b)$ that look similar to malignant images, while D_M aims to distinguish between generated samples $G_B(b)$ and real samples m. During the training, as G_B generates a malignant lesion, D_M verifies if the translated image is actually a real malignant lesion or a generated one. The data distributions of benign and malignant are $p_{\text{data}}(b)$ and $p_{\text{data}}(m)$, respectively. Similarly, the second adversarial loss is given by

$$\begin{aligned}\mathcal{L}_{GAN}(G_M, D_B, M, B) = {} & \mathbb{E}_{b \sim p_{\text{data}}(b)}[\log D_B(b)] \\ & + \mathbb{E}_{m \sim p_{\text{data}}(m)}[\log(1 - D_B(G_M(m)))],\end{aligned} \tag{2.3}$$

where G_M takes a malignant image m from M as input, and tries to generate a realistic image $G_M(m)$ in B that tricks the discriminator D_B. Hence, the goal of G_M is to generate a benign lesion such that it fools the discriminator D_B to label it as a real benign lesion.

The third loss function is the cycle consistency loss given by

$$\begin{aligned}\mathcal{L}_{cyc}(G_B, G_M) = {} & \mathbb{E}_{b \sim p_{\text{data}}(b)}[\|G_M(G_B(b)) - b\|_1] \\ & + \mathbb{E}_{m \sim p_{\text{data}}(m)}[\|G_B(G_M(m)) - m\|_1],\end{aligned} \tag{2.4}$$

which basically quantifies the difference between the input image and the generated one using the ℓ_1-norm. The idea of the cycle consistency loss it to enforce $G_M(G_B(b)) \approx b$ and $G_B(G_M(m)) \approx m$. In other words, the objective of CycleGAN is to learn two bijective generator mappings by solving the following optimization problem

$$G_B^*, G_M^* = \arg \min_{G_B, G_M} \max_{D_B, D_M} \mathcal{L}(G_B, G_M, D_M, D_B). \tag{2.5}$$

We adopt the U-Net architecture [5] for the generators and PatchGAN [21] for the discriminators. The U-Net architecture consists of an encoder subnetwork and decoder subnetwork that are connected by a bridge section, while PatchGAN is basically a convolutional neural network classifier that determines whether an image patch is real or fake.

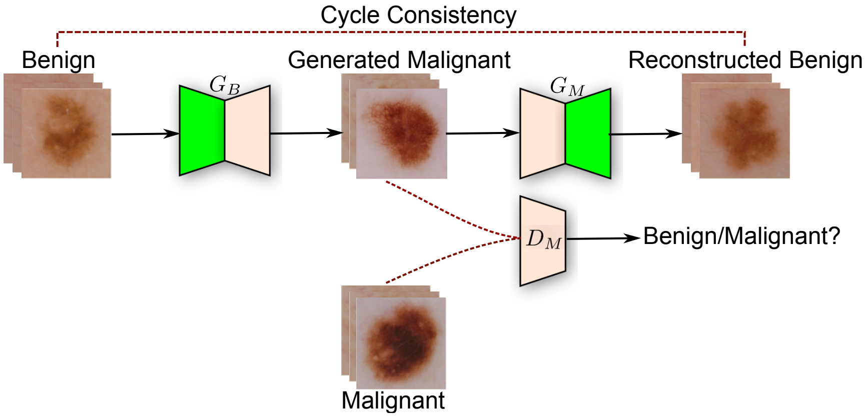


Figure 2.1: Illustration of the generative adversarial training process for unpaired image-to-image translation. Lesions are translated from benign to malignant and then back to benign to ensure cycle consistency in the forward pass. The same procedure is applied in the backward pass from malignant to benign.

2.2.2 Pre-trained Model Architecture

Due to limited training data, it is standard practice to leverage deep learning models that were pre-trained on large datasets [73]. The proposed melanoma classification model uses the pre-trained VGG-16 convolutional neural network without the fully connected (FC) layers, as illustrated in Figure 2.2. The VGG-16 network consists of 16 layers with learnable weights: 13 convolutional layers, and 3 fully connected layers [74]. As shown in Figure 2.2, the proposed architecture, dubbed VGG-GAP, consists of five blocks of convolutional layers, followed by a global average pooling (GAP) layer. Each of the first and second convolutional blocks is comprised of two convolutional layers with 64 and 128 filters, respectively. Similarly, each of the third, fourth and fifth convolutional blocks consists of three convolutional layers with 256, 512, and 512 filters, respectively. The GAP layer, which is widely used in classification tasks, computes the average output of each feature map in the previous layer and helps minimize overfitting by reducing the total number of parameters in the model. GAP turns a feature map into a single number by taking the average of the numbers in that feature map. Similar to max pooling layers, GAP layers have no trainable parameters and are used to reduce the spatial dimensions of a three-dimensional tensor. The GAP layer is followed by a single FC layer with a softmax function (i.e. a dense softmax layer of two units for the binary classification case) that yields the probabilities of predicted classes.

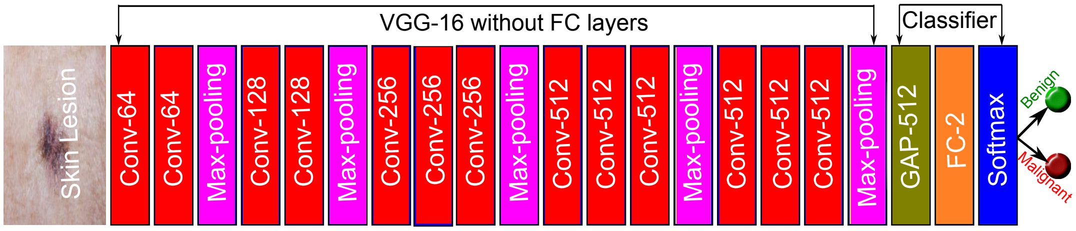


Figure 2.2: VGG-GAP architecture with a GAP layer, followed by an FC layer that in turn is fed into a softmax layer of two units.

Since we are addressing a binary classification problem with imbalanced data, we learn the weights of the VGG-GAP network by minimizing the focal loss function [75] defined as

$$\mathrm{FL}(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t), \tag{2.6}$$

where p_t and α_t are given by

$$p_t = \begin{cases} p & \text{if } y = 1, \\ 1 - p & \text{otherwise} \end{cases} \quad \text{and} \quad \alpha_t = \begin{cases} \alpha & \text{if } y = 1, \\ 1 - \alpha & \text{otherwise,} \end{cases}$$

with $y \in \{-1, 1\}$ denoting the ground truth for negative and positive classes, and $p \in [0, 1]$ denoting the model's predicted probability for the class with label $y = 1$. The weight parameter $\alpha \in [0, 1]$ balances the importance of positive and negative labeled samples, while the nonnegative tunable focusing parameter γ smoothly adjusts the rate at which easy examples are down-weighted. Note that when $\gamma = 0$, the focal loss function reduces to the cross-entropy loss. A positive value of the focusing parameter decreases the relative loss for well-classified examples, focusing more on hard, misclassified examples.

Intuitively, the focal loss function penalizes hard-to-classify examples. It basically down-weights the loss for well-classified examples so that their contribution to the total loss is small even if their number is large.

2.2.3 Algorithm

The main algorithmic steps of our approach are summarized in Algorithm 1. The input is a training set consisting of skin lesion dermoscopic images, along with their associated class labels. In the first stage, the different classes are grouped together (e.g. for binary classification, we have two groups), and we resize each image to $256 \times 256 \times 3$. Then, we balance the inter-class data samples by performing undersampling. We train CycleGAN to learn a function of the interclass variation between the two groups, i.e. we learn a transformation between melanoma and non-melanoma lesions. We apply CycleGAN to the over-represented class samples in order to synthesize the target class samples (i.e. under-represented class). After this transformation is applied, we acquire a balanced dataset, composed of original training data and generated data. In the second stage, we employ the VGG-GAP classifier with the focal loss function. Finally, we evaluate the trained model on the test set to generate the predicted class labels.

Algorithm 1 MelaNet classifier

```
Input: Training set $\mathcal{D} = \{(\mathbf{I}_1, y_1), \dots, (\mathbf{I}_n, y_n)\}$ of dermoscopic images, where $y_i$ is a class label
    of the input $\mathbf{I}_i$.
Output: Vector $\hat{\mathbf{y}}$ containing predicted class labels.
for $i = 1$ to $n$ do
    Group each lesion image according to class label.
    Resize each image to $256 \times 256 \times 3$.
end for
Balance the inter-class data samples.
Train CycleGAN on unpaired and balanced interclass data.
for $i = 1$ to $n$ do
    if class label benign then
        Translate to malignant using the generator network
    else
        pass
    end if
end for
Merge synthetic under-represented class outputs and original training set.
Shuffle.
Train VGG-GAP on the balanced training set
Evaluate the model on the test set and generate predicted class labels.
```

We formulate the detection of COVID-19 as a binary classification problem. For the Normal vs. COVID-19 and Pneumonia vs. COVID-19 tasks, we train two translation models and synthesize COVID-19 images for each task in order to adjust the skew in the training data by over-sampling the minority class. For the sake of clarity and unless otherwise expressly indicated, we refer to the source domain of the two tasks as *Non-COVID-19* instead of *Normal* and *Pneumonia* separately.

We use Algorithm 1 to translate Non-COVID-19 images for each case (i.e. normal or pneumonia) to COVID-19 and then train a classifier on the over-sampled training set.

2.2.4 Training procedure

Since we are tackling a binary classification problem with imbalanced data, we use the focal loss function for the training of the VGG-GAP model. The focal loss is designed to address class imbalance problem by down-weighting easy examples, and focusing more on training the hard examples. Fine-tuning is essentially performed through re-training the whole VGG-GAP network by iteratively minimizing the focal loss function.

For COVID-19 detection, the training for the generators and discriminators are carried out in the same way. First, we balance the inter-class data samples by performing undersampling. Then, we train Cycle-GAN to learn a function of the interclass variation between the two groups, i.e. we learn a transformation between Non-COVID-19 and COVID-19 radiographs. We apply CycleGAN to the over-represented class samples in order to synthesize the target class samples (i.e. under-represented class).

After training, we apply the generators G_A and G_B on the training datasets of Normal vs. COVID-19 and Pneumonia vs. COVID-19. We apply G_A on the majority class of Normal vs. COVID-19, which consists of normal images in order to synthesize 16,537 COVID-19 images. We denote this synthesized dataset as $\mathcal{G}_{NC}$, which consists of generated images by performing image-to-image translation from normal to COVID-19.

Similarly, for Pneumonia vs. COVID-19, we synthesize 4,758 COVID-19 images by applying G_B on the majority class consisting of pneumonia images and we denote the synthesized dataset as $\mathcal{G}_{PC}$, which is comprised of generated images by performing image-to-image translation from pneumonia to COVID-19.

2.3 Experiments

In this section, extensive experiments are conducted to evaluate the performance of the proposed two-stage approach on standard benchmark dataset for imbalanced medical image classification. The source code to reproduce the experimental results is made publicly available on GitHub[1]. An online demo of the model is also available[2]. We also make our synthetic dataset publicly available[3].

[1] https://github.com/hasibzunair/adversarial-lesions

[2] aiderm.herokuapp.com/

[3] https://github.com/hasibzunair/synthetic-covid-cxr-dataset

2.3.1 Datasets

ISIC-2016 Skin Lesion Benchmark. The effectiveness of MelaNet is evaluated on the ISIC-2016 dataset, a publicly accessible dermatology image analysis benchmark challenge for skin lesion analysis towards melanoma detection [76], which leverages annotated skin lesion images from the International Skin Imaging Collaboration (ISIC) archive. The dataset contains a representative mix of images of both malignant and benign skin lesions, which were randomly partitioned into training and test sets, with 900 images in the training set and 379 images in the test set, respectively. These images consist of different types of textures in both background and foreground, and also have poor contrast, making the task of melanoma detection a challenging problem. It is also noteworthy to mention that in the training set, there are 727 benign cases and only 173 malignant cases, resulting in an inter-class ratio of 1:4. Sample benign and malignant images from the ISIC-2016 dataset are depicted in Figure 2.3, which shows that both categories have a high visual similarity, making the task of melanoma detection quite arduous. Note that there is a high intra-class variation among the malignant samples. These variations include color, texture and shape. On the other hand, it is important to point out that benign samples are not visually very different, and hence they exhibit low inter-class variation. Furthermore, there are artifacts present in the images such as ruler markers and fine hair, which cause occlusions. Notice that most malignant images show more diffuse boundaries owing to the possibility that before image acquisition, the patient was already diagnosed with melanoma and the medical personnel acquired the dermoscopic images at a deeper level in order to better differentiate between the benign and malignant classes.

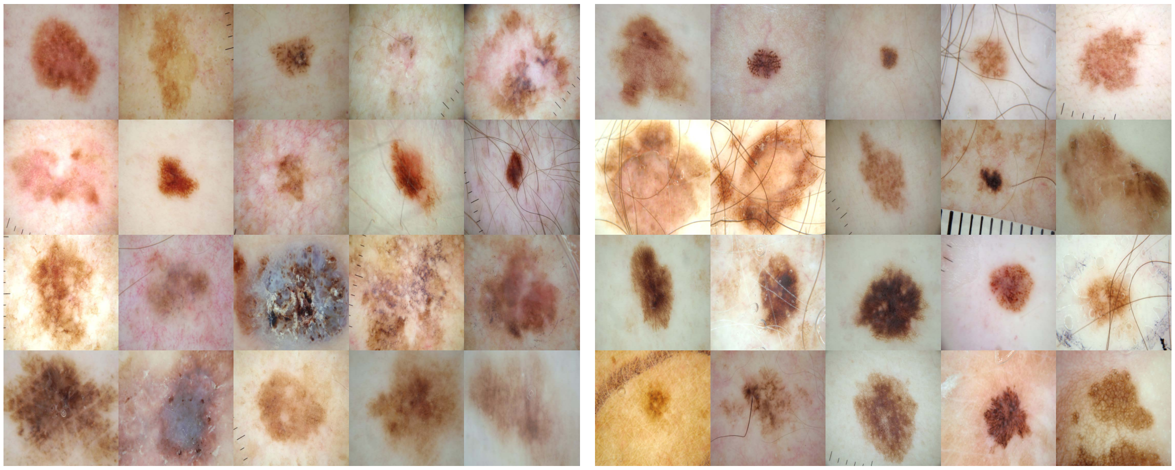


Figure 2.3: Sample malignant and benign images from the ISIC-2016 dataset. Notice a high intra-class variation among the malignant samples (left), while benign samples (right) are not visually very different.

The histogram of the training data is displayed in Figure 2.4, showing the class imbalance problem, where the number of images belonging to the minority class ("malignant") is far smaller than the number of the images belonging to the majority class ("benign"). Also, the number of benign and malignant cases in the test set are 304 and 75, respectively, with an inter-class ratio of 1:4.

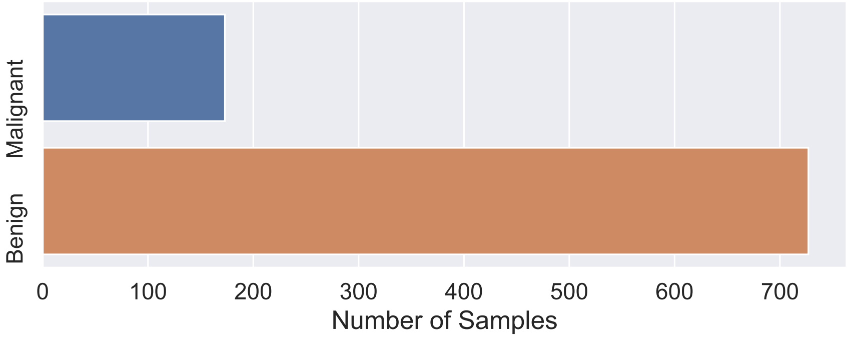


Figure 2.4: Histogram of the ISIC-2016 training set, showing the class imbalance between malignant and benign cases.

Since the images in the ISIC-2016 dataset are of varying sizes, we resize them to 256×256 pixels after applying padding to make them square in order to retain the original aspect ratio.

We use two publicly available datasets of chest X-rays:

COVID-19 Image Data Collection. This dataset comprises 226 images of pneumonia cases with chest X-ray or CT images, specifically COVID-19 cases as well as MERS, SARS, and ARDS. Data are scraped from publications and websites such as Radiopaedia.org, Italian Society of Medical and Interventional Radiology[4], and Figure1.com[5]. From this dataset, we discard the CT images and retain the 226 images positive for COVID-19 and their corresponding labels.

RSNA Pneumonia Detection Challenge. This dataset originated from a Kaggle challenge[6] and consists of publicly available data from [69]. It is composed of 26684 images, and each image was annotated by a radiologist for the presence of lung opacity; thereby providing a label for two classes. This label is included as both lung opacity and pneumonia.

Dataset Splits and Preprocessing. In order to achieve faster convergence, feature standardization is usually performed, i.e. we rescale the images to have values between 0 and 1. Given a data matrix $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\intercal$, the standardized feature vector is given by

$$\mathbf{z}_i = \frac{\mathbf{x}_i - \min(\mathbf{x}_i)}{\max(\mathbf{x}_i) - \min(\mathbf{x}_i)}, \quad i = 1, \dots, n, \tag{2.7}$$

[4] https://www.sirm.org/category/senza-categoria/covid-19/

[5] https://www.figure1.com/covid-19-clinical-cases

[6] https://www.kaggle.com/c/rsna-pneumonia-detection-challenge

where $\mathbf{x}_i$ is the i-th input data point, denoting a row vector. It is important to note that in our approach, no domain specific or application specific pre-processing or post-processing is employed.

On the other hand, data augmentation is usually carried out on medical datasets to improve performance in classification tasks [16,77]. This is often done by creating modified versions of the input images in a dataset through random transformations, including horizontal and vertical flip, Gaussian noise, brightness and zoom augmentation, horizontal and vertical shift, sampling noise once per pixel, color space conversion, and rotation.

We do not perform on-the-fly data augmentation (random) during training, as it may add an unnecessary layer of complexity to training and evaluation. When designing our configurations, we first augment the data offline and then we train the classifier using the augmented data. Also, we do not apply data augmentation in the proposed two-stage approach, as it would not give us an insight on which of the two approaches has more contribution in the performance (data augmentation or image synthesis?). Hence, we keep these two configurations independent from each other.

We partition the three classes from COVID-19 Image Data Collection and RSNA Pneumonia Detection Challenge into two sets, namely "Normal vs. COVID-19" and "Pneumonia vs COVID-19". A patient level split is then applied using 80% as training set and the remaining 20% as test set to assess algorithm performance, and we follow the same evaluation protocol laid out in [36,47]. We define the skew ratio as follows:

$$\text{Skew} = \frac{\text{Negative Examples}}{\text{Positive Examples}}, \tag{2.8}$$

where Skew $= 1$ represents a fully balanced dataset, Skew > 1 shows that the negative samples are the majority, and Skew < 1 represents positive sample dominance in the distribution.

The data distributions of Normal vs. COVID-19 and Pneumonia vs. COVID-19 are displayed in Figure 2.5, which illustrates the class imbalance in the training dataset. For Pneumonia vs. COVID-19, the skew ratio is around 22.9, while the skew for Normal vs. COVID-19 is almost four times larger, indicating high imbalance in the classes.

We also resize all images to 256×256 pixels, and scale the pixel values to $[0, 1]$ for the training of classifiers. It is important to mention that when we use the term *synthetic data*, we refer to COVID-19 CXR images only.

2.3.2 Baselines

We compare the proposed MelaNet approach against VGG-GAP, VGG-GAP + Augment-5x, and VGG-GAP + Augment-10x. The VGG-GAP network is trained on the original training set, which consists of 900 samples. The VGG-GAP + Augment-5x model uses the same VGG-GAP architecture, but is trained on an augmented dataset composed of 5400 training samples, i.e. we increase

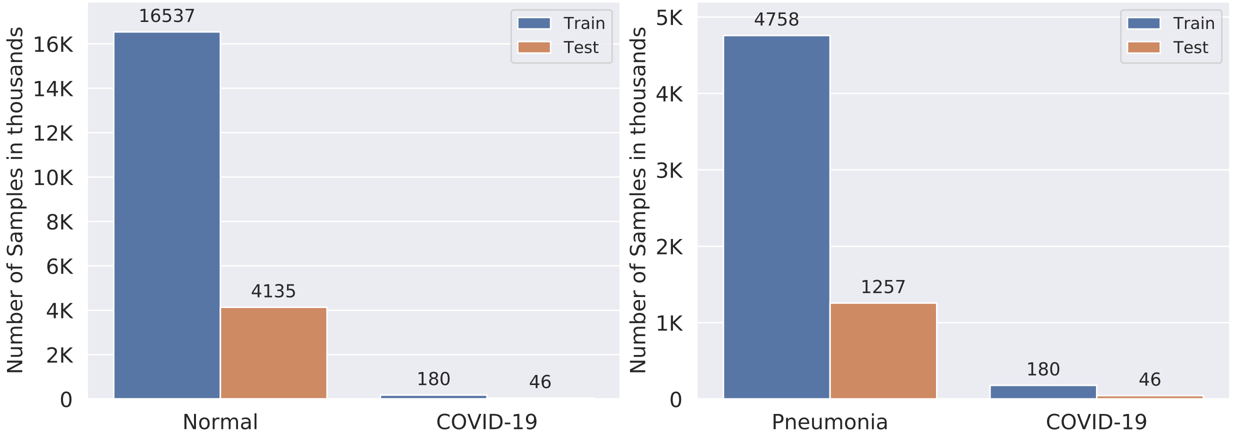


Figure 2.5: Data distributions of Normal vs. COVID-19 (top) and Pneumonia vs. COVID-19 (bottom) with skew ratios of 91.87 and 22.9, respectively.

the training set 5 times from 900 to 5400 samples using image augmentation. Similarly, the VGG-GAP + Augment-10x network is trained on an augmented set of 99000 training samples (i.e. 10 times the original set). We also ran experiments with augmented training sets higher than 10x the original one, but we did not observe improved performance as the network tends to learn redundant representations.

For COVID-19 detection, since our primary goal is to provide a dataset to be used as a training set for the minority class, we test the effectiveness of several deep CNN architectures, including VGG-16 [74], ResNet-50 [78] and DenseNet-102 [79], on the detection of the minority class. These pretrained networks were trained on more than a million images from the ImageNet database[7]. More specifically, we investigate the contribution of the synthetic datasets $\mathcal{G}_{NC}$ and $\mathcal{G}_{PC}$, which consist of COVID-19 CXR images, to the overall performance of these deep learning models. The last layer of each of these models consists of a global average pooling (GAP) layer, which computes the average output of each feature map in the previous layer and helps minimize overfitting by reducing the total number of parameters in the model. The GAP layer turns a feature map into a single number by taking the average of the numbers in that feature map. Similar to max-pooling layers, GAP layers have no trainable parameters and are used to reduce the spatial dimensions of a 3D tensor. The GAP layer is followed by a single fully connected (FC) layer with a softmax function (i.e. a dense softmax layer of two units for the binary classification case), which yields the predicted classes' probabilities that sum to one.

[7] http://www.image-net.org

2.3.3 Implementation Details

All experiments are carried out on a Linux server with 2x Intel Xeon E5-2650 V4 Broadwell @ 2.2GHz, 256 GB RAM, 4x NVIDIA P100 Pascal (12G HBM2 memory) GPU cards. The algorithms are implemented in Keras with TensorFlow backend.

We train CycleGAN for 500 epochs using Adam optimizer [80] with learning rate 0.0002 and batch size 1. We set the regularization parameter λ to 10. The VGG-GAP classifier, on the other hand, is trained using Adadelta optimizer [81] with learning rate 0.001 and mini-batch 16. A factor of 0.1 is used to reduce the learning rate once the loss stagnates. For the VGG-GAP model, we set the focal loss parameters to $\alpha = 0.25$ and $\gamma = 2$, meaning that $\alpha_t = 0.25$ for positive labeled samples, and $\alpha_t = 0.75$ for negative labeled samples. Training of VGG-GAP is continued on all network layers until the focal loss stops improving, and then the best weights are retained. For fair comparison, use used the same set of hyper-parameters for VGG-GAP and baseline methods. We choose Adadelta as an optimizer due to its robustness to noisy gradient information and minimal computational overhead.

For COVID-19 detection, we perform training/testing on both COVID-19 Image Data Collection and RSNA Pneumonia Detection Challenge. For training the models, we use the Adadelta optimization algorithm [81] to minimize the binary cross-entropy loss function with a learning rate of 0.001 and batch size of 16. We initialize the weights using ImageNet and train all layers until the loss stagnates using an early stopping mechanism. For each dataset, we follow the same evaluation protocol laid out in [36] for testing the contribution of newly added data. In this evaluation protocol, both training and test sets are used. The training set varies, as new data are added to each configuration. The deep CNN classifiers are trained on this data and evaluated on the held-out test set. For fair evaluation and comparison purposes, the size of the test set remains constant. It is important to mention that the test set does not contain any synthetic examples. Moreover, the hyper-parameters are not tuned and hence do not require a separate validation set.

2.3.4 Evaluation Metrics

The effectiveness of the proposed classifier is assessed by conducting a comprehensive comparison with the baseline methods using several performance evaluation metrics [10, 11, 82], including the receiver operating characteristic (ROC) curve, sensitivity, and the area under the ROC curve (AUC). Sensitivity is defined as the percentage of positive instances correctly classified, i.e.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{2.9}$$

where TP, FP, TN and FN denote true positives, false positives, true negatives and false negatives, respectively. TP is the number of correctly predicted malignant lesions, while TN is the number of correctly predicted benign lesions. A classifier that reduces FN (ruling cancer out in cases that do have it) and FP (wrongly diagnosing cancer where there is none) indicates a better performance. Sensitivity, also known as recall or true positive rate (TPR), indicates how often a classifier misses a positive prediction. It is one of the most common measures to evaluate a classifier in medical image classification tasks [83]. We use a threshold of 0.5.

Another common metric is AUC that summarizes the information contained in the ROC curve, which plots TPR versus $\text{FPR} = \text{FP}/(\text{FP}+\text{TN})$, the false positive rate, at various thresholds. Larger AUC values indicate better performance at distinguishing between melonoma and non-melanoma images. It is worth pointing out that the accuracy metric is not used in this study, as it provides no interpretable information and may lead to a false sense of superiority of classifying the majority class.

Due to high class imbalance in the datasets, the choice of evaluation metrics plays a vital role in the comparison of classifiers. Threshold metrics such as accuracy and rank metrics (e.g. area under the ROC curve) may lead to a false sense of superiority and mask poor performance [84], thereby introducing bias. Since we are interested in the detection of the minority class (COVID-19), we follow the recommendations provided in [84, 85] and perform quantitative evaluations using sensitivity and false negatives in the same vein as [62]. Sensitivity is the percentage of positive instances correctly classified and is defined as

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{2.10}$$

where TP, FP, TN and FN denote true positives, false positives, true negatives and false negatives, respectively. TP is the number of correctly predicted malignant lesions, while TN is the number of correctly predicted benign lesions. A classifier that reduces FN (ruling COVID-19 out in cases that do have it) and FP (wrongly diagnosing COVID-19 where there is none) indicates a better performance. A false negative COVID-19 result can be a serious problem due to the fact that we lose the benefits of early intervention. A false positive result can also cause significant issues for both an individual and the community. Even from an epidemiologicial perspective, a high number of false positives can lead to a wrong understanding of the spread of COVID-19 in the community. Sensitivity, also known as recall or true positive rate, indicates how often a classifier misses a positive prediction. It is one of the most common measures to evaluate a classifier in medical image classification tasks [83]. A larger value of Sensitivity indicates a better performance of the classification model.

2.3.5 Melanoma Detection Results

In this section, we conduct extensive experiments to evaluate the performance of the proposed framework for the task of melanoma detection from dermascopic images.

Few synthetic images are better than many augmented images. The performance comparison results of MelaNet and the baseline methods using AUC, FN and Sensitivity are depicted in Figure 2.6. We observe that our approach outperforms the baselines, achieving an AUC of 81.18% and a sensitivity of 91.76% with performance improvements of 2.1% and 7.3% over the VGG-GAP baseline. Interestingly, MelaNet yields the lowest number of false negatives, which were reduced by more than 50% compared to the baseline methods, meaning it picked up on malignant cases that the baselines had missed. In other words, MelaNet caught instances of melanoma that would have otherwise gone undetected. This is a significant performance in the potential for early melanoma detection, albeit MelaNet was trained on only 1627 samples composed of 900 images from the original dataset and 727 synthesized images (benign and malignant) obtained via generative adversarial training.

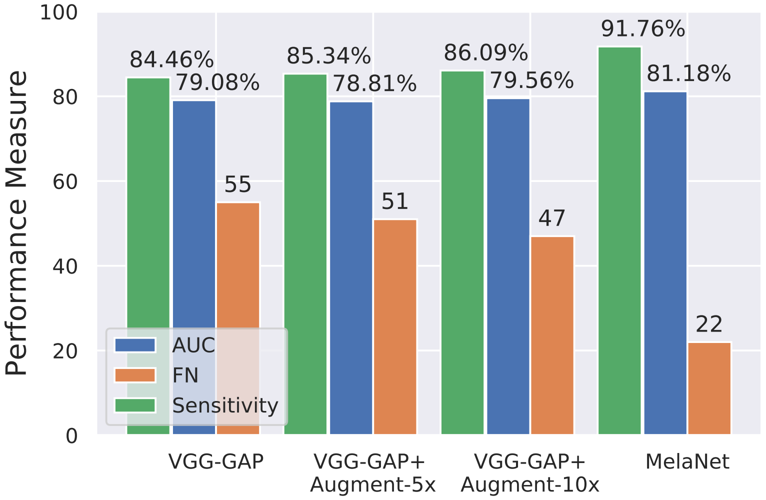


Figure 2.6: Classification performance of MelaNet and the baseline methods using AUC, FN and Sensitivity as evaluation metrics on the ISIC-2016 test set.

Figure 2.7 displays the ROC curve, which shows the better performance of our proposed MelaNet approach compared to the baseline methods. Each point on ROC represents different trade-off between false positives and false negatives. An ROC curve that is closer to the upper right indicates a better performance (TPR is higher than FPR). Even though during the early and last stages, the ROC curve of MelaNet seems to fluctuate at certain points, the overall performance is much higher than the baselines, as indicated by the AUC value. This better performance demon-

Table 2.1: Classification evaluation results of MelaNet and baseline methods. Boldface numbers indicate the best performance.

	Performance Measures		
Method	AUC (%)	Sensitivity (%)	FN
Gutman *et al.* [10]	80.40	50.70	–
Yu *et al.* [11] (*without segmentation*)	78.20	42.70	–
Yu *et al.* [11] (*with segmentation*)	78.30	54.70	–
VGG-GAP	79.08	84.46	55
VGG-GAP + Augment-5x (*ours*)	78.81	85.34	51
VGG-GAP + Augment-10x (*ours*)	79.56	86.09	47
MelaNet (*ours*)	**81.18**	**91.76**	**22**

strates that the conditional image synthesis procedure plays a crucial role and enables our model to learn effective representations, while mitigating data scarcity and class imbalance.

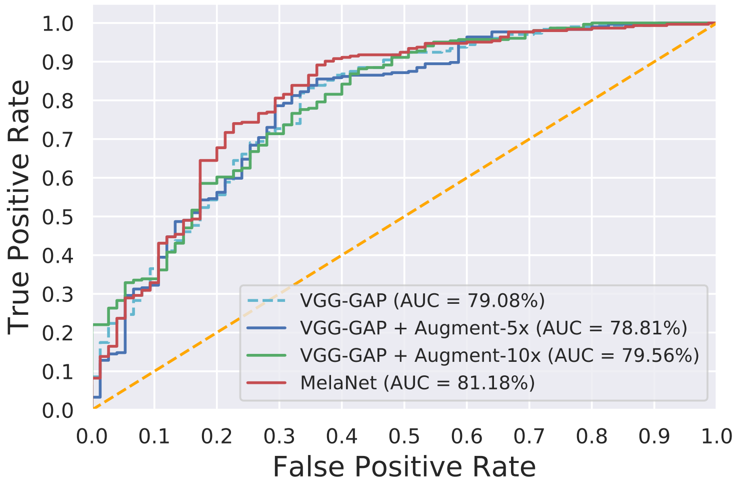


Figure 2.7: ROC curves for MelaNet and baseline methods, along with the corresponding AUC values.

MelaNet outperforms current state-of-the-art. We also compare MelaNet to two other standard baseline methods [10, 11]. The top evaluation results on the ISIC-2016 dataset to classify images as either being benign or malignant are reported in [10]. The method presented in [11] is also a two-stage approach consisting of a fully convolutional residual network for skin lesion segmentation, followed by a very deep residual network for skin lesion classification. The classification results are displayed in Table 2.2, which shows that the proposed approach achieves significantly better results than the baseline methods.

Feature visualization and analysis. Understanding and interpreting the predictions made by a deep learning model provides valuabe insights into the input data and the features learned by the model so that the results can be easily understood by human experts. In order to visually explain the decisions made by the proposed classifier and baseline methods, we use gradient-weighted class activation map (Grad-CAM) [86] to generate the saliency maps that highlight the most influential features affecting the predictions. Since convolutional feature maps retain spatial information and each pixel of the feature map indicates whether the corresponding visual pattern exists in its receptive field, the output from the last convolutional layer of the VGG-16 network shows the discriminative region of the image.

The class activation maps displayed in Figure 2.8 show that even though the baseline methods demonstrate high activations for the region consisting of the lesion, they still fail to correctly classify the dermoscopic image. For our proposed MelaNet approach, we observe that the area surrounding the skin lesion is highly activated. Notice that most of the borders of the whole input image are highlighted, due largely to the fact the classifiers are not looking at the regions of interest, and hence result in misclassification.

We can also see in Figure 2.9 that while the proposed approach shows similar visual patterns as the baselines when correctly classifying the input image, it, however, outputs high activations for the regions surrounding the skin lesion in many cases. These regions consist of shapes and edges. Hence, our approach not only focus on the skin lesion, but also captures its context, which helps in the final detection. This context-based approach is commonly used by expert dermatologists [83]. This observation is of great significance, and further shows the effectiveness of our approach.

In order to get a clear understanding of the data distribution, the learned features from both the original training set and the balanced dataset (i.e. with the additional synthesized data using adversarial training) are visualized using Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) [87], which is a dimensionality reduction technique that is particularly well-suited for embedding high-dimensional data into a two- or three-dimensional space. The UMAP embeddings shown in Figure 2.10 were generated by running the UMAP algorithm on the original training set with 900 samples (benign and malignant) and the balanced dataset consisting of 1627 samples (benign and malignant).

From Figure 2.10 (left), it is evident that the inter-class variation is significantly small due in large part to the very high visual similarity between malignant and benign skin lesions. Hence, the task of learning a decision boundary between the two categories is challenging. We can also see that the synthesized samples (malignant lesions shown in green) lie very close to the original data distribution. It is important to note that the outliers present in the dataset are not due to

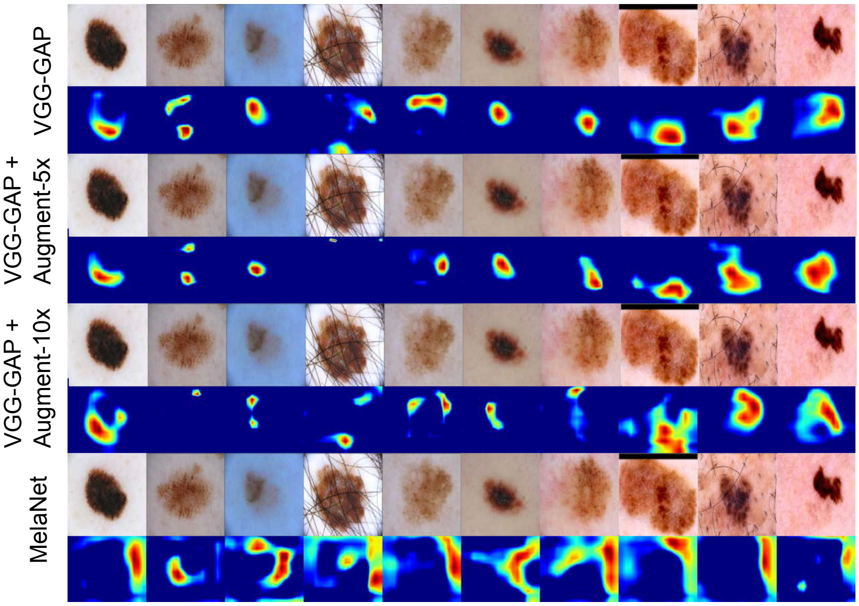


Figure 2.8: Grad-CAM heat maps for the misclassified malignant cases by MelaNet and baseline methods.

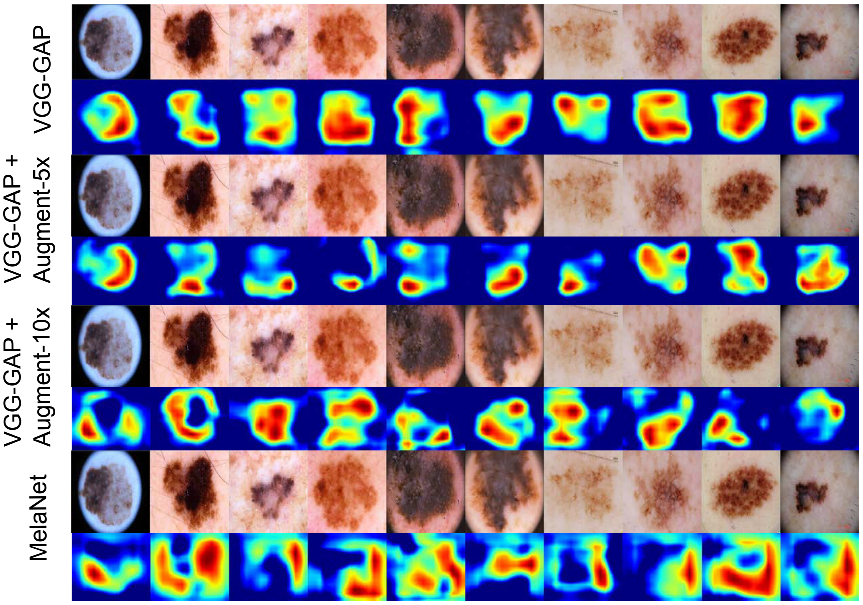


Figure 2.9: Grad-CAM heat maps for the correctly classified malignant cases by MelaNet and baseline methods.

the image synthesis procedure, but this is rather a characteristic present in the original training set. Therefore, the synthetically generated data are representative of the original under-represented class (i.e. malignant skin lesions).

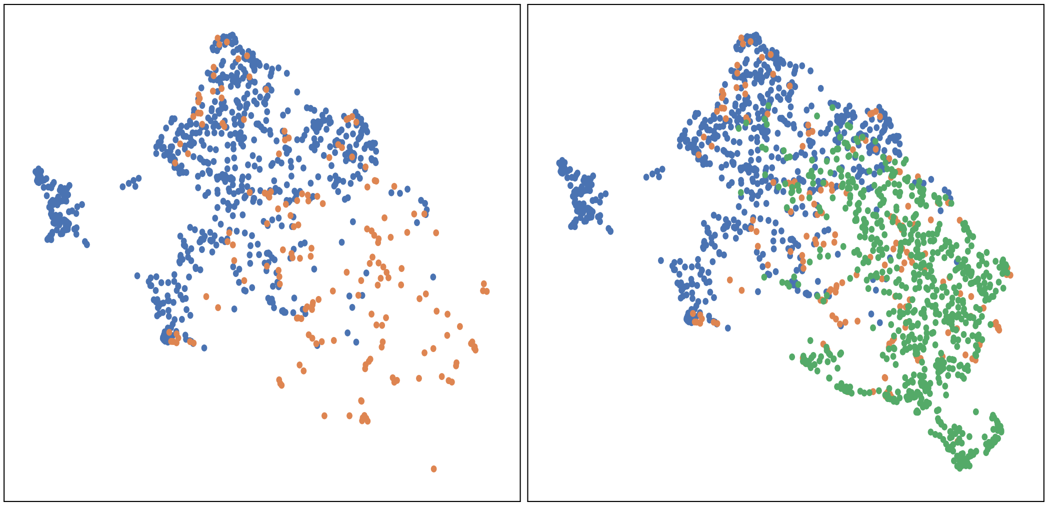


Figure 2.10: Two-dimensional UMAP embeddings using the original ISIC-2016 training set (left) consisting of 900 samples (benign shown in blue and malignant in orange) and with additional synthesized malignant data samples (shown in green) consisting of a total 1627 samples (right).

2.3.6 COVID-19 Detection Results

In this section, we conduct extensive experiments to evaluate the performance of the proposed data generation framework for COVID-19 detection from chest X-ray images.

Over-sampling with synthetic data. We demonstrate the effectiveness of the synthetic sets $\mathcal{G}_{NC}$ and $\mathcal{G}_{PC}$ in Tables 2.2 and 2.3 using four deep learning models, namely VGG-16 [74], ResNet-50 [78], DenseNet-102 [79], and DenseNet-121 with a bagging tree classifier (DenseNet121 + BGT) [62]. For each task, we can observe that when $\mathcal{G}_{NC}$ is added, there is a significant increase in performance. While the addition of $\mathcal{G}_{PC}$ also results in an increase in performance, such an increase is not quite large compared to adding $\mathcal{G}_{NC}$ in some cases. We hypothesize that this is due to the number of COVID-19 examples in $\mathcal{G}_{NC}$ (16,537), which enables the models to learn better representations for COVID-19, whereas $\mathcal{G}_{PC}$ is comprised of only 4,758 COVID-19 examples. Further, an increase in performance using both metrics is observed when the skew in the training dataset decreases. The relative improvement seems to drop as the model complexity increases, which is in line with the findings in [88] due to the problem of over-parametrization. When synthetic data are used as additional training set, the detection performance significantly increases. However, the relative improvement drops when the architectural complexity of the model increases. Note that despite its simplicity, the VGG-16 network outperforms all the other baseline methods, while the "DenseNet121 + BGT" model yields the second best performance. For less complex models, we can see that using only synthetic dataset performs better than the original data. Moreover, Table 2.3 shows that with the exception of VGG-16, all models achieve sub-optimal performance when using synthetic data only.

Table 2.2: COVID-19 detection performance results on Normal vs. COVID-19 test set when trained on real data; real + synthetic data; and only synthetic data (i.e. only $\mathcal{G}_{NC}$ is used for positive class examples in training each model). SEN is short for Sensitivity. Boldface numbers indicate the best performance.

	Real			Real + $\mathcal{G}_{NC}$			Real + $\mathcal{G}_{NC}$ + $\mathcal{G}_{PC}$			Only Synthetic	
Model	SEN (%) ↑	FN ↓	Skew↓	SEN (%) ↑	FN ↓	Skew↓	SEN (%) ↑	FN ↓	Skew↓	SEN (%) ↑	FN ↓
VGG-16	19.56	37	91.87	54.34	21	0.98	**63.04**	**17**	0.79	50.00	23
ResNet-50	32.61	31	91.87	41.30	27	0.98	**43.47**	**26**	0.79	10.86	41
DenseNet-102	26.08	34	91.87	28.27	33	0.98	**34.73**	**30**	0.79	8.69	42
DenseNet-121 + BGT	36.95	29	91.87	45.65	25	0.98	**52.17**	**22**	0.79	21.73	36

Table 2.3: COVID-19 detection performance results on Pneumonia vs. COVID-19 test set when trained on real data; real + synthetic data; and only synthetic data (i.e. only $\mathcal{G}_{PC}$ is used for positive class examples in training each model). Boldface numbers indicate the best performance.

	Real			Real + $\mathcal{G}_{PC}$			Real + $\mathcal{G}_{PC}$ + $\mathcal{G}_{NC}$			Only Synthetic	
Model	SEN (%) ↑	FN ↓	Skew↓	SEN (%) ↑	FN ↓	Skew↓	SEN(%) ↑	FN ↓	Skew↓	SEN (%) ↑	FN ↓
VGG-16	8.69	42	22.9	29.50	24	0.95	**52.17**	**22**	0.19	39.13	28
ResNet-50	21.73	36	22.9	36.95	29	0.95	**41.30**	**27**	0.19	13.04	40
DenseNet-102	4.34	44	22.9	21.74	36	0.95	**32.43**	**32**	0.19	6.52	43
DenseNet-121 + BGT	32.60	31	22.9	41.30	27	0.95	**47.82**	**24**	0.19	32.60	31

Training on anonymized synthetic data. We also evaluate the performance when the classifiers are trained on only synthetic COVID-19 images, as shown in Tables 2.2 and 2.3 for each dataset. Sub-optimal performance is achieved for both tasks for different CNNs, except for VGG-16 which shows performance improvement compared to when using the original COVID-19 examples. Since a new data sample is not attributed to an individual patient, but it is rather an instance which is conditioned on the training data, it does not entirely reflect the original data. This suggests that synthetic data alone cannot to used to achieve optimal performance. In other words, the synthetic data can be used as a form of pre-training, which often requires a small amount of real data to achieve comparable performance. In addition, the relatively large margin between the evaluation scores suggests that the observed difference between the models is actually real, and not due to a statistical chance.

Detecting target class with high confidence. The output of the softmax function describes the probability (or confidence) of the learning model that a particular sample belongs to a certain class. The softmax layer takes the raw values (logits) of the last FC layer and maps them into probability scores by taking the exponents of each output and then normalize each number by the sum of those exponents so that all probabilities sum to one. Figure 2.11 shows the probability scores for the VGG-16 model on unseen test set of COVID-19 for the two binary classification tasks of Normal vs. COVID-19 and Pneumonia vs. COVID-19 using original data only. The red dashed line depicts the 0.5 probability threshold. Notice that Figure 2.11(left) shows low confidence scores, while Figure 2.11(right) shows sub-optimal performance for COVID-19 detection when using original training data only.

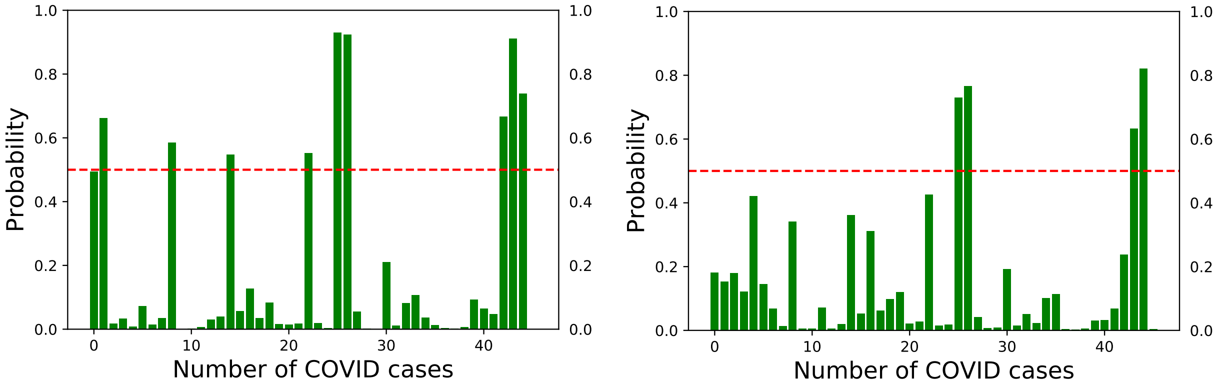


Figure 2.11: Confidence scores for the VGG-16 model on unseen test set of COVID-19 for the two binary classification tasks of Normal vs. COVID-19 (left) and Pneumonia vs. COVID-19 (right) using original data only.

Figure 2.12 shows that synthetic data can be used without the original examples. When using synthetic data as additional training set, we observe that not only the number of correctly detected instances of COVID-19 increases, but also the predictions tend to improve, as demonstrated by the high probability scores.

Figures 2.13 and 2.14 show improved detection performance when the synthetic data are used as additional training set. A similar trend was observed with the ResNet-50 and DenseNet-102 models.

Generating anonymized synthetic images with variation. Data visualization based on dimension reduction plays an important role in data analysis and interpretation. The objective of dimension reduction is to map high-dimensional data into a low-dimensional space (usually 2D or 3D), while preserving the overall structure of the data as much as possible. A commonly used dimension reduction method is the Uniform Manifold Approximation and Projection (UMAP) algorithm, which is non-linear technique based on manifold learning and topological data analysis. UMAP is capable of preserving both local and most of the global structure of the data when an appropriate initialization of the embedding is used. The two-dimensional UMAP embeddings of the features

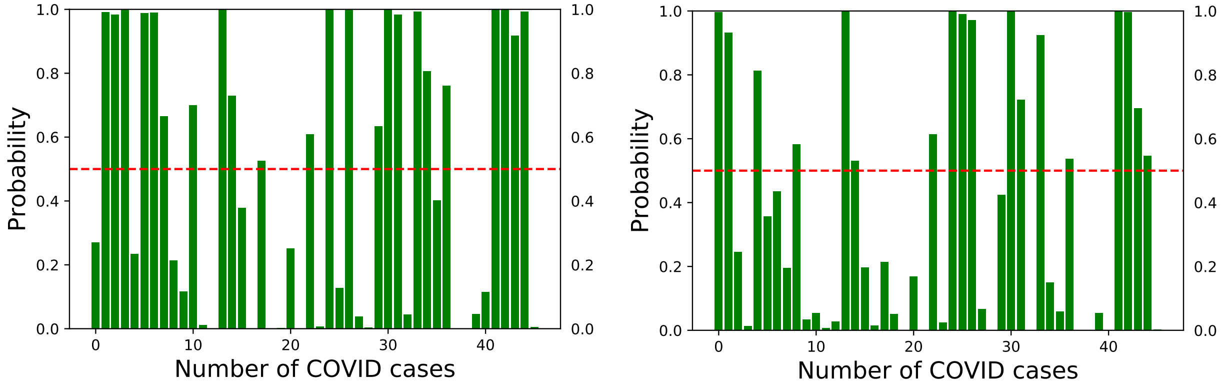


Figure 2.12: Confidence scores for the VGG-16 model on unseen test set of COVID-19 for the two binary classification tasks of Normal vs. COVID-19 (left) and Pneumonia vs. COVID-19 (right) using synthetic data without the original examples. Notice that synthetic data increase the confidence scores.

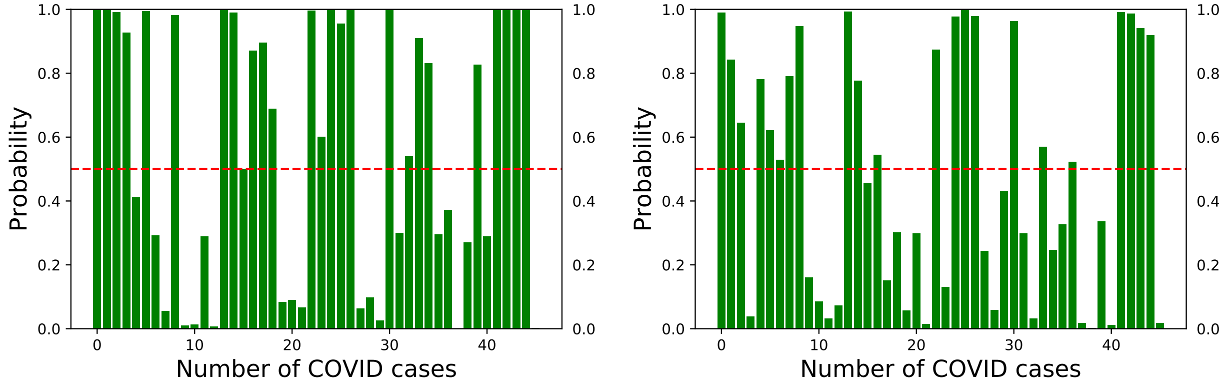


Figure 2.13: Confidence scores for the VGG-16 model on unseen test set of COVID-19 for the two binary classification tasks of Normal vs. COVID-19 (left) and Pneumonia vs. COVID-19 (right) with synthetic data as additional training set. Left: adding 16,537 COVID-19 examples of $\mathcal{G}_{NC}$ to the original COVID-19 dataset. Right: adding 4,758 COVID-19 examples of $\mathcal{G}_{PC}$ to the original COVID-19 dataset.

are shown in Figure 2.15 to visualize the difference between the original and synthetic data. Notice that the synthetic samples are in a different distribution in the feature space, enabling a decision boundary between the classes. The original examples in Figure 2.15(a) exhibit low inter-class variation and consist of outliers. In Figure 2.15(b), we can see that the synthetic examples of the $\mathcal{G}_{NC}$ dataset are in a different distribution in the feature space. While the UMAP embeddings may not be interpreted as a justification that the synthetic examples actually consist of COVID-19 symptoms from a clinical perspective, it is, however, important to note that the distribution of the synthetic images is significantly different than that of normal images; thereby enabling a proper decision boundary. A similar trend can be observed in Figures 2.15(d), (e) and (f). The overlapping features

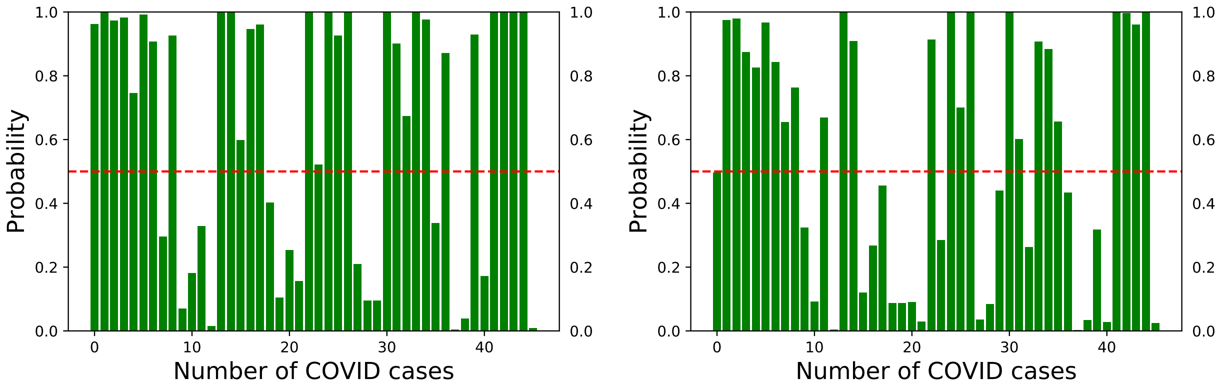
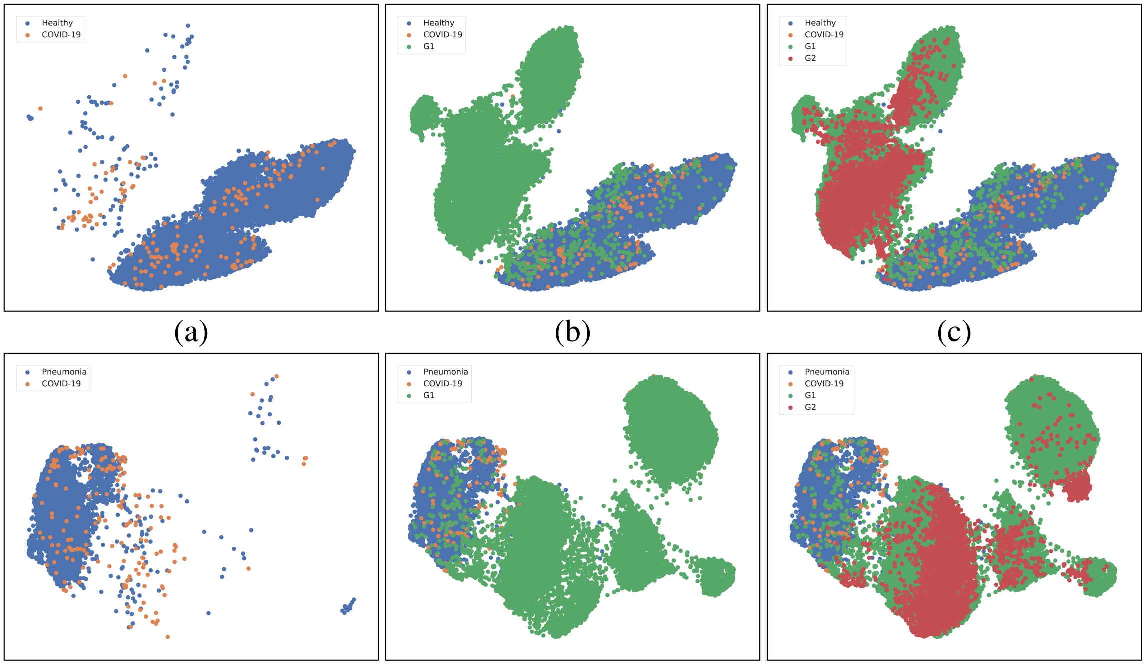


Figure 2.14: Confidence scores for the VGG-16 model on unseen test set of COVID-19 for the two binary classification tasks of Normal vs. COVID-19 (left) and Pneumonia vs. COVID-19 (right) with synthetic data as additional training set. Both $\mathcal{G}_{NC}$ and $\mathcal{G}_{PC}$ are added to the original COVID-19 dataset.

for Pneumonia vs. COVID-19 can be explained by the fact that the findings of X-ray imaging in COVID-19 are not specific, and tend to overlap with other infections such as Pneumonia in this case.

Discussion. With a training set consisting of only 1627 images, our proposed MelaNet approach is able to achieve improved performance. This better performance is largely due to the fact that by leveraging the inter-class variation in medical images, the mapping between the source and target distribution for conditional image synthesis can be easily learned. Moreover, it is much easier to generate target images given prior information, rather than generating from noise which often results in training instability and artifacts [82]. It is important to note that even though image-to-image translation schemes are considered to hallucinate images by adding or removing image features [32], we showed that in our scheme the partition of the inter-classes does not result in a bias or unwanted feature hallucination. Figure 2.16 shows the benign lesions sampled from the ISIC-2016 training set, which are translated to malignant samples using MelaNet. As can be seen, the benign and the corresponding synthesized malignant images have a high degree of visual similarity. This is largely due to the nature of the dataset, which is known to have a low inter-class variation.

In the synthetic minority over-sampling technique (SMOTE), when drawing random observations from its k-nearest neighbors, it is possible that a "border point" or an observation very close to the decision boundary may be selected, resulting in synthetically-generated observations lying too close to the decision boundary, and as a result the performance of the classifier may be degraded. The advantage of our approach over SMOTE is that we learn a transformation between



(a) (b) (c)

(d) (e) (f)

Figure 2.15: Two-dimensional UMAP embeddings: (a) Normal vs. COVID-19; (b) Normal vs. COVID-19 + $\mathcal{G}_{NC}$; (c) Normal vs. COVID-19 + $\mathcal{G}_{NC}$ + $\mathcal{G}_{PC}$; (d) Pneumonia vs. COVID-19; (e) Pneumonia vs COVID-19 + $\mathcal{G}_{NC}$; (f) Pneumonia vs. COVID-19 + $\mathcal{G}_{NC}$ + $\mathcal{G}_{PC}$. Here, G1 and G2 denote $\mathcal{G}_{NC}$ and $\mathcal{G}_{PC}$, respectively

a source and a target domain by solving an optimization problem in order to determine two bijective mappings. This enables the generator to synthesize observations, which help improve the classification performance while learning the transformation/decision boundary.

In order to gain a deeper insight on the performance of the proposed approach, we sample all the original benign lesions and a subset of the synthesized malignant lesions, consisting of 727 and 10 samples, respectively. For the benign group of images, the proposed MelaNet model yields a sensitivity score of 89%, with 77 misclassified images. By contrast, a 100% sensitivity score is obtained when performing predictions on the synthesized malignant group of images. In addition, the F-score values for MelaNet on the benign and synthesized malignant groups are 94% and 21%, respectively.

For COVID-19 detection, since the generative and classification models are trained to learn representations in the training data distribution, it is likely that a bias might occur toward that data. In light of the class imbalance problem, the generator is trained by under-sampling the majority class. This under-sampling process often leaves a relatively small number of data points (180 samples for each domain) to learn from. While a boost in performance is achieved when using the synthetic datasets, it is not conclusive enough to confirm whether our approach can be

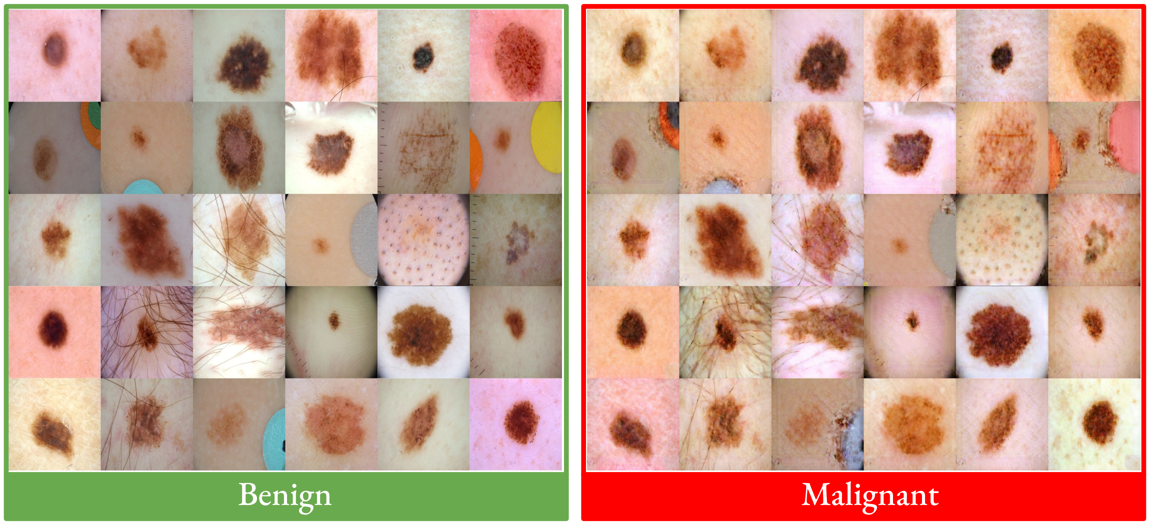


Figure 2.16: Sample benign images from the ISIC-2016 dataset that are translated to malignant images using the proposed approach. Notice that the synthesized images display a reasonably good visual quality.

generalized across other COVID-19 datasets due largely to the lack of such benchmarks. While the improvements we have achieved using our proposed framework are encouraging, it is important to mention that a key objective of this work is not to claim state-of-the-art results, but rather to release an open source dataset to the research community in an effort to further improve COVID-19 detection.

CHAPTER 3

Depthwise ConvNet for Medical Image Segmentation

In this chapter, we introduce a simple, yet effective end-to-end depthwise encoder-decoder fully convolutional network architecture, called Sharp U-Net, for binary and multi-class biomedical image segmentation. The key rationale of Sharp U-Net is that instead of applying a plain skip connection, a depthwise convolution of the encoder feature map with a sharpening kernel filter is employed prior to merging the encoder and decoder features, thereby producing a sharpened intermediate feature map. Using this sharpening filter layer, we are able to not only fuse semantically less dissimilar features, but also to smooth out artifacts throughout the network layers during the early stages of training. Extensive experiments on six datasets show that the proposed Sharp U-Net model consistently outperforms or matches the recent state-of-the-art baselines in both binary and multi-class segmentation tasks, while adding no extra learnable parameters. Furthermore, Sharp U-Net outperforms baselines that have more than three times the number of learnable parameters.

3.1 Introduction

Semantic segmentation is a fundamental task in biomedical imaging [89–91], with numerous clinical applications including the detection of the novel coronavirus disease 2019 (COVID-19) in computed tomography (CT) images. It refers to the process of classifying each pixel in a biomedical image into one of the pre-defined semantic categories or classes. The goal is to semantically understand the role of each pixel in the image in an effort to distinguish between regions of interests (ROIs) in the image such as tumors or organs [76, 92].

The task of semantic segmentation is tantamount to performing classification at a pixel level, allowing the homogeneous pixels to be clustered together. This motivates the development of automated computer-aided diagnosis systems, which enable physicians to analyze specific image regions, especially in multimodal medical images [93].

Deep neural networks, and in particular convolutional neural networks (CNNs), have been successfully applied to image segmentation, showing promising results in comparison with shallow methods [94, 95]. Cireşan *et al.* [94] segment biological neuron membranes using a deep neural network as a pixel classifier, where the label of each pixel is predicted from raw pixel values in a sliding window centered around it. Yuan *et al.* [95] present a fully convolutional network for skin lesion segmentation by leveraging a deep convolutional neural network trained end-to-end using Jaccard distance as a loss function. Owing to the recent developments in fully convolutional networks [96], there has been a surge of interest in the adoption of encoder-decoder networks, particularly U-Net, for biomedical image segmentation [5, 40, 42, 97–99]. The U-Net model, which is trained in an end-to-end fashion for pixel-wise prediction, has emerged as a very powerful encoder-decoder architecture due largely to its state-of-the-art performance in segmenting biomedical images even when the labelled training data is small. The U-Net architecture is composed of an encoder subnetwork for capturing context by encoding the input image into low-level feature representations at multiple levels and a decoder subnetwork for semantically projecting these feature representations into the pixel space in an effort to enable precise localization via transposed convolutions. While U-Net is built upon the fully convolutional network, it differs from the latter in the sense that it is a symmetric architecture and uses skip connections between the encoder and decoder subnetworks in order to merge low- and high-level features with the goal of preserving more refined image details.

As a core component of encoder-decoder networks, skip connections combine the deep, semantic and course-grained features from the decoder with the shallow, low-level and fine-grained features from the encoder. This feature fusion process has proven to be effective in recovering details of ROIs [100] and in producing accurate segmentation maps, even on a complex image background [101]. While skip connections work relatively well in dense prediction tasks such as image segmentation, the encoder and decoder feature combinations may, however, not match. This mismatch problem is largely attributed to the fact that the encoder features are low-level and fine-grained since they are computed from the early layers of the network, whereas the decoder features are high-level, semantic and course-grained as they are computed from much deeper layers in the network. Consequently, the feature mismatch between the encoder and decoder subnetworks is likely to occur, leading to fusing semantically dissimilar features and hence resulting in blurred feature maps throughout the learning process and also adversely affecting the output segmentation map by under- and/or over-segmenting ROIs.

Prior to fusing with the decoder features, the encoder features undergo a depthwise convolution (i.e. spatial convolution operation performed independently over each channel of the encoder features) by using a sharpening spatial kernel, with the aim to reduce feature mismatch.

To address these limitations, we propose Sharp U-Net, a novel end-to-end depthwise convolutional network architecture for biomedical image segmentation. The key idea behind the proposed framework is to emphasize the fine details of the early level features generated by the encoder via depthwise convolution of the encoder feature map (i.e. spatial convolution operation performed independently over each channel of the encoder features) with sharpening spatial filter prior to performing fusion with the decoder features. Sharpening the features from the encoder subnetwork not only enables better feature fusion, but also help the network progressively learn better representations of the data. Moreover, a sharpening spatial filter helps smooth out artifacts throughout the network during the early stages of training due to untrained parameters. In this chapter, we demonstrate that Sharp U-Net yields significantly improved performance over the vanilla U-Net model for both binary and multi-class segmentation of medical images from different modalities, including electron microscopy (EM), endoscopy, dermoscopy, nuclei, and computed tomography (CT). This better performance is achieved without additional learnable parameters in comparison with recent architectures that have more than three times the number of learnable parameters. Finally, the key idea proposed in this work can naturally be used to generalize many other encoder-decoder architectures by incorporating sharpening filters in a similar fashion. To summarize, the main contributions in this chapter are as follows:

- We introduce a novel Sharp U-Net architecture by designing new connections between the encoder and decoder subnetworks using a depthwise convolution of the encoder feature maps with a sharpening spatial filter to address the semantic gap issue between the encoder and decoder features.
- We show that the Sharp U-Net architecture can be scaled for improved performance, outperforming baselines that have three times the number of learnable parameters.
- We demonstrate through extensive experiments the ability of the proposed model to learn efficient representations for both binary and multi-class segmentation tasks on a variety of medical images from different modalities.

The remainder of this chapter is structured as follows. In Section 3.2, we outline the motivation behind the proposed framework and present the problem formulation. Then, we introduce an end-to-end depthwise encoder-decoder convolutional network architecture for both binary and multi-class biomedical image segmentation. In Section 3.3, we present experimental results to demonstrate the competitive performance of our approach on six standard benchmarks.

3.2 Method

In this section, we start with the motivation behind introducing a depthwise convolution for feature maps sharpening, followed by the problem formulation. Then, we present the main building blocks of our proposed network architecture for binary and multi-class segmentation.

3.2.1 Motivation and Problem Statement

In order to alleviate the fusion of dissimilar features between the encoder and decoder sub-networks of the U-Net model, we extend the U-Net architecture by introducing a depthwise convolution for sharpening the encoder features prior to fusing them with the decoder features. This sharpening operation not only helps balance the semantic gap introduced by the high-level process in the decoder sub-network, but also sharpens the details in the feature maps. In addition, sharpening helps improve the training process in the early stages by reducing low-frequency noise propagated by the untrained parameters in the feature space. It is important to point out that our Sharp U-Net architecture does not introduce any additional learnable parameters.

Problem Formulation. Semantic segmentation refers to the process of classifying each pixel in an image into its semantic class and hence can be regarded as a classification problem at the pixel level. A lung segmentation task, for example, can be thought of as a binary segmentation problem with two semantic classes: lung and background. However, unlike image classification whose goal is to assign an input image to one label from a fixed set of categories or classes, the output in semantic segmentation is an image, typically of the same size as the input image, such that each pixel is classified to a particular class.

For a multi-class segmentation problem consisting of C classes, we denote by $\mathcal{X} = \{(\mathbf{x}_i, y_i) : i = 1, \dots, N\}$ a set of N samples, where $\mathbf{x}_i$ is the ith training sample and $y_i \in \{1, \dots, C\}$ is the corresponding true label. The true label of the ith sample can be represented as a one-hot encoding vector $\mathbf{y}_i = (y_{i1}, \dots, y_{iC})^\intercal$, such that $y_{ic} = 1$ if $i = c$ and 0 otherwise for each class c.

3.2.2 Proposed Neural Architecture

Skip connections in encoder-decoder networks, particularly in U-Net, play a crucial role in recovering fine-grained details in the prediction. However, these skip connections tend to fuse low- and high-level convolutional features that are semantically different, and hence resulting in blurred feature maps. In order to address these issues, we introduce a Sharp U-Net architecture for both binary and multi-class segmentation, as shown in Figure 3.1. The key rationale of our Sharp U-Net framework is to mitigate the semantic gap between the encoder and decoder features, and in the meanwhile, to smooth out artifacts throughout the network layers during the early stages of training. Similar to U-Net, the proposed Sharp U-Net architecture consists of a contracting or downsampling path (encoder) for capturing context using convolutions and an expanding or upsampling path (decoder) for enabling precise localization using transposed convolutions (also known as up-convolutions).

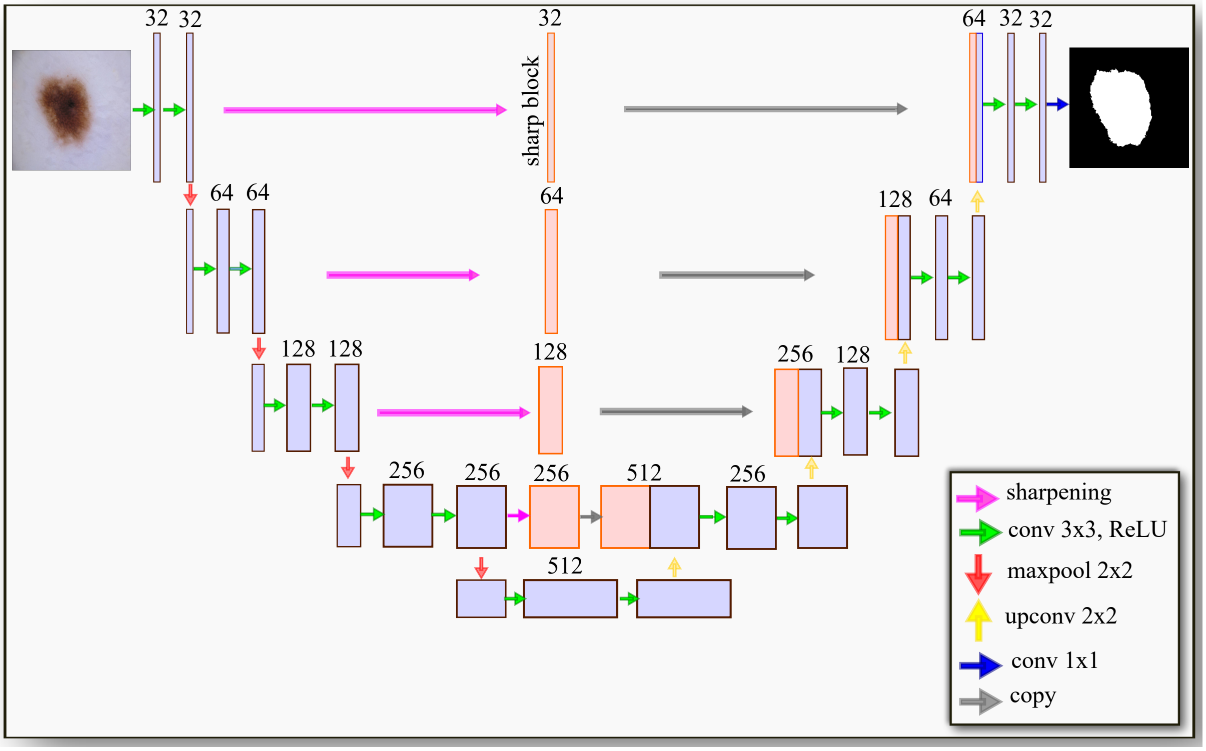


Figure 3.1: Schematic layout of the proposed Sharp U-Net architecture. Prior to fusing with the decoder features, the encoder features undergo a depthwise convolution (i.e. spatial convolution operation performed independently over each channel of the encoder features) by using a sharpening spatial kernel, with the aim to reduce feature mismatch. These additional operations do not increase the number of learnable parameters, and hence no additional computational cost is incurred.

The encoder, which is composed of five blocks, down-samples the input feature maps to extract low-level features. Each block of the encoder consists of two 3 × 3 convolutional layers with rectified linear unit (ReLU) activations, followed by a 2 × 2 max-pooling layer, except the fifth block, which does not include a pooling layer. We use 32, 64, 128, 256 and 512 filters for the convolutional layers in the encoder blocks, i.e. the number of features maps is doubled after each block.

The convolutional layer applies a filter to an input to create a feature map that summarizes the presence of extracted features from the input, whereas the pooling layer downsamples or reduces the size of each feature map by a factor of 2.

On the other hand, the decoder, which also consists of five blocks, may be regarded as an operator that performs the reverse of the downsampling path. Every block in the expanding path comprises a 2×2 up-convolution (i.e. upsampling the features), followed by two 3×3 convolutions with ReLU activations, except the fifth block, which has an additional 1×1 convolution. We use 256, 128, 64 and 32 filters for the convolutional layers in the decoder blocks, i.e. the number of features maps is halved after each block. The output of the Sharp U-Net model is a pixel-by-pixel mask that shows the class of each pixel.

In order to localize the upsampled features, we design a new connection mechanism, called sharp block, between the encoder and decoder subnetworks in an effort to fuse the low- and high-level features from the encoder and decoder, respectively, while mitigating the semantic gap problem. To this end, instead of using plain skip connections between the encoder and decoder, the encoder features undergo a spatial convolution operation conducted independently on each channel of the encoder features using a sharpening spatial kernel prior to performing fusion with the decoder features. This sharpening filter layer not only fuses semantically less dissimilar features, but also helps reduce the high-frequency noise components that are propagated throughout the network layers during the early stages of training as the channels are Gaussian smoothed before applying the Laplacian filter. Moreover, scaling strategies can be leveraged to further improve the performance of the proposed Sharp U-Net framework by uniformly increasing the learnable parameters and/or using pre-trained CNN models as encoders.

3.2.3 Sharpening Spatial Kernel

Spatial filtering is a low-level neighborhood-based image processing technique, which performs operations on the neighborhood of every pixel in an image for the sake of image enhancement such as sharpening. The objective of sharpening spatial filtering is to preserve or emphasize high-frequency components representing fine-grained image details by highlighting transitions in intensity within the image. The image sharpening process, also referred to as high-pass filtering, is usually performed via image convolution with kernels or masks. Convolution kernels, also known as filters, are discrete approximations of the image Laplacian, which is a second-order derivative operator that is capable of responding to intensity transitions in any direction. A commonly-used Laplacian high-pass filtering kernel for image sharpening, which takes into account all eight neighbors of the reference pixel in the input image, is given by

$$\mathbf{K} = \begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}. \tag{3.1}$$

Note that convolving an image with the Laplacian filter kernel $\mathbf{K}$ increases the brightness of the center pixel relative to neighboring pixels, as the 3×3 kernel matrix is comprised of a positive value in the center and negative off-center values. Also, the sum of all elements of the high-pass filter kernel is always zero.

In order to obtain a sharpened image, the input image is added to its convolution with the kernel. More precisely, let **I** be the input image, then the resulting sharpened image **S** is given by

$$\mathbf{S} = \mathbf{I} + \mathbf{K} * \mathbf{I}, \tag{3.2}$$

where $*$ denotes convolution, a neighborhood-based operator that processes an image by adding each value of a pixel to its local neighbors, weighted by the kernel.

3.2.4 Sharp Blocks

Since the encoder features are multi-dimensional, generally of size $W \times H \times M$ with W, H and M denoting the width, height and number of encoder feature maps respectively, we perform a depthwise convolution on each feature map using a sharpening spatial kernel given by the Laplacian filter kernel $\mathbf{K}$. Similar to a kernel in a convolutional layer, the depthwise convolution layer is parametrized by the sharpening spatial kernel $\mathbf{K}$. This can also be thought of as initializing the weights of the depthwise convolutional layer with the weights given by $\mathbf{K}$. Since we are interested in transforming the input encoder features, no additional bias terms are used for this initialization. Instead of using a single filter of a particular size (i.e. $3 \times 3 \times 3$) in convolutions, we use M filters, which act on each input channel separately. Each channel of the input is convolved with the kernel $\mathbf{K}$ separately with a stride of 1. Each of these convolutions yields a feature map of size $W \times H \times 1$. During the feature fusion of the encoder and decoder subnetworks, we also perform padding to keep the output dimension the same as that of the input so that it matches the size of the decoder features in all stages of the connection. We then stack these maps together to obtain the final output of size $W \times H \times M$ from the depthwise convolution layer. We refer to this proposed feature connection as a sharp block. A visualization of the flow of operations in the sharp block is depicted in Figure 3.2.

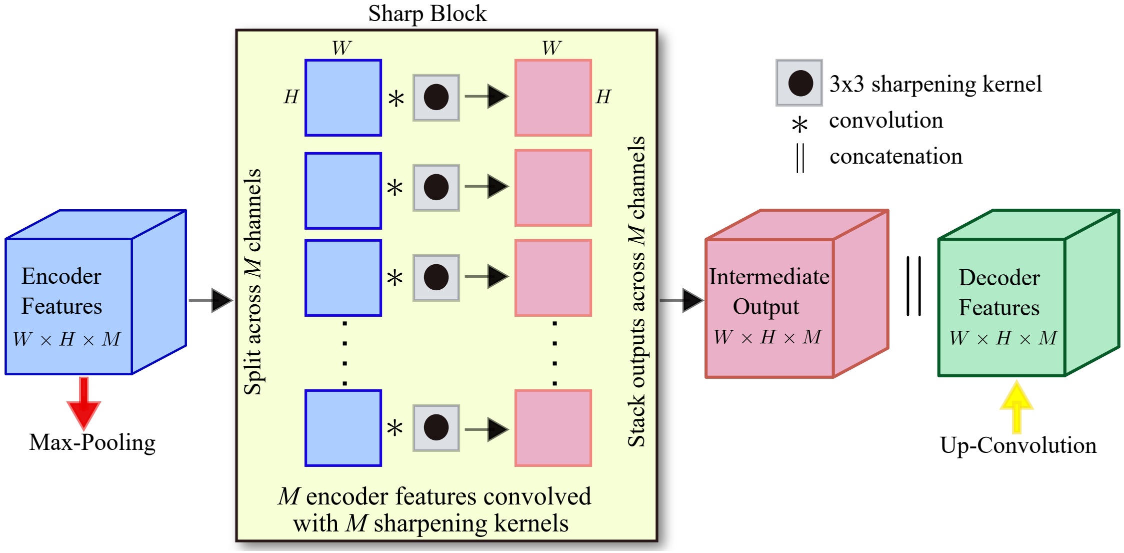


Figure 3.2: Illustration of the proposed Sharp Block. Given a multi-channel encoder feature map of size $W \times H \times M$ as input, the sharpening kernel layer performs a depthwise convolution, resulting in a sharpened intermediate feature map of the same size as the input.

3.2.5 Scaling Sharp U-Net

The proposed Sharp U-Net architecture can be scaled using recent scaling strategies such as uniformly increasing the number of learnable parameters [40], and/or using deep convolutional neural networks in the encoder subnetwork [42,78]. Uniformly increasing the number of learnable parameters is achieved by increasing the number of convolutional kernels while maintaining the kernel size. We increase the number of kernels of both the encoder and decoder subnetworks from 32, 64, 128, 256 and 512 to 35, 70, 140, 280 and 560, respectively. We refer to the resulting network architecture as Wide Sharp U-Net. We also scale Sharp U-Net by replacing the encoder with a pretrained convolutional neural network (e.g., VGG [42] or ResNet [78]) on the 1000-class Imagenet dataset, as the encoder plays the role of a feature extractor. In the different configurations of the scaled Sharp U-Net, an increase in the number of learnable parameters comes from the scaling strategies.

3.2.6 Model Training

Similar to the U-Net architecture, the proposed Sharp U-Net for multi-class segmentation is trained in an end-to-end fashion. The parameters (i.e. weights and biases for different layers) of the Sharp U-Net model are learned by minimizing the following categorical cross-entropy loss for all N

samples

$$\mathcal{L} = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic}\log\hat{y}_{ic}, \tag{3.3}$$

where C is the total number of classes, y_{ic} is the indicator that the ith sample belongs to the cth class, and $\hat{y}_{ic}$ is the predicted probability that the model associates the ith input with class c. During training, the network parameters are updated using the Adam optimizer [80]. Training is continued on all network layers until the validation loss stops improving, and then the best weights are retained using early an stopping mechanism. It is worth mentioning that other loss functions such as the focal Tversky loss [102] and distance-based losses [103, 104] can also be used for training.

3.3 Experiments

In this section, we conduct extensive experiments to assess the performance of the proposed image segmentation framework in comparison with strong baseline methods on several benchmark datasets. The source code to reproduce the experimental results is made publicly available on GitHub[1].

3.3.1 Datasets

We demonstrate and analyze the performance of the proposed image segmentation model on four datasets: Lung Segmentation[2], Data Science Bowl 2018 segmentation challence[3], ISIC-2018, COVID-19 CT Segmentation[4], ISBI-2012 [105], and CVC-ClinicDB [106]. We use the COVID-19 CT Segmentation dataset for multi-class image segmentation and the remaining datasets for binary image segmentation. The summary descriptions of these benchmark datasets are as follows:

- **Lung Segmentation** is a collection of 2D and 3D CT images with manually segmented lungs. The image size is 512×512, with variable depth. In our experiments, we use the slices of the CT images and their corresponding masks. For data preprocessing, we follow the same procedure as [107].

- **Data Science Bowl 2018** is a dataset comprised of 670 segmented nuclei images from different modalities, brightfield and fluorescence in particular. This dataset also consists of

[1] https://github.com/hasibzunair/sharp-unets

[2] https://www.kaggle.com/kmader/finding-lungs-in-ct-data

[3] https://www.kaggle.com/c/data-science-bowl-2018

[4] https://medicalsegmentation.com/covid19

instance-level annotations, where each cell is marked with a unique color or label. These 670 images are of various resolution, and we resize them to 256×256 while maintaining the aspect ratio.

- **ISIC-2018** is a dataset composed of skin lesions with a total of 2594 images with both segmentation masks expert labels from the HAM10000 dataset [108] and ISIC-2017 dataset [76]. The dataset consists of images of various resolutions, which we resize them to 256×192 while retaining aspect ratio.
- **COVID-19 CT Segmentation** is a multi-class segmentation dataset, which consists of 100 axial CT images from about 40 patients with COVID-19. These images were segmented by a radiologist using three labels: ground-glass, consolidation and and pleural effusion. All images are of size 512×512.
- **ISBI-2012** is an electron microscopy image modality dataset, which comprises 30 images from a serial section Transmission electron Microscopy (ssTEM) of Drosophila first instar larva ventral nerve cord XX. The images consist of alignment errors and also have noisy examples. We resize these images to 256×256.
- **CVC-ClinicDB** is a colonoscopy image dataset, which consists of endoscopy images. These images were extracted from video sequences of colonoscopy. Since only images of polyp are considered for the task of segmentation, this results in a total of 612 images. We resize these images to 256×192 while preserving the aspect ratio while maintaining the aspect ratio.

3.3.2 Baselines

We evaluate the performance of the proposed method against various baselines, including U-Net [5], Wide U-Net [40], TernausNet-16 [42], and U-Net + ResNet-50 [78]. We further improve these networks using our proposed architecture to demonstrate the extensibility of our approach, while avoiding additional learnable parameters. For baselines, we mainly consider methods that are closely related to U-Net and/or the ones that are state-of-the-art biomedical image segmentation frameworks. A brief description of these strong baselines can be summarized as follows:

- **U-Net** is a fully convolutional network for binary and multi-class biomedical image segmentation. It consists of a contracting path to capture context and a symmetric expanding path that enables precise localization. The five convolutional layers in the contracting path consist of 32, 64, 128, 256 and 512 filters, while the convolutional layers in the expanding path are comprised of 256, 128, 64 and 32 filters. For an input image of size 192×256, the U-Net model has approximately 7.8 million learnable parameters.

- **Wide U-Net** is a scaled version of U-Net, and is obtained by increasing the number of filters in both contracting and expanding paths (i.e. encoder and decoder). The convolutional layers in the contracting path have 35, 70, 140, 280 and 560 filters. The same numbers of filters are used in the expanding path. The network has roughly 9.1 million learnable parameters.

- **TernausNet-16** is a variant of U-Net, in which the contracting path is replaced with a VGG-16 pretrained model on ImageNet. The VGG-16 model consists of 16 layers with learnable weights: 13 convolutional layers, and 3 fully connected layers. Each of the first and second convolutional blocks is comprised of two convolutional layers with 64 and 128 filters, respectively. Similarly, each of the third, fourth and fifth convolutional blocks consists of three convolutional layers with 256, 512, and 512 filters, respectively. The network has about 23.7 million learnable parameters.

- **U-Net + ResNet-50** is a segmentation network, where the encoder part of the U-Net is replaced with the ResNet-50 pretrained model on ImageNet. The network has approximately 32.5 million learnable parameters.

It is important to note that in all experiments, no data augmentation or post-processing such as conditional random fields (CRF) or median filting are employed, as our aim is to introduce a new architecture and demonstrate the performance improvements attributed to sharp blocks.

3.3.3 Implementation Details

All experiments are run on a Linux Workstation featuring an AMD Ryzen Threadripper 2950X processor with 16 cores and 64 processing threads, 4.4 GHz Max Boost, 64GB RAM, and an NVIDIA GeForce RTX 2080 Ti with 11 GB Memory. We use k-fold cross-validation with $k = 5$ to evaluate the segmentation results by different methods using two evaluation metrics. The cross-validation experiments are conducted 5 times with different random splits of the data. The average and standard deviation scores are reported. We use the Adam optimizer with learning rate 0.001.

3.3.4 Evaluation Metrics

In order to evaluate the performance of our proposed framework against the baseline methods, we use the Jaccard index and Dice coefficient as evaluation metrics. The Jaccard index, also known as Intersection-Over-Union (IoU), is one of the most commonly used metrics in semantic segmentation. Given two sets G and P, denoting the ground truth binary and predicted labels, respectively, the Jaccard similarity index is defined as

$$\mathcal{J}(G,P) = \frac{|G \cap P|}{|G \cup P|}, \tag{3.4}$$

where $|\cdot|$ denotes the cardinality of a set. Thus, the Jaccard index is the area of overlap between the predicted segmentation and the ground truth divided by the area of union between the predicted segmentation and the ground truth. For binary or multi-class segmentation, the mean IoU is computed by taking the IoUs of all classes and averaging them. The Jaccard index ranges from 0 to 1, with 1 indicating perfect match between the true and predicted labels, while 0 indicates a complete mismatch between them.

The Dice similarity coefficient is defined as the area of overlap between the predicted segmentation and the ground truth divided by the average of the areas of the predicted segmentation and the ground truth:

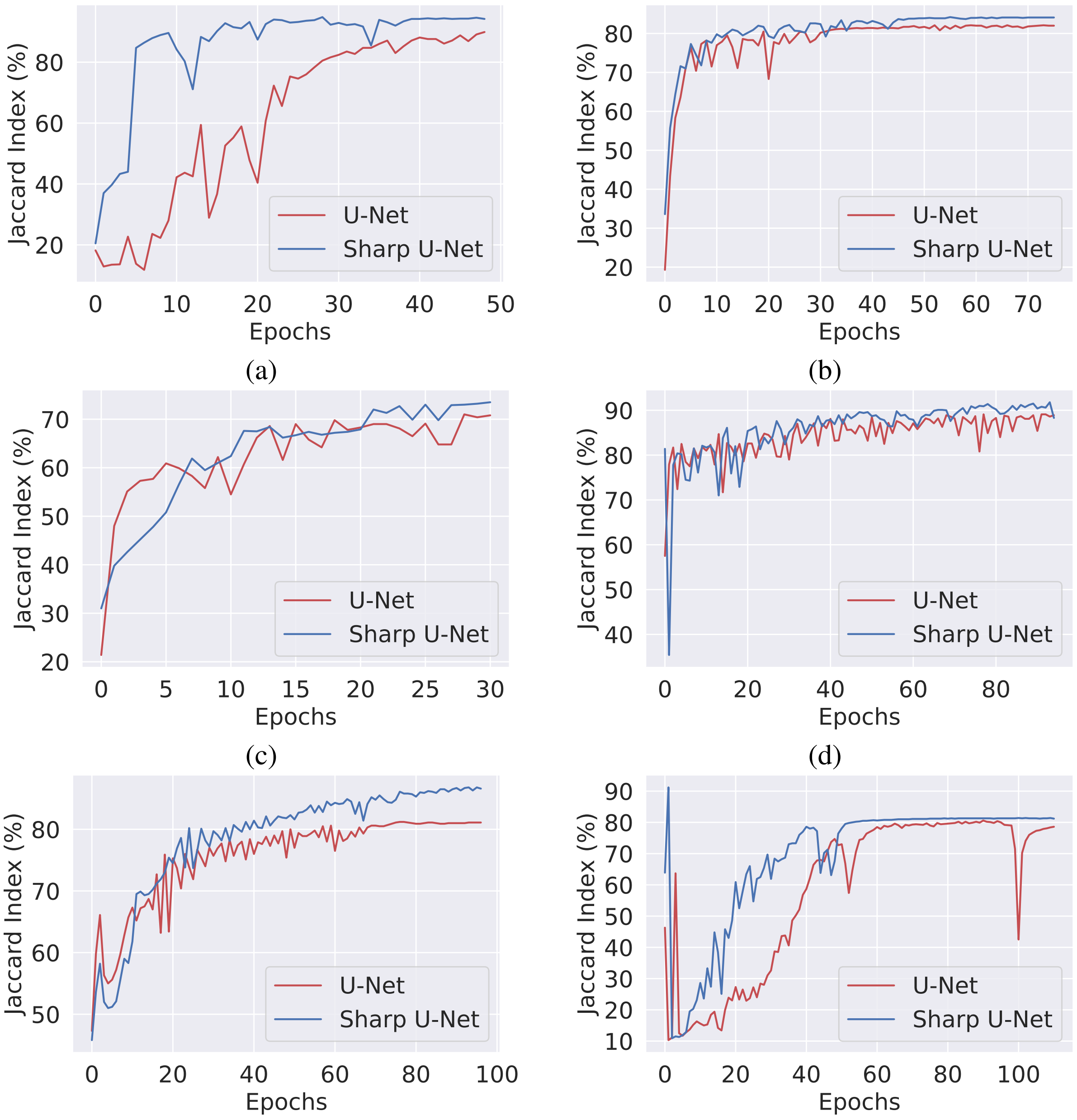
$$\mathcal{S}(G,P) = \frac{2|G \cap P|}{|G| + |P|}. \tag{3.5}$$

The Dice similarity coefficient also has values in the range [0, 1]. A similarity of 1 means that the segmentations are a perfect match.

3.3.5 Multimodal Medical Image Segmentation Results

We start by showing model validation history comparison between U-Net and the proposed Sharp U-Net model on all datasets. Then, we provide experimental results for both binary and multi-class segmentation on biomedical images from multiple modalities. We also show qualitative results of the predictions on some hard examples.

Model validation performance. The performance comparison between U-Net and Sharp U-Net over validation epochs on the validation set of each dataset is illustrated in Figure 3.3, which shows that the proposed Sharp U-Net model yields higher Jaccard index scores, indicating better performance. As can be seen, Sharp U-Net achieves better performance using 5-fold cross-validation on all cases for the same number of epochs. This better performance is largely attributed to the prevention of noise that is propagated during the early stages of training due to untrained parameters. In Figure 3.3(a), we can see that Sharp U-Net achieves better results much faster than U-Net. In Figure 3.3(f), both U-Net and Sharp U-Net display major fluctuations during training, but Sharp U-Net converges much faster, whereas U-Net lags behind in the early stages of the training process. These results demonstrate that Sharp U-Net yields superior results using the same number of epochs and even faster in some cases compared to the U-Net model.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 3.3: Progress of the best validation performance with the number of epochs from 5-fold cross-validation tests: (a) Lung Segmentation, (b) Data Science Bowl 2018, (c) ISIC-2018, (d) COVID-19 CT Segmentation, (e) ISBI-2012, and (f) CVC-ClinicDB. We record the value of the Jaccard index on the validation set after each epoch.

Sharp U-Net outperforms U-Net. We evaluate the performance of Sharp U-Net in comparison with the U-Net baseline across various datasets consisting of multiple modalities for both binary and multi-class segmentation tasks. For each task, we perform 5-fold cross-validation tests. The best results in each run are recorded and then averaged across all runs to get the scores of the

Jaccard Index and Dice for each model. We report the results in Table 3.1, which shows that the proposed Sharp U-Net model performs better than the U-Net architecture in segmenting medical images from different modalities. Significant performance improvements are observed in the case of endoscopy and EM images. A relative improvement of 12.6% and 3.63% on the Jaccard Index is observed on the CVC-ClinicDB and ISBI-2012 datasets. Note that this performance improvement is achieved with no additional learnable parameters. For dermoscopy images, Sharp U-Net achieves a relative improvement of 1.79% on the ISIC-2018 dataset. For CT and Nuclei images, Sharp U-Net matches the performance of U-Net, albeit Sharp U-Net performs slightly better with relative improvements of 0.48% and 0.63% on the Lung Segmentation and Data Science Bowl 2018 dataset, respectively. A similar pattern can be observed with the Dice score evaluations.

We also test our proposed Sharp U-Net model in multi-class segmentation, which is more challenging than binary segmentation, using the COVID-19 CT Segmentation dataset. As expected, Sharp U-Net outperforms U-Net by 2.52%. We visualize some multi-class segmentation results in Figure 3.4, which shows that Sharp U-Net is able to generate smoother predictions compared to U-Net.

Table 3.1: Evaluation results averaged over 5 folds for all datasets. We also report the standard deviation. Boldface numbers indicate the best segmentation performance. Notice that Sharp U-Net consistently outperforms U-Net.

	U-Net		Sharp U-Net	
Dataset	Jaccard (%)	Dice (%)	Jaccard (%)	Dice (%)
CVC-ClinicDB	70.02 ± 9.37	78.75 ± 8.10	$\mathbf{75.89} \pm 2.06$	$\mathbf{83.65} \pm 1.88$
ISBI-2012	83.95 ± 3.47	91.23 ± 2.11	$\mathbf{87.00} \pm 4.95$	$\mathbf{92.96} \pm 2.95$
COVID-19 CT Segmentation	85.24 ± 1.96	91.85 ± 1.17	$\mathbf{87.39} \pm 3.07$	$\mathbf{93.01} \pm 1.85$
ISIC-2018	74.74 ± 4.46	83.25 ± 3.73	$\mathbf{76.07} \pm 1.03$	$\mathbf{84.34} \pm 0.86$
Data Science Bowl 2018	84.95 ± 0.26	91.47 ± 0.18	$\mathbf{85.26} \pm 0.27$	$\mathbf{91.75} \pm 0.15$
Lung Segmentation	94.75 ± 0.58	97.03 ± 0.31	$\mathbf{95.22} \pm 0.67$	$\mathbf{97.25} \pm 0.36$

Scaling Sharp U-Net improves performance. In Figure 3.5, we illustrate the effectiveness of adding sharp blocks in comparison with adding traditional skip connections in Wide U-Net [40] on the CVC-ClinicDB dataset for the task of polyp segmentation. Endoscopy images consist of homogeneous ROIs and background which makes it very challenging to differentiate between the two. Both U-Net and Sharp U-Net yield the lowest Jaccard and Dice scores among all the other datasets, showing the difficulty in segmenting polyps. From Figure 3.5, it can be seen that when replacing the skip connections in Wide U-Net [40] with sharp blocks, a relative improvement on

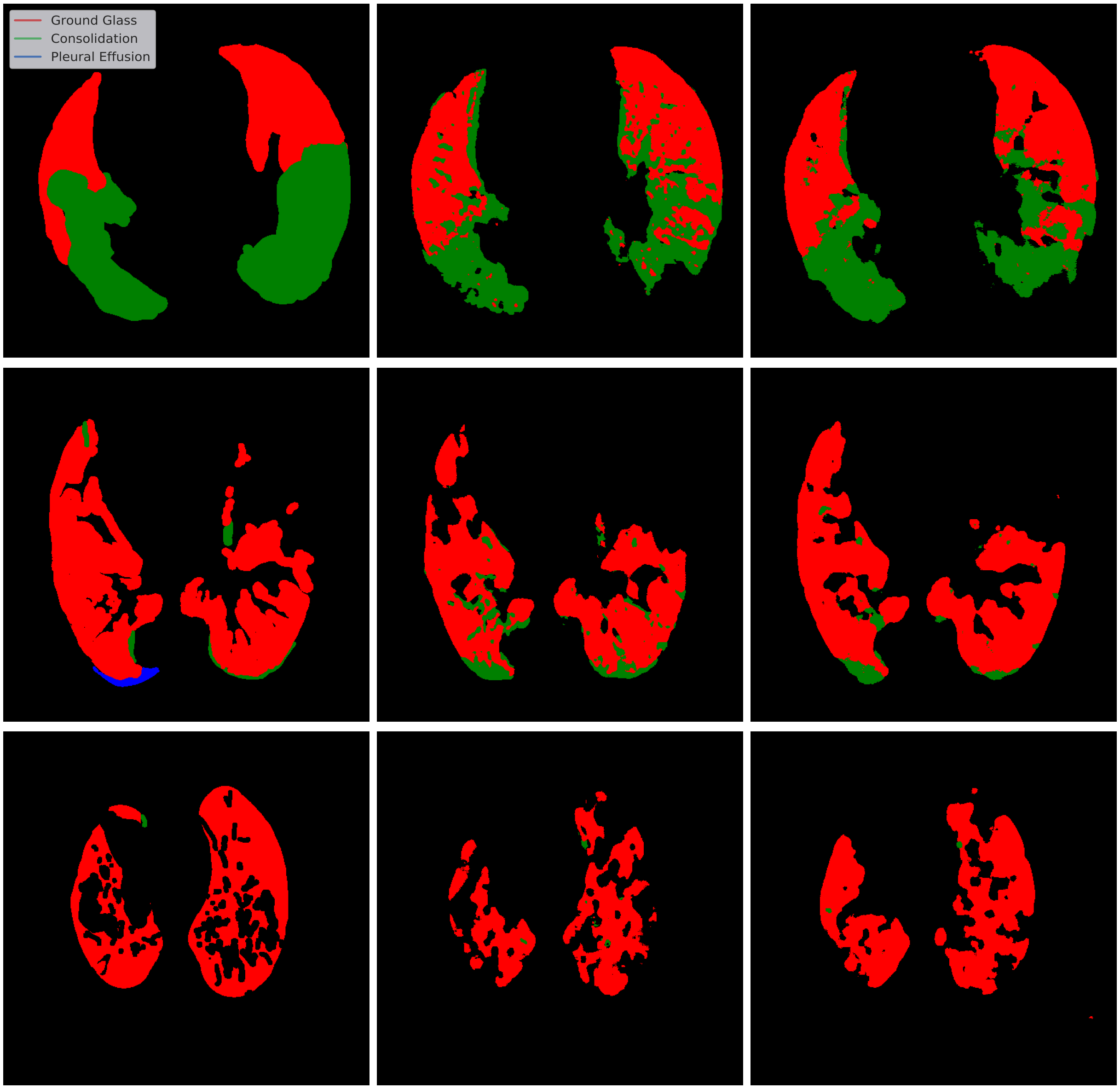


Figure 3.4: Multi-class segmentation results on the COVID-19 CT Segmentation dataset: Ground truth (left); U-Net (center); and Sharp U-Net (right). Notice how Sharp U-Net is able to generate smoother predictions compared to U-Net.

the Jaccard index of 2.63% is achieved. Also, the standard deviation of Wide Sharp U-Net is much lower than Wide U-Net (1.8024 and 3.4125, respectively). A similar pattern can be observed using the Dice coefficient. Another interesting finding from this experiment is that Sharp U-Net performs better than Wide U-Net, even though the latter has 9.1 million learnable parameters, whereas Sharp U-Net has only 7.8 million learnable parameters similar to U-Net.

Since recent works aim to improve performance on segmentation tasks by replacing the encoder subnetwork with pretrained convolutional networks such as VGG and ResNet, we also test the ex-

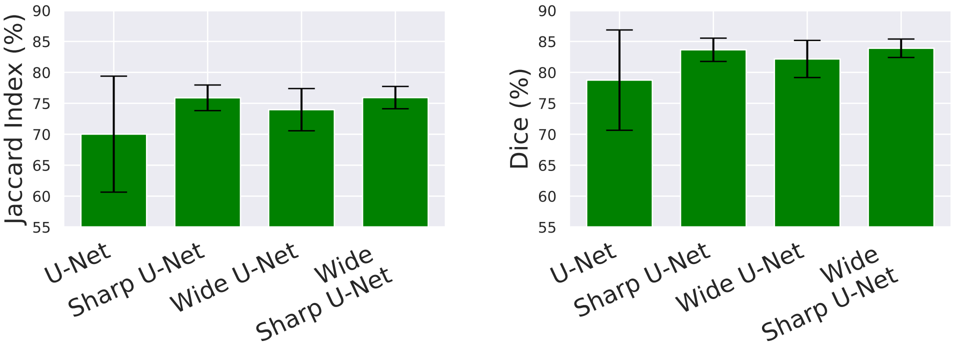


Figure 3.5: Bar graphs with error bars in terms of Jaccard and Dice metrics (in percent) on the CVC-ClinicDB dataset using 5-fold cross-validation. For both metrics, Sharp U-Net performs better than Wide U-Net, albeit the latter has much more learnable parameters.

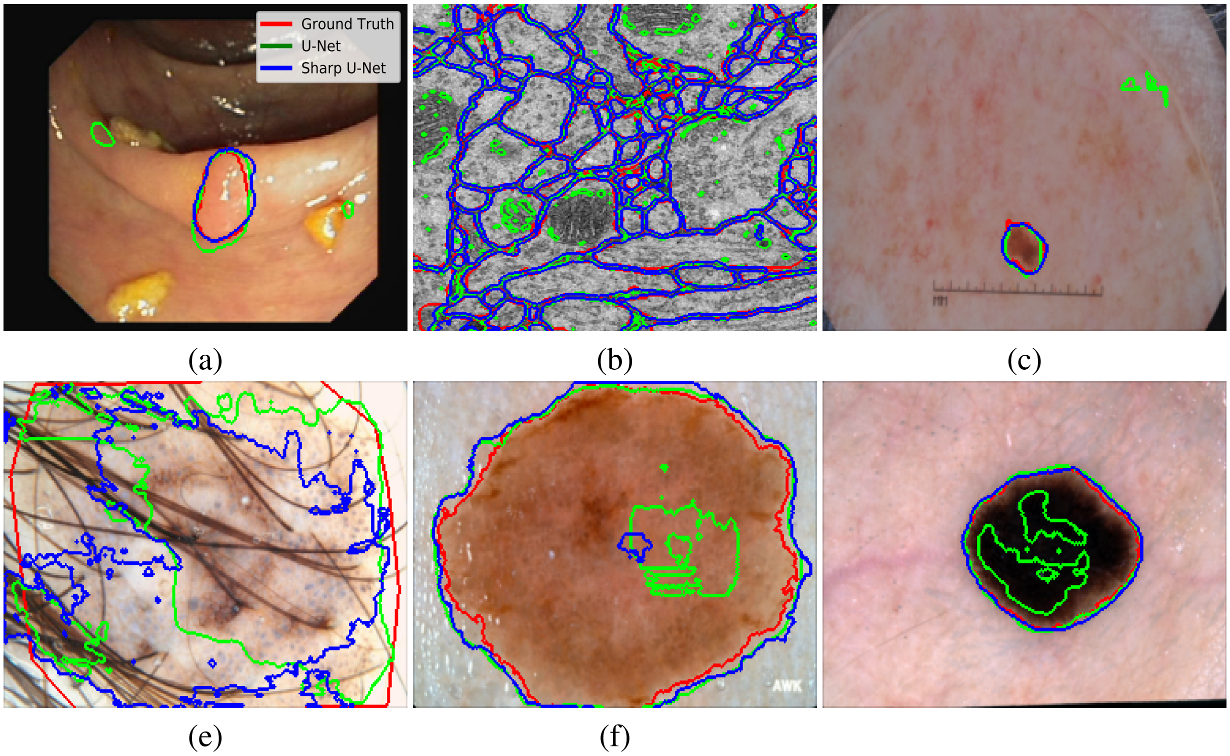
tensibility of training these pretrained models together with sharp blocks instead of the traditional skip connections. Table 3.2 shows 5-fold cross-validation test results of U-Net and U-Nets with VGG [42] and ResNet [78] encoders when training with skip connections and sharp blocks on the CVC-ClinicDB dataset. Notice that as the depth of the encoder increases, the Jaccard and Dice scores increase when trained with skip connection and sharp blocks. This is in line with findings reported in [42]. As the encoder complexity increases, the performance gains when adding sharp blocks decreases. This is intuitive as the closer the predicted segmentation is to the perfect segmentation, the harder it is to improve it further, and can be largely attributed to the lower improvement of sharp U-Net as the complexity of encoder increases as compared to U-Net. When using the VGG encoder, it can be seen that training with sharp blocks yields almost comparable performance, but with sharp blocks better results are obtained. A significant reduction in the standard deviation can be observed (0.8090 vs. 3.6895) when using sharp blocks, suggesting that sharp blocks are more reliable compared to the traditional skip connections. The same pattern can be observed when using the ResNet [78] encoder, which consists of 32.7 million learnable parameters. When comparing Sharp U-Net with VGG Sharp U-Net, we can see that the performance is comparable, but the standard deviation of Sharp U-Net is much lower than that of VGG Sharp U-Net, demonstrating the effectiveness of sharp blocks. This is an interesting result as Sharp U-Net has three times fewer learnable parameters (7.8 million) compared to VGG Sharp U-Net (23.7 million).

Sharp U-Net is robust to image irregularities. In Figure 3.6, we show qualitative results for some hard examples. These examples consist of segmenting ROIs that are occluded by artifacts such as hair or when the ROI and background are quite similar. While the segmented ROI predicted by Sharp U-Net is not perfect and kind of under-segmented, it yields substantially less amount of noise and fractured segmented ROIs compared to U-Net. For example, in Figure 3.6(b), it can be

Table 3.2: Comparison results in terms of Jaccard and Dice metrics (in percent) on the CVC-ClinicDB dataset using 5-fold cross-validation tests.

	CVC-ClinicDB	
Model	Jaccard (%)	Dice (%)
U-Net [5]	70.02 ± 9.37	78.75 ± 8.10
Sharp U-Net	75.89 ± 2.06	83.65 ± 1.88
TernausNet-16 [42]	76.07 ± 3.69	83.95 ± 2.95
Sharp TernausNet-16	76.33 ± 0.81	83.84 ± 0.75
U-Net + ResNet-50 [109]	83.88 ± 1.26	89.83 ± 1.25
Sharp U-Net + ResNet-50	**83.98** ± 0.27	**90.05** ± 0.29

observed that U-Net tends to segment non ROI regions in many areas of the image, whereas Sharp U-Net successfully segments most of the ROIs. Figures 3.6(e) and (f) show cases where U-Net tends to segment multiple non ROIs regions forming small clusters even though the ROI is easily distinguishable.



(a) (b) (c)

(e) (f)

Figure 3.6: Segmentation results of Sharp U-Net vs. U-Net on images with irregularities and hard-to-distinguish ROIs. Note that Sharp U-Net yields smoother predictions, while U-Net creates clusters of segmented regions in different parts of the images.

Mitigating under- and over-segmentation. In Figures 3.7 and 3.8, we illustrate some cases in which U-Net tends to over-segment or under-segment the ROIs. For example, we can see in Figure 3.7(b) that the U-Net tends to over-segment the ROIs even when there is a clear distinction between the foreground and background. More cases are shown in Figures 3.7(e) and (f), where the foreground is quite similar to the background. Even in such cases, Sharp U-Net is able to segment the ROIs better than U-Net, although Sharp U-Net tends to over-segment the ROIs in Figure 3.7(f). Figure 3.8(b) shows a low-contrast image from the CVC-ClinicDB dataset, and it can be seen that both Sharp U-Net and U-Net suffer from under-segmentation. However, Sharp U-Net is able to segment almost half of the ROIs, whereas U-Net forms two separate clusters. A similar pattern is observed in Figures 3.8(d) and (f), which show cases from the ISIC-2018 dataset for the task of segmenting lesions. Figures 3.8(a), (c) and (e) show cases, where Sharp U-Net yields near perfectly segmentation of the ROIs, even with almost no difference between the foreground and background image in Figure 3.8(a). This suggests that Sharp U-Net is able to achieve better performance, while tackling the over- and under-segmentation problems.

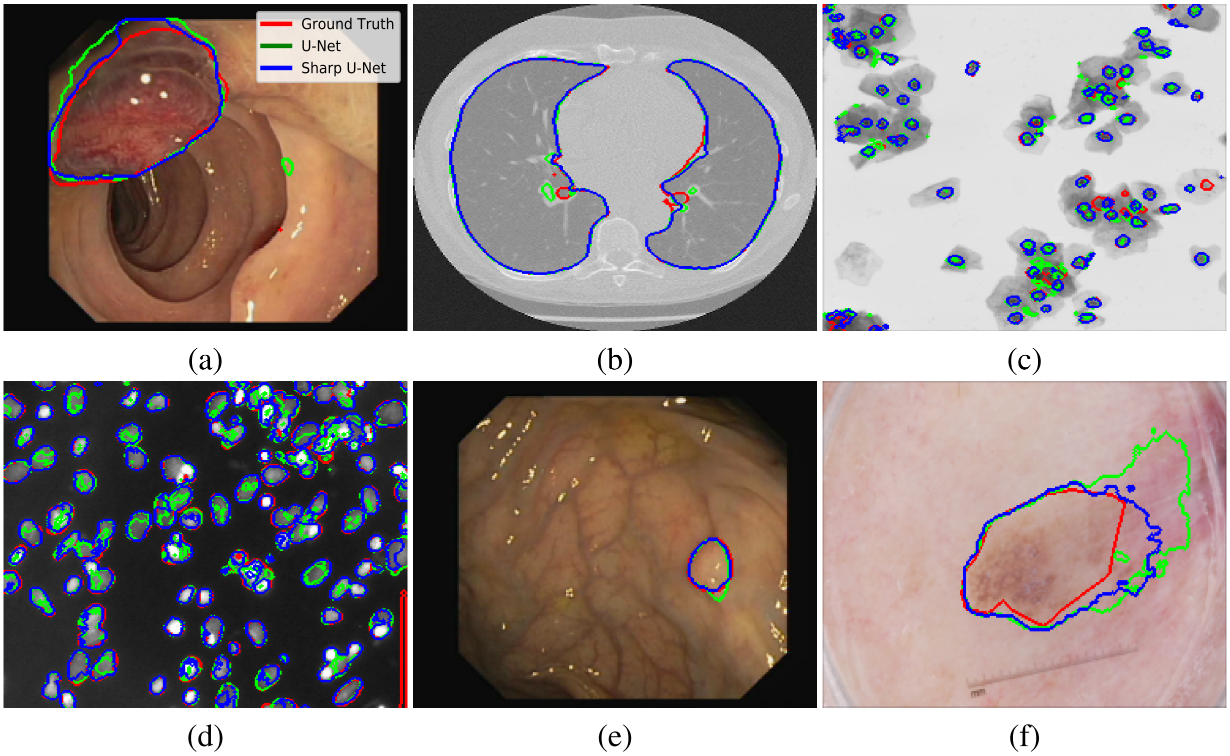
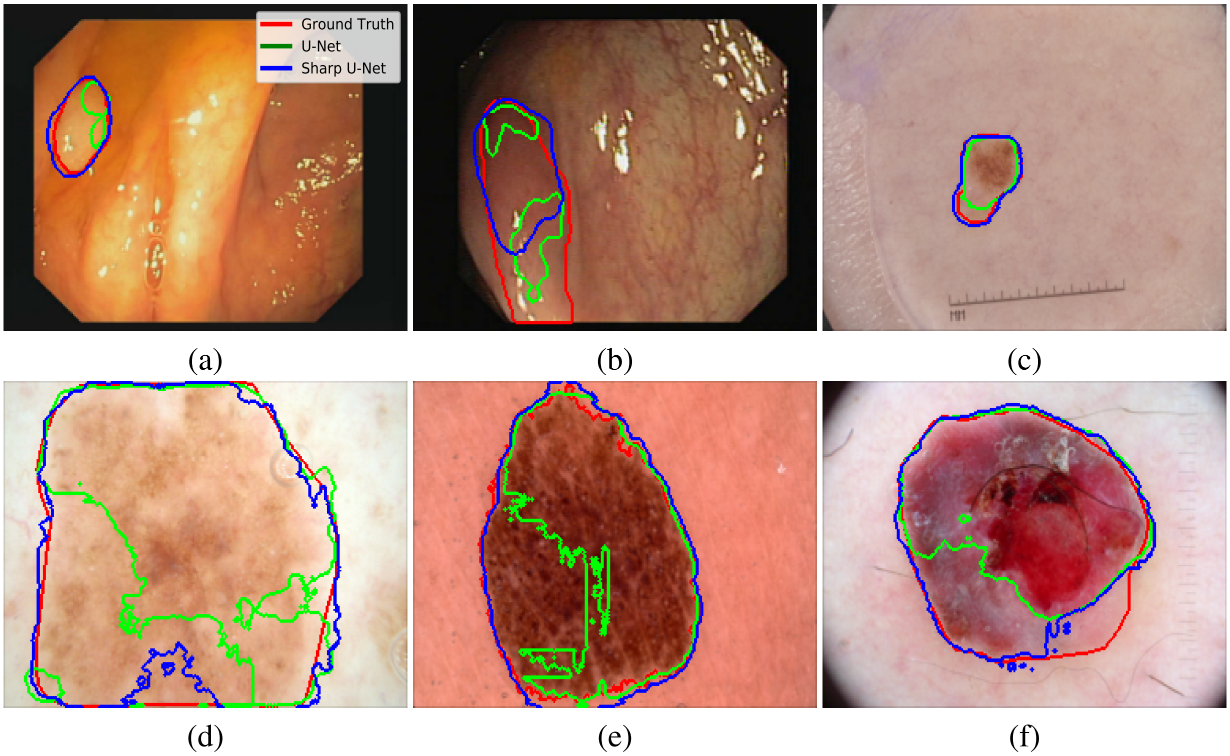


Figure 3.7: Segmentation results of Sharp U-Net vs. U-Net on hard examples. Notice that U-Net suffers from over-segmentation even though in some cases the ROI is easily differentiable from the background, while Sharp U-Net yields a segmentation result that closely matches the ground truth.

Feature visualization. Understanding and interpreting the predictions made by a deep learning



(a) (b) (c)

(d) (e) (f)

Figure 3.8: Segmentation results in case where U-Net tends to under-segment the ROI. In some cases that are easily differentiable, U-Net performs better to some extent, but Sharp U-Net tends to predict the ROIs much better.

model provides valuabe insights into the input data and the features learned by the model so that the results can be easily understood by human experts. In order to visually explain the decisions made by the proposed segmentation architecture and baseline methods for segmenting the ROIs, we use the gradient-weighted class activation map (Grad-CAM) [86] to generate the saliency maps, which highlight the most influential features affecting the predictions. Since convolutional features retain spatial information and that each feature value indicates whether the corresponding visual pattern exists in its receptive field, the output of the convolutional layer of the network shows the discriminative regions of the image.

In Figure 3.9, we visualize the class activation maps of the upsampling layers of two lung segmentation cases for both Sharp U-Net and U-Net. As can be seen, U-Net demonstrates high activations around the ROIs on the upsampling layer (i.e. first concatenation layer), which is the shallowest skip connection bridging the encoder decoder subnetworks. As the skip connections get deeper (i.e. second and third concatenation layers), the network shows high activations on the ROIs. For the deepest upsampling layer (i.e. fourth concatenation layer), lightly noticeable activations for the ROIs are observed. As shown in Figure 3.9, the opposite scenario is observed when using Sharp U-Net. It can be seen that the early upsampling layers show barely perceptible activations for the ROIs compared to U-Net, but exhibit high activations for the deeper upsampling layer, which shows high activations similar to the ground truth mask. Ideally, after the last upsampling layer operation and since the network outputs the final predictions, we should expect high activations for the ROIs similar to the ground truth mask. This better performance of Sharp U-Net suggests that the sharp blocks are able to better fuse the encoder-decoder features, yielding the merging of less semantically dissimilar features. This, in turn, enables the network to learn better representations of the input data and segment ROIs better than U-Net.

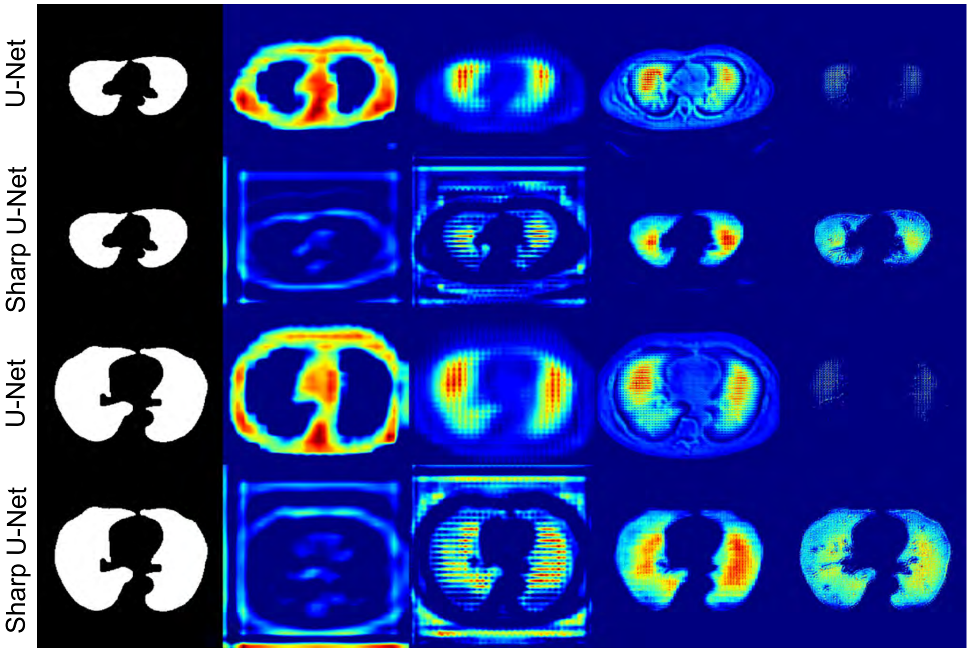


Figure 3.9: GradCAM heatmaps showing high activations for the regions affecting the predicted segmentation. For comparison between skip connections and sharp blocks, we visualize the upsampling layers of the decoder subnetwork. From left to right; ground truth and first to fourth concatenation layers.

CHAPTER 4

Conclusions and Future Work

This thesis has presented efficient deep learning models for imbalanced medical image classification and biomedical image segmentation. The proposed classification model uses generative adversarial networks to synthesize realistic looking target samples to over-sample the minority class for classifier training. The proposed image segmentation model, on the other hand, builds on the popular U-Net architecture by incorporating traditional image processing kernels in the U-Net skip connection during optimization. We have demonstrated through extensive experiments the superior performance of the proposed methods in comparison with existing techniques in the literature. In Section 4.1, the contributions made in each of the previous chapters and the concluding results drawn from the associated research work are presented. The limitations of the proposed approaches are discussed in Section 4.2. Suggestions for future research directions related to this thesis are also provided in Section 4.3.

4.1 Contributions of the Thesis

4.1.1 Deep Adversarial and Transfer Learning for Imbalanced Image Classification

In Chapter 2, we proposed a two-stage framework for imbalanced medical image classification, specifically melanoma and COVID-19 detection. The first stage addresses the problem of data scarcity and class imbalance by formulating inter-class variation as conditional image synthesis for over-sampling in order to synthesize under-represented class samples (e.g. melanoma from non-melanoma lesions). The newly synthesized samples are then used as additional data to train a deep convolutional neural network. We demonstrate through extensive experiments that the proposed MelaNet approach improves sensitivity by a margin of 13.10% and the AUC by 0.78% from only 1627 dermoscopy images compared to the baseline methods on the ISIC-2016 dataset. In the case of COVID-19 detection, we also demonstrated how the data generation procedure can serve as an anonymization tool by achieving comparable detection performance when trained only on synthetic data versus real data in an effort to alleviate data privacy concerns. Our experiments reveal that synthetic data can significantly improve the COVID-19 detection performance results, that as the amount of synthetic data is increased, sensitivity improves considerably and the number of false negatives decreases.

4.1.2 Sharpening Spatial Kernel based Encoder-Decoder Networks for Multimodal Image Segmentation

In Chapter 3, we introduced Sharp U-Net, a new encoder-decoder depthwise fully convolutional network architecture for binary and multi-class segmentation. The core idea is to convolve the output of the encoder feature map with a sharpening spatial filter prior to performing fusion with the decoder features. The proposed segmentation framework is not only able to make the encoder and decoder features semantically less dissimilar, but also helps smooth out artifacts throughout the network layers during the early stages of training due to untrained parameters. Experimental results demonstrate that our proposed architecture consistently outperforms or matches the state-of-the-art baselines on various benchmarks for binary and multi-class segmentation on biomedical images from different modalities. More importantly, our approach achieved significant performance improvements without adding any extra learnable parameters. In addition, we showed that Sharp U-Net can be easily scaled for improved performance, outperforming baselines that have three times the number of learnable parameters. For future work, we plan to explore effective techniques for handling the semantic gap between the encoder and decoder features, as well as to extend our approach to volumetric medical image segmentation.

4.2 Limitations

A key advantage of the proposed deep learning frameworks, namely MelaNet and Sharp U-Net, for multimodal medical image classification and segmentation described in this thesis is their ability to exploit discriminative information by learning several deep hierarchical nonlinear mappings, resulting in improved classification and segmentation performance. While deep learning models encode features more efficiently than shallow models, they are, however, prone to over-fitting due largely to the added layers of abstraction. In addition, the features learned by deep learning methods are in most situations not easily interpretable, as is the case with most neural networks. Even though we show salient features in an effort to explain the predictions made by these models, the lack of insight into the features may be considered one of the main disadvantages that the proposed deep learning methods have in comparison with classical methods. Another limitation of the MelaNet method we presented in Chapter 2 is that it requires an independent generative model for each domain, leading to prohibitive computational overhead for adversarial training.

4.3 Future Work

Several interesting research directions, motivated by this thesis, are discussed below:

4.3.1 MelaNet for Multi-class Medical Image Classification

We plan to address the multi-class classification problem, which requires an independent generative model for each domain, leading to prohibitive computational overhead for adversarial training. We aim to build an end-to-end model that can jointly tackle the synthesis and classification tasks in a multitask learning setting.

4.3.2 Traditional Image Processing Kernels to Reduce Semantic Gap

We plan to explore the integration of other types of image processing kernels (i.e Laplacian of Gaussian LoG operator) for handling the semantic gap between the encoder and decoder features in U-Net like architectures. We intend to conduct a comprehensive study on which kernel works best by performing extensive ablation studies.

4.3.3 Sharp U-Net for 3D Medical Image Segmentation

Volumetric medical image segmentation is an interesting problem in medical image analysis, where the goal is the segment not only a two-dimensional image, but also a sequence of images, similar to video data. We plan to extend our Sharp U-Net framework to 3D medical image segmentation.

References

[1] H. Ganster, P. Pinz, R. Rohrer, E. Wildling, M. Binder, and H. Kittler, “Automated melanoma recognition,” *IEEE Transactions on Medical Imaging*, vol. 20, no. 3, pp. 233–239, 2001.

[2] Y. Cheng, R. Swamisai, S. E. Umbaugh, R. H. Moss, W. V. Stoecker, S. Teegala, and S. K. Srinivasan, “Skin lesion classification using relative color features,” *Skin Research and Technology*, vol. 14, no. 1, pp. 53–64, 2008.

[3] Z. Liu and J. Zerubia, “Skin image illumination modeling and chromophore identification for melanoma diagnosis,” *Physics in Medicine & Biology*, vol. 60, pp. 3415–3431, 2015.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

[5] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, 2015.

[6] H. R. Roth, L. Lu, A. Seff, K. M. Cherry, J. Hoffman, S. Wang, J. Liu, E. Turkbey, and R. M. Summers, “A new 2.5 d representation for lymph node detection using random sets of deep convolutional neural network observations,” in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 520–527, 2014.

[7] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou, “Lung pattern classification for interstitial lung diseases using a deep convolutional neural network,” *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1207–1216, 2016.

[8] K. Matsunaga, A. Hamada, A. Minagawa, and H. Koga, “Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble,” *arXiv preprint arXiv:1703.03108*, 2017.

[9] N. C. Codella, Q.-B. Nguyen, S. Pankanti, D. A. Gutman, B. Helba, A. C. Halpern, and J. R. Smith, “Deep learning ensembles for melanoma recognition in dermoscopy images,” *IBM Journal of Research and Development*, vol. 61, no. 4/5, pp. 5–1, 2017.

[10] D. Gutman, N. C. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, and A. Halpern, “Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC),” *arXiv preprint arXiv:1605.01397*, 2016.

[11] L. Yu, H. Chen, Q. Dou, J. Qin, and P.-A. Heng, “Automated melanoma recognition in dermoscopy images via very deep residual networks,” *IEEE Transactions on Medical Imaging*, vol. 36, no. 4, pp. 994–1004, 2016.

[12] C.-K. Shie, C.-H. Chuang, C.-N. Chou, M.-H. Wu, and E. Y. Chang, “Transfer representation learning for medical image analysis,” in *Proc. International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 711–714, 2015.

[13] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, “Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning,” *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

[15] D. Nie, R. Trullo, J. Lian, C. Petitjean, S. Ruan, Q. Wang, and D. Shen, “Medical image synthesis with context-aware generative adversarial networks,” in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 417–425, 2017.

[16] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, “Synthetic data augmentation using gan for improved liver lesion classification,” in *Proc. IEEE International Symposium on Biomedical Imaging*, pp. 289–293, 2018.

[17] X. Yi, E. Walia, and P. Babyn, “Generative adversarial network in medical imaging: A review,” *Medical Image Analysis*, 2019.

[18] P. Costa, A. Galdran, M. I. Meyer, M. Niemeijer, M. Abràmoff, A. M. Mendonça, and A. Campilho, "End-to-end adversarial retinal image synthesis," *IEEE Transactions on Medical Imaging*, vol. 37, no. 3, pp. 781–791, 2017.

[19] T. Zhang, H. Fu, Y. Zhao, J. Cheng, M. Guo, Z. Gu, B. Yang, Y. Xiao, S. Gao, and J. Liu, "SkrGAN: Sketching-rendering unconditional generative adversarial networks for medical image synthesis," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 777–785, 2019.

[20] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE International Conference on Computer Vision*, pp. 2223–2232, 2017.

[21] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1125–1134, 2017.

[22] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Advances in Neural Information Processing Systems*, pp. 700–708, 2017.

[23] A. M. Lamb, D. Hjelm, Y. Ganin, J. P. Cohen, A. C. Courville, and Y. Bengio, "GibbsNet: Iterative adversarial inference for deep graphical models," in *Advances in Neural Information Processing Systems*, pp. 5089–5098, 2017.

[24] A. Ben-Cohen, E. Klang, S. P. Raskin, M. M. Amitai, and H. Greenspan, "Virtual PET images from CT data using deep convolutional networks: initial results," in *Proc. International Workshop on Simulation and Synthesis in Medical Imaging*, pp. 49–57, 2017.

[25] G. Yang, S. Yu, H. Dong, G. Slabaugh, P. L. Dragotti, X. Ye, F. Liu, S. Arridge, J. Keegan, Y. Guo, *et al.*, "DAGAN: deep de-aliasing generative adversarial networks for fast compressed sensing mri reconstruction," *IEEE Transactions on Medical Imaging*, vol. 37, no. 6, pp. 1310–1321, 2017.

[26] J. M. Wolterink, A. M. Dinkla, M. H. Savenije, P. R. Seevinck, C. A. van den Berg, and I. Išgum, "Deep MR to CT synthesis using unpaired data," in *Proc. International Workshop on Simulation and Synthesis in Medical Imaging*, pp. 14–23, 2017.

[27] T. Russ, S. Goerttler, A.-K. Schnurr, D. F. Bauer, S. Hatamikia, L. R. Schad, F. G. Zöllner, and K. Chung, "Synthesis of CT images from digital body phantoms using CycleGAN," *In-*

ternational Journal of Computer Assisted Radiology and Surgery, vol. 14, no. 10, pp. 1741–1750, 2019.

[28] M. T. Shaban, C. Baur, N. Navab, and S. Albarqouni, “StainGAN: Stain style transfer for digital histological images,” in *Proc. IEEE International Symposium on Biomedical Imaging*, pp. 953–956, 2019.

[29] T. de Bel, M. Hermsen, J. Kers, J. van der Laak, and G. Litjens, “Stain-transforming cycle-consistent generative adversarial networks for improved segmentation of renal histopathology,” in *Proc. International Conference on Medical Imaging with Deep Learning*, vol. 102, pp. 151–163, 2018.

[30] A. Bissoto, F. Perez, E. Valle, and S. Avila, “Skin lesion synthesis with generative adversarial networks,” in *Proc. International Workshop on Computer-Assisted and Robotic Endoscopy*, pp. 294–302, 2018.

[31] I. S. Ali, M. F. Mohamed, and Y. B. Mahdy, “Data augmentation for skin lesion using self-attention based progressive generative adversarial network,” *arXiv preprint arXiv:1910.11960*, 2019.

[32] J. P. Cohen, M. Luck, and S. Honari, “Distribution matching losses can hallucinate features in medical image translation,” in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 529–536, 2018.

[33] S. Kazeminia, C. Baur, A. Kuijper, B. van Ginneken, N. Navab, S. Albarqouni, and A. Mukhopadhyay, “GANs for medical image analysis,” *arXiv preprint arXiv:1809.06222*, 2018.

[34] P. Costa, A. Galdran, M. I. Meyer, M. D. Abràmoff, M. Niemeijer, A. M. Mendonça, and A. Campilho, “Towards adversarial retinal image synthesis,” *arXiv preprint arXiv:1701.08974*, 2017.

[35] S. U. Dar, M. Yurt, L. Karacan, A. Erdem, E. Erdem, and T. Çukur, “Image synthesis in multi-contrast MRI with conditional generative adversarial networks,” *IEEE Transactions on Medical Imaging*, vol. 38, no. 10, pp. 2375–2388, 2019.

[36] H.-C. Shin, N. A. Tenenholtz, J. K. Rogers, C. G. Schwarz, M. L. Senjem, J. L. Gunter, K. P. Andriole, and M. Michalski, “Medical image synthesis for data augmentation and anonymization using generative adversarial networks,” in *Proc. International Workshop on Simulation and Synthesis in Medical Imaging*, pp. 1–11, 2018.

[37] J. T. Guibas, T. S. Virdi, and P. S. Li, “Synthetic medical images from dual generative adversarial networks,” *arXiv preprint arXiv:1709.01872*, 2017.

[38] D. Korkinof, T. Rijken, M. O’Neill, J. Yearsley, H. Harvey, and B. Glocker, “High-resolution mammogram synthesis using progressive generative adversarial networks,” *arXiv preprint arXiv:1807.03401*, 2018.

[39] B. Teixeira, V. Singh, T. Chen, K. Ma, B. Tamersoy, Y. Wu, E. Balashova, and D. Comaniciu, “Generating synthetic X-ray images of a person from the surface geometry,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9059–9067, 2018.

[40] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, “UNet++: Redesigning skip connections to exploit multiscale features in image segmentation,” *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1856–1867, 2020.

[41] Y. Chen, B. Ma, and Y. Xia, “α-UNet++: A data-driven neural network architecture for medical image segmentation,” in *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*, pp. 3–12, 2020.

[42] A. A. Kalinin, V. I. Iglovikov, A. Rakhlin, and A. A. Shvets, “Medical image segmentation using deep neural networks with pre-trained encoders,” in *Deep Learning Applications*, pp. 39–52, 2020.

[43] N. Ibtehaz and M. S. Rahman, “MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation,” *Neural Networks*, vol. 121, pp. 74–87, 2020.

[44] J. M. J. Valanarasu, V. A. Sindagi, I. Hacihaliloglu, and V. M. Patel, “KiU-Net: Towards accurate segmentation of biomedical images using over-complete representations,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 363–373, 2020.

[45] Z. Mirikharaji, Y. Yan, and G. Hamarneh, “Learning to segment skin lesions from noisy annotations,” in *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*, pp. 207–215, 2019.

[46] Y. Ji, R. Zhang, Z. Li, J. Ren, S. Zhang, and P. Luo, “UXNet: Searching multi-level feature aggregation for 3D medical image segmentation,” in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 346–356, 2020.

[47] H. Zunair and A. B. Hamza, “Melanoma detection using adversarial training and deep transfer learning,” *Physics in Medicine & Biology*, vol. 65, 2020.

[48] H. Zunair and A. B. Hamza, “Synthesis of COVID-19 chest X-rays using unpaired image-to-image translation,” *Social Network Analysis and Mining*, vol. 11, no. 1, pp. 1–12, 2021.

[49] E. Saito and M. Hori, “Melanoma skin cancer incidence rates in the world from the cancer incidence in five continents XI,” *Japanese Journal of Clinical Oncology*, vol. 48, no. 12, pp. 1113–1114, 2018.

[50] R. Siegel, K. Miller, , and A. Jemal, “Cancer statistics, 2019,” *CA: A Cancer Journal for Clinicians*, vol. 69, pp. 7–34, 2019.

[51] M. Binder, M. Schwarz, A. Winkler, A. Steiner, A. Kaider, K. Wolff, and H. Pehamberger, “Epiluminescence microscopy: a useful tool for the diagnosis of pigmented skin lesions for formally trained dermatologists,” *Archives of Dermatology*, vol. 131, no. 3, pp. 286–291, 1995.

[52] N. K. Mishra and M. E. Celebi, “An overview of melanoma detection in dermoscopy images using image processing and machine learning,” *arXiv preprint arXiv:1601.07843*, 2016.

[53] L. Ballerini, R. B. Fisher, B. Aldridge, and J. Rees, “A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions,” in *Color Medical Image Analysis*, pp. 63–86, Springer, 2013.

[54] T. Tommasi, E. La Torre, and B. Caputo, “Melanoma recognition using representative and discriminative kernel classifiers,” in *Proc. International Workshop on Computer Vision Approaches to Medical Image Analysis*, pp. 1–12, 2006.

[55] M. E. Celebi, H. A. Kingravi, B. Uddin, H. Iyatomi, Y. A. Aslandogan, W. V. Stoecker, and R. H. Moss, “A methodological approach to the classification of dermoscopy images,” *Computerized Medical Imaging and Graphics*, vol. 31, no. 6, pp. 362–373, 2007.

[56] G. Schaefer, B. Krawczyk, M. E. Celebi, and H. Iyatomi, “An ensemble classification approach for melanoma diagnosis,” *Memetic Computing*, vol. 6, no. 4, pp. 233–240, 2014.

[57] T. Ai, Z. Yang, H. Hou, C. Zhan, C. Chen, W. Lv, Q. Tao, Z. Sun, and L. Xia, “Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases,” *Radiology*, 2020.

[58] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, *et al.*, "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China," *The Lancet*, vol. 395, no. 10223, pp. 497–506, 2020.

[59] M.-Y. Ng, E. Y. Lee, J. Yang, F. Yang, X. Li, H. Wang, M. M.-s. Lui, C. S.-Y. Lo, B. Leung, P.-L. Khong, *et al.*, "Imaging profile of the COVID-19 infection: radiologic findings and literature review," *Radiology: Cardiothoracic Imaging*, vol. 2, no. 1, 2020.

[60] M. Karim, T. Döhmen, D. Rebholz-Schuhmann, S. Decker, M. Cochez, O. Beyan, *et al.*, "DeepCOVIDExplainer: Explainable COVID-19 predictions based on chest X-ray images," *arXiv preprint arXiv:2004.04582*, 2020.

[61] L. Wang and A. Wong, "COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest radiography images," *arXiv preprint arXiv:2003.09871*, 2020.

[62] S. H. Kassani, P. H. Kassasni, M. J. Wesolowski, K. A. Schneider, and R. Deters, "Automatic detection of coronavirus disease (COVID-19) in X-ray and CT images: A machine learning-based approach," *arXiv preprint arXiv:2004.10641*, 2020.

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

[64] X. Li, C. Li, and D. Zhu, "COVID-MobileXpert: On-device COVID-19 screening using snapshots of chest X-Ray," *arXiv preprint arXiv:2004.03042*, 2020.

[65] M. Farooq and A. Hafeez, "COVID-ResNet: A deep learning framework for screening of COVID19 from radiographs," *arXiv preprint arXiv:2003.14395*, 2020.

[66] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald, "Preparing a collection of radiology examinations for distribution and retrieval," *Journal of the American Medical Informatics Association*, vol. 23, no. 2, pp. 304–310, 2016.

[67] A. E. W. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C. ying Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng, "MIMIC-CXR-JPG: A large publicly available database of labeled chest radiographs," *arXiv preprint arXiv:1901.07042*, 2019.

[68] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, *et al.*, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI Conference on Artificial Intelligence*, vol. 33, pp. 590–597, 2019.

[69] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2097–2106, 2017.

[70] A. Bustos, A. Pertusa, J.-M. Salinas, and M. de la Iglesia-Vayá, "PadChest: A large chest X-ray image dataset with multi-label annotated reports," *arXiv preprint arXiv:1901.07441*, 2019.

[71] J. P. Cohen, P. Morrison, and L. Dao, "COVID-19 image data collection," *arXiv preprint arXiv:2003.11597*, 2020.

[72] Z. Zhao, Q. Sun, H. Yang, H. Qiao, Z. Wang, and D. O. Wu, "Compression artifacts reduction by improved generative adversarial networks," *EURASIP Journal on Image and Video Processing*, vol. 2019, no. 1, p. 62, 2019.

[73] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Advances in Neural Information Processing Systems*, pp. 3320–3328, 2014.

[74] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[75] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017.

[76] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, *et al.*, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)," in *Proc. IEEE International Symposium on Biomedical Imaging*, pp. 168–172, 2018.

[77] T. Araújo, G. Aresta, E. Castro, J. Rouco, P. Aguiar, C. Eloy, A. Polónia, and A. Campilho, "Classification of breast cancer histology images using convolutional neural networks," *PloS one*, vol. 12, no. 6, p. e0177544, 2017.

[78] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[79] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, 2017.

[80] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *International Conference on Learning Representations*, 2015.

[81] M. D. Zeiler, “ADADELTA: an adaptive learning rate method,” *arXiv preprint arXiv:1212.5701*, 2012.

[82] F. Perez, C. Vasconcelos, S. Avila, and E. Valle, “Data augmentation for skin lesion analysis,” in *Proc. International Workshop on Computer-Assisted and Robotic Endoscopy*, pp. 303–311, 2018.

[83] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, no. 7639, p. 115, 2017.

[84] L. A. Jeni, J. F. Cohn, and F. De La Torre, “Facing imbalanced data–recommendations for the use of performance metrics,” in *Proc. Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 245–251, 2013.

[85] J. Brabec and L. Machlica, “Bad practices in evaluation methodology relevant to class-imbalanced problems,” *arXiv preprint arXiv:1812.01388*, 2018.

[86] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in *Proc. IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

[87] L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform manifold approximation and projection for dimension reduction,” *The Journal of Open Source Software*, 2018.

[88] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, “Transfusion: Understanding transfer learning for medical imaging,” in *Advances in Neural Information Processing Systems*, pp. 3342–3352, 2019.

[89] L. Liu, F.-X. Wu, Y.-P. Wang, and J. Wang, “Multi-receptive-field CNN for semantic segmentation of medical images,” *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 11, pp. 3215–3225, 2020.

[90] L. Li, M. Wei, B. Liu, K. Atchaneeyasakul, F. Zhou, Z. Pan, S. A. Kumar, J. Y. Zhang, Y. Pu, D. S. Liebeskind, and F. Scalzo, “Deep learning for hemorrhagic lesion detection and segmentation on brain CT images,” *IEEE Journal of Biomedical and Health Informatics*, 2020.

[91] B. Wang, S. Wang, S. Qiu, W. Wei, H. Wang, and H. He, “CSU-Net: A context spatial U-Net for accurate blood vessel segmentation in fundus images,” *IEEE Journal of Biomedical and Health Informatics*, 2020.

[92] J. Yang, H. Veeraraghavan, S. G. Armato III, K. Farahani, J. S. Kirby, J. Kalpathy-Kramer, W. van Elmpt, A. Dekker, X. Han, X. Feng, *et al.*, “Autosegmentation for thoracic radiation treatment planning: A grand challenge at AAPM 2017,” *Medical Physics*, 2018.

[93] S. Naik, S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski, “Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology,” in *Proc. IEEE International Symposium on Biomedical Imaging*, pp. 284–287, 2008.

[94] D. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, “Deep neural networks segment neuronal membranes in electron microscopy images,” in *Advances in Neural Information Processing Systems*, pp. 2843–2851, 2012.

[95] Y. Yuan, M. Chao, and Y.-C. Lo, “Automatic skin lesion segmentation using deep fully convolutional networks with Jaccard distance,” *IEEE Transactions on Medical Imaging*, vol. 36, pp. 1876–1886, 2017.

[96] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, 2015.

[97] P. F. Christ, M. E. A. Elshaer, F. Ettlinger, S. Tatavarty, M. Bickel, P. Bilic, M. Rempfler, M. Armbruster, F. Hofmann, M. D?Anastasi, *et al.*, “Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields,” in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 415–423, 2016.

[98] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 424–432, 2016.

[99] K. Sirinukunwattana, J. P. Pluim, H. Chen, X. Qi, P.-A. Heng, Y. B. Guo, L. Y. Wang, B. J. Matuszewski, E. Bruni, U. Sanchez, *et al.*, "Gland segmentation in colon histology images: The glas challenge contest," *Medical Image Analysis*, vol. 35, pp. 489–502, 2017.

[100] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," in *Deep Learning and Data Labeling for Medical Applications*, pp. 179–187, 2016.

[101] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2117–2125, 2017.

[102] N. Abraham and N. M. Khan, "A novel focal Tversky loss function with improved attention U-Net for lesion segmentation," in *Proc. IEEE International Symposium on Biomedical Imaging*, pp. 683–687, 2019.

[103] D. Karimi and S. E. Salcudean, "Reducing the hausdorff distance in medical image segmentation with convolutional neural networks," *IEEE Transactions on Medical Imaging*, vol. 39, no. 2, pp. 499–513, 2019.

[104] F. Caliva, C. Iriondo, A. M. Martinez, S. Majumdar, and V. Pedoia, "Distance map loss penalty term for semantic segmentation," *arXiv preprint arXiv:1908.03679*, 2019.

[105] A. Cardona, S. Saalfeld, S. Preibisch, B. Schmid, A. Cheng, J. Pulokas, P. Tomancak, and V. Hartenstein, "An integrated micro-and macroarchitectural analysis of the drosophila brain by computer-assisted serial section electron microscopy," *PLoS Biology*, vol. 8, no. 10, pp. 1–17, 2010.

[106] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, 2015.

[107] R. Azad, M. Asadi-Aghbolaghi, M. Fathy, and S. Escalera, "Bi-directional ConvLSTM u-net with densley connected convolutions," in *Proc. IEEE International Conference on Computer Vision Workshops*, pp. 406–415, 2019.

[108] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, 2018.

[109] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet-a: a deep learning framework for semantic segmentation of remotely sensed data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 94–114, 2020.