Deep Learning for Skin Lesion Segmentation

by

Zahra Mirikharaji

M.Sc., University of New Brunswick, 2015
B.Sc., Isfahan University of Technology, 2013

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
School of Computing Science
Faculty of Applied Sciences

SIMON FRASER UNIVERSITY
Summer 2022

Declaration of Committee

Name:	**Zahra Mirikharaji**
Degree:	**Doctor of Philosophy**
Thesis title:	**Deep Learning for Skin Lesion Segmentation**
Committee:	**Chair:** Mo Chen Assistant Professor, Computing Science
	Ghassan Hamarneh Supervisor Professor, Computing Science
	Manolis Savva Committee Member Assistant Professor, Computing Science
	Greg Mori Examiner Professor, Computing Science
	Jinman Kim External Examiner Associate Professor, School of Computer Science University of Sydney

Abstract

Skin cancer is a major public health problem requiring computer aided diagnosis to reduce the burden of disease's high incidence ratio and the associated expenses by assisting clinicians. Image segmentation, the task of decomposing an image into multiple regions by per pixel labeling, is a crucial step toward skin cancer diagnosis and treatments. However, the existence of natural and artificial artifacts (e.g. hair and air bubbles), intrinsic factors (e.g. lesion shape and contrast), and variation in image conditions originating from imaging tools and environments make skin lesion segmentation a challenging task. Recently, several efforts have been made to leverage the demonstrated superior performance of deep learning models in the segmentation of skin lesions from the surrounding healthy skin.

In this thesis, after a thorough examination of the studies leveraging the capability of deep learning models in skin lesion segmentation, we propose novel segmentation prediction models advancing state-of-the-art skin lesion segmentation techniques. First, we introduce deep learning based models that leverage the auxiliary information in the form of domain knowledge, contextual information, and labels consistency to regularize model parameters toward a more generalizable solution. Specifically, we encode high order shape prior knowledge into the loss function and also incorporate high-level semantic information in learning a sequence of deep models. Second, we study the limitations of ground truth pixels level annotations to effectively leverage limited reliable annotations. Specifically, we propose a robust to noise network by learning spatially adaptive weight maps associated with training images encoding the level of annotation noise to reduce the requirement of careful labeling. Also, we avoid single annotator bias, by training in an ensemble paradigm that handles inter-annotator disagreements and learns from all available annotations.

Keywords: skin cancer, image segmentation, deep learning, auxiliary information, annotation limitation

Acknowledgements

First, I would like to thank Prof. Ghassan Hamarneh, my senior supervisor for his constant support, patience, and encouragement and especially for his insightful feedback over the past few years. I would also like to thank my other collaborators, Prof. Emre Celebi, Dr. Sandra Avila, Dr. Catarina Barata, Prof. Eduardo Valle for their contributions to the survey.

I would also like to thank other members of my examination committee, Dr. Manolis Savva, Prof. Greg Mori, Dr. Jinman Kim, for their time and feedback on this Thesis.

I would further like to thank all members of the Medical Image Analysis Lab especially Kumar Abhishek, Saeed Izadi, Dr. Jeremy Kawahara for their help and support.

Finally, great thanks to my family for their love and endless support.

Table of Contents

List of Tables

List of Figures

List of Acronyms

Notation	Description
AC	Accuracy
ADV	Adversarial
ASM	Active Shape Models
BA	Balanced Accuracy
BNN	Bayesian Neural Network
Br	Brier score
CAD	Computer-Aided Diagnosis
CE	Cross Entropy
CNN	Convolutional Neural Network
CRF	Conditional Random Field
DCNN	Deep Convolutional Neural Network
DS	Deep Supervision
DUC	Dense Upsampling Convolution
EPE	End Point Error
F	F-measure
FCA	Factorized Channel Attention
FCN	Fully Convolutional Network
FN	False Negative
FP	False Positive
GAN	Generative Adversarial Network
GCN	Global Convolutional Network
GM	G_mean
IPCA	Intra-Patient Comparative Analysis
ISIC	International Skin Imaging Collaboration

Notation	Description
J	Jaccard index
LBP	Local-Binary Pattern
LFA	Lesion-Focused Analysis
LSTM	Long Short-Term Memory
MC	Monte Carlo
MCC	Matthews Correlation Coefficient
MLP	Multi-Layer Perceptron
MSE	Mean Squared Error
NLL	Negative Log-Likelihood
NPRI	Normalized Probabilistic Rand Index
OCT	Optical Coherence Tomography
PP	Pyramid Pooling
PR	Precision
RCL	Recurrent Convolutional Layer
RCNN	Recurrent Convolutional Neural Network
RE	Recall
SE	Sensitivity
SP	Specificity
TN	True Negative
TP	True Positive
WCE	Weighted Cross Entropy

Chapter 1

Introduction

1.1 Background and Motivation

The abnormal growth of skin cells known as skin cancer happening anywhere on skin is a global widespread health problem. Manual skin lesion screening and diagnosis is exhaustively tedious and time-consuming and suffers from inter- and intra-experts variations in diagnosis and treatments as well as limited reproducibility among clinicians. In addition, shortage and maldistribution of dermatologists affect diagnostic accuracy and lead to delayed and improper treatments. To broaden the expertise of dermatologists, artificial intelligence has been recently utilized to develop skin disease diagnostic tools toward assisting practitioners.

While in automatic analysis of skin lesions, several modalities including dermoscopy, clinical, optical coherence tomography (OCT), histopathology and patient meta-data have been utilized, in this thesis, we focus on dermosocopy and clinical images. Dermoscopy is a skin imaging technology widely practiced as a non-invasive diagnostic technique. Providing the magnified illuminated images as well as skin reflection suppression in dermoscopic images enables dermatologists to observe multiple features of skin surfaces and improve the diagnostic accuracy [196]. While dermoscopic images provide observable sub-surfaces of skin that cannot be seen by naked eye, they are not always available even for dermatologists [112]. On the other hand, clinical images acquired by conventional cameras are easily accessible but suffer from lower quality.

Segmentation, the task of partitioning the image into multiple meaningful objects with a set of specific labels, is an intermediate step in the dermatological analysis pipeline. While training end-to-end systems toward the final tasks (e.g., predicting diagnoses [187] or treatments [10]) has multiple advantages like computational efficiency and ease of optimization, the superior performance of end-to-end models requires sufficient labeled training data [249], with the performance of segmentation models shown to improve logarithmically with the volume of training data [317]. In medical image analysis tasks with small size datasets, incorporating prior knowledge including segmentation masks in training reduces the complexities of understanding the images by machines via extracting representative features

from lesions and leads to an improved diagnosis performance [369]. Further, estimation of diagnostic criteria of lesions such as shape, size, and boundary irregularities from segmentation maps are the bases of rule-based diagnostic systems like the ABCD rule [246] and its derivatives (ABCDE [5] and ABCDEF [172]). Skin lesion segmentation maps are required as input to some other image understanding tasks, such as synthesizing new image data given the segmentation mask [7] and adding 2D lesions to 3D meshes for analysis of 3D total-body skin surface scans [390]. Moreover, segmentation maps also improve the interpretability and the trust in machine decisions on diagnoses. However, as clinicians' and users' trust in CAD reach higher levels, it is possible that segmentation as an intermediate step in the medical image analysis pipeline may not be needed anymore.

Recently, researchers have applied successfully deep learning-based approaches by integrating automatic extraction of the representative features into a skin lesion segmentation network in a computationally efficient manner. In this thesis, we explore and tackle the challenges of deep segmentation models to delineate skin lesions from the healthy region. Fig. 1.1 depicts a simplified representation of the different components of a deep learning-based image segmentation model as well as our contribution, in context, highlighted in different colors. In particular, our main goal motivating this work is to advance the state-of-the-art of deep skin lesion segmentation techniques by

1) leveraging the auxiliary information in the form of domain knowledge, contextual information, and labels consistency in deep segmentation models to regularize model parameters toward a more generalizable solution and,

2) addressing the limitation of pixel-level annotations when learning a deep model.

Our contributions are extensively discussed in the following section.

1.2 Thesis Contributions

In this thesis, we first review the state-of-the-art on deep learning models applied to skin lesion dermoscopy and clinical images (subsection 1.2.1). Then, we introduce different approaches of encoding auxiliary information into deep skin lesion segmentation models (subsections 1.2.2-1.2.4). First, We train a sequence of deep models using high-level contextual features as well as image appearance. Degraded probability maps generated in early stopped models are utilized to prevent overfitting (subsection 1.2.2). We also leverage generative adversarial networks to impose high-order consistency in predicted segmentation maps by looking into the joint configuration of labels (subsection 1.2.3). Then, we introduce a differentiable form of the star shape prior as a regularization term in deep models' loss function to enforce learning plausible skin lesion segmentation (subsection 1.2.4). In the second part of the thesis, we study the limitation of pixel-level annotation to effectively leverage them toward learning a generalizable model. First, given a small set of reliable expert-level segmentation annotations and a large set of unreliable annotations, we train a robust to noise

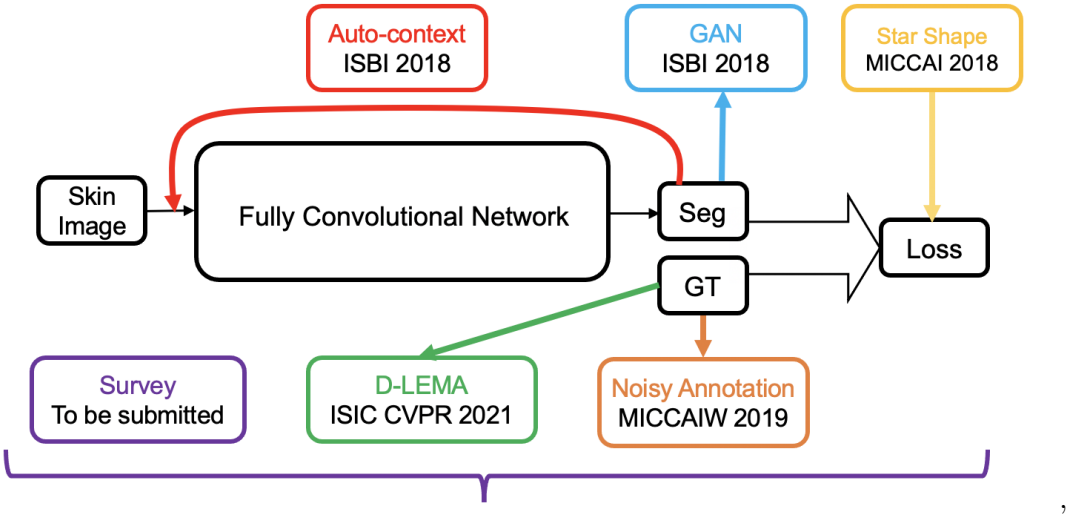


Figure 1.1: Our contributions in the thesis. Purple: surveying different components of deep skin lesion segmentation models (subsection 1.2.1), red: contextual information encoding in a sequence of models (subsection 1.2.2), blue: augmenting the segmentation model with a discriminator model (subsection 1.2.3), yellow: incorporation of star shape prior into the loss function(subsection 1.2.4), orange: learning from unreliable data (subsection 1.2.5), green: learning from multiple contradictory annotations (subsection 1.2.6).

model that learns spatially adaptive weight maps associated with training data to adjust the contribution of each pixel annotation in the loss function (subsection 1.2.5). Finally, we investigate inter-annotator disagreement and propose an ensemble paradigm modeling multiple experts' opinions toward learning from all available annotations (subsection 1.2.6). Here, we present a brief summary of each of the research contributions in this thesis.

1.2.1 Deep Learning for Skin Lesion Segmentation

Recently, several efforts have been made to leverage the demonstrated superior performance of deep learning models in the segmentation of skin lesions from the surrounding healthy skin. We cross-examine 134 research papers for the automatic segmentation of skin lesions in both clinical and dermoscopic images and present a thorough survey of the studies leveraging the capability of deep learning models in skin lesion segmentation approaches. We review the contributions of existing literature and analyze the works from several dimensions, comprising input data (datasets, pre-processing and synthetic data generation), model design (architecture, modules and losses), and evaluation aspects (data annotation and evaluation metrics)(see Fig. 1.2). We discuss those dimensions both from the viewpoint of selected seminal or influential works, and from a systemic viewpoint, examining how those choices have dictated current trends, and how their limitations should be addressed in the future. We summarize all examined works into a comprehensive table encoding the analyzed dimensions.

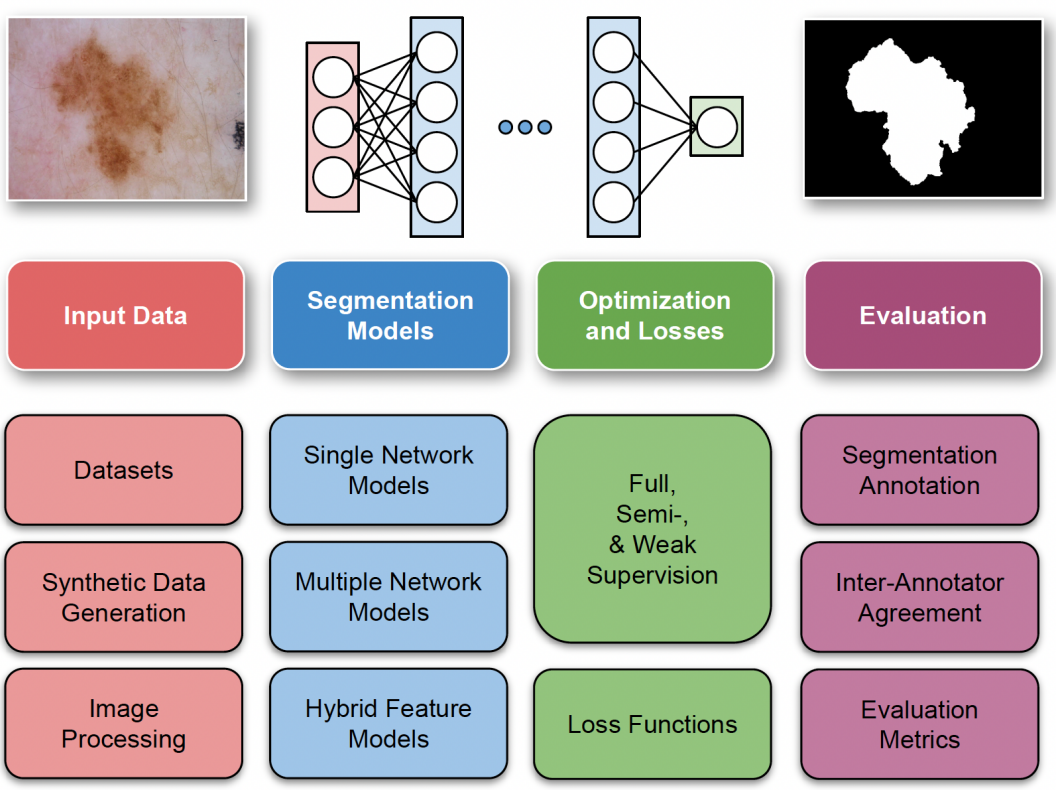


Figure 1.2: An overview of the topics covered in Chapter 2.

Contributions:

- The first thorough survey of 134 research papers on deep learning for skin lesion segmentation.
- Analyze the works from several aspects of deep models including input data model design (architecture, modules and losses), and evaluation aspects.
- Discuss different challenges facing automatic lesion delineation and future opportunities in this field.

The manuscript is submitted to Medical Image Analysis (MedIA) journal.

1.2.2 Deep Auto-context Fully Convolutional Neural Network for Skin Lesion Segmentation

Incorporating prior information of image context into image understanding models substantially improves dense prediction tasks. Auto context, originally proposed for patch-based segmentation, is an iterative learning algorithm for structural refinement, which incorporates contextual information into classical image understanding models [336]. Auto-context takes as input appearance information as well as features from the predicted probability

maps of the previous iteration into the current iteration. By iterating this process, classifiers can gradually correct earlier mistakes by using new contextual features. We propose an auto-context deep framework that sequentially learns improved skin lesion segmentation maps given RGB skin images. We train a sequence of FCNs so that each takes as input the original images as well as the degraded a posteriori probability map estimated by the previous early-stopped FCN. Feeding the whole contextual information into a CNN leads to automatic learning of deep multi-scale contextual features.

Contributions:

- The first work to encode explicitly the contextual information into deep networks in an auto-context fashion.
- Prevent overfitting in the subsequent models using the degraded probability maps generated by early-stopped FCNS.

This work was published in the IEEE International Symposium on Biomedical Imaging conference and has 16 citations to date.

[242] Zahra Mirikharaji, Saeed Izadi, Jeremy Kawahara, and Ghassan Hamarneh. Deep auto-context fully convolutional neural network for skin lesion segmentation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 877–880. IEEE, 2018

1.2.3 Generative Adversarial Networks to Segment Skin Lesions

GANs propose a new approach of learning parameters of a generative model while regularizing them via a discriminator model [133]. The generative model captures a data distribution by transforming a noise variable into a data sample and the discriminative model differentiates the generative model distribution from the data distribution. We propose to use a generative adversarial network to segment skin lesions.

Our aim is to practically examine the role of a critic network in improving the performance of an existing model. To this end, we use a fully convolutional segmentation model and augment it with a critic neural network model. The critic network receives the synthetic or real segmentation mask along with the input dermoscopy image and learns to distinguish between these two cases. We then backpropagate the error of the critic into the segmenter training procedure to encourage more realistic segmentation masks. Utilizing adversarial training while learning the segmentation model parameters encourages the high-order consistency in predicted segmentation masks by looking implicitly into the joint configuration of labels and distinguishing ground truth segmentation masks and model generated label maps.

Contributions:

- The first skin lesion deep model regularizing the model parameters by differentiating the fake and real data distributions

- Imposing high-order consistency in predicted segmentation masks by looking into joint configuration of labels

This work was published in the IEEE International Symposium on Biomedical Imaging conference and has 43 citations to date.
[164] Saeed Izadi, Zahra Mirikharaji, Jeremy Kawahara, and Ghassan Hamarneh. Generative adversarial networks to segment skin lesions. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 881–884. IEEE, 2018

1.2.4 Star Shape Prior in Fully Convolutional Networks for Skin Lesion Segmentation

Optimizing individual pixel-level class predictions in the FCNs loss function assigns independent class labels to image pixels without considering high-level label dependencies. On the other hand, incorporating prior knowledge about the structure of target objects to regularize plausible solutions with anatomically meaningful constraints has proven effective in traditional energy-based segmentation approaches to obtain more reliable delineations.

We propose a new loss term that encodes the star shape prior into the loss function of an FCN framework. We aim to harness the powerful proven capabilities of deep learning in automatically extracting learned (i.e., not hand-crafted) pixel-driven image features (i.e., likelihood) and augment them with demonstrably useful shape priors without requiring the knowledge of the target object pose. We penalize non-star shape segments in prediction maps and preserve global structures in the output space. Integration of the star shape prior to the loss function makes it possible to train the whole FCN framework in an end-to-end manner. In contrast to energy-based models incorporating the star shape prior, our approach to star shape prior in a deep learning setting not only eliminates the need for manually setting object centers, but also alleviates, at inference time, the computationally intensive optimization associated with the energy minimizing approaches.
Contributions:

- The first work that formulates a differentiable form of star shape prior in the loss function of an end-to-end trainable FCN framework.
- Penalize non-star shape segments in FCN prediction maps to preserve global structures in the output space and generate plausible skin lesion segmentation.

This work was published in the International Conference on Medical Image Computing and Computer-Assisted Intervention and has 78 citations to date.
[241] Zahra Mirikharaji and Ghassan Hamarneh. Star shape prior in fully convolutional networks for skin lesion segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 737–745. Springer, 2018

1.2.5 Learning to Segment Skin Lesions from Noisy Annotations

Requiring a large collection of images and their associated annotations is one of the main bottlenecks limiting the adoption of deep networks. In the task of medical image segmentation, requiring pixel-level semantic annotations performed by human experts exacerbate this difficulty. On the other hand, FCNs assume that reliable ground truth annotations are abundant, which is not always the case in practice, not only because collecting pixel-level annotation is time-consuming, but also since human-annotations are inherently noisy.

We propose a new framework to train a fully convolutional segmentation network from a large set of cheap unreliable annotations and a small set of expert-level clean annotations. Spatially adaptive weight maps associated with training images are learned to adjust the contribution of each pixel and treat clean and noisy pixel-level annotations in the loss function. The importance weights are assigned to pixels based on the pixel-wise loss gradient directions. A meta-learning approach is integrated at every training iteration to approximate the optimal weight maps of the current batch based on the CE loss on a small set of skin lesion images annotated by experts. Learning the deep skin lesion segmentation network and spatially adaptive weight maps are performed in an end-to-end manner.

Contributions:

- The first robust to noise deep network for segmentation task.
- Leveraged a limited amount of cleanly-annotated data to learn a robust-to-noise deep segmentation network.
- The first work to learn spatially adaptive weight maps to effectively leverage different levels of annotation reliability in learning.

This work was published in the International Conference on Medical Image Computing and Computer Assisted Interventions, workshop of Medical Image Learning with Less Labels and Imperfect Data and has 48 citations to date.

[243] Zahra Mirikharaji, Yiqi Yan, and Ghassan 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*, pages 207–215. Springer, 2019

1.2.6 Deep Learning Ensembles from Potentially Contradictory Multiple Annotations

Medical image segmentation annotations suffer from inter- and intra-observer variations even among experts due to intrinsic differences in human annotators and ambiguous boundaries. That is why evaluation using manual segmentations outlined by multiple experts is important. Although training deep models in a supervised setting with a single annotation per image has been extensively studied, generalizing their training to work with datasets containing multiple annotations per image remains a fairly unexplored problem.

We propose an approach to handle annotators' disagreements when training a deep model. An ensemble of Bayesian fully convolutional networks (FCNs) is proposed for the segmentation task by considering two major factors in the aggregation of multiple ground truth annotations: (1) handling contradictory annotations in the training data originating from inter-annotator disagreements and (2) improving confidence calibration through the fusion of base models' predictions. Our hypothesis is that given a new image, leveraging different experts' skills independently and fusing them in an ensemble model, while considering their estimated uncertainty, makes for a more reliable final prediction. To handle contradictory annotations arising from having multiple annotations per image during the training, we partition the entire dataset into M disjoint subsets containing one unique annotation for each image. Then in an ensemble setting, we train M robust-to-annotation-noise deep model to efficiently leverage the multiple experts' opinions toward learning from all available annotations. Our model also captures two types of uncertainty, aleatoric uncertainty modeled in the training loss function and epistemic uncertainty modeled in the ensemble paradigm, to improve confidence calibration.

Contributions:

- Propose an ensemble paradigm to: (1)model different experts' skills independently. (2) deal with discrepancies in segmentation annotations.
- A robust-to annotation-noise learning scheme is utilized to efficiently leverage experts' opinions toward learning from all available annotations.

This work was published in the IEEE Computer Vision and Pattern Recognition (IEEE CVPR) ISIC Skin Image Analysis Workshop (CVPR ISIC), won the best paper award and has 7 citations to date.
[240] Zahra Mirikharaji, Kumar Abhishek, Saeed Izadi, and Ghassan Hamarneh. D-LEMA: Deep learning ensembles from multiple annotations-application to skin lesion segmentation. pages 1837–1846, 2021

1.2.7 Other Contributions

In addition to the contributions listed above, a number of other publications are completed during my doctoral studies. These works are listed below in chronological order.

- **Zahra Mirikharaji**, Mengliu Zhao, and Ghassan Hamarneh. Globally-Optimal Anatomical Tree Extraction from 3D Medical Images using Pictorial Structures and Minimal Paths (Mirikharaji and Zhao: Joint first authors). In *Lecture Notes in Computer Science, Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 10434, pages 242-250, 2017.

Table 1.1: Public skin lesion datasets with segmentation annotations.

dataset	year	modality	size	train/val./test	class distribution	additional info	section
DermoFit [33]	2013	clinical	1300	-	1224 non-melanoma 76 melanoma	8-bit RGB images of sizes ranging from 177×189 to 2176×2549 pixels captured at a controlled lighting situation and the same distance	subsections 1.2.3, 1.2.6
Pedro Hispano Hospital (PH^2) [237]	2013	Dermoscopy	200	-	160 benign nevi 40 melanoma	8-bit RGB images of sizes 553×763 to 577×769 pixels acquired at 20× magnification.	subsection 1.2.6
ISIC2016 [144]	2016	Dermoscopy	1279	900/-/379	Train: 727 non-melanoma 173 melanoma Test: 304 non-melanoma 75 melanoma	8-bit RGB images of sizes ranging from 566×679 to 2848×4288 pixels.	subsection 1.2.2
ISIC2017 [89]	2017	Dermoscopy	2750	2000/150/600	Train: 1626 non-melanoma 374 melanoma Test: 483 non-melanoma 117 melanoma	8-bit RGB images of sizes ranging from 540×722 to 4499×6748 pixels.	subsections 1.2.4, 1.2.5
ISIC-archive [1, 89, 88]	2016-now	Dermoscopy	70,000	-	-	8-bit RGB images of sizes up to 5 segmentation ground truth.	subsection 1.2.6

- Saeede Afshari, Aicha BenTaieb, **Zahra Mirikharaji**, and Ghassan Hamarneh. Weakly Supervised Fully Convolutional Network for PET Lesion Segmentation. In *SPIE Medical Imaging*, volume 10949, pages 1-7, 2019

- Saeed Izadi, **Zahra Mirikharaji**, Mengliu Zhao, and Ghassan Hamarneh. WhiteNNer - Blind Image Denoising via Noise Whiteness Priors. In *International Conference on Computer Vision workshop on Visual Recognition for Medical Images (ICCV VRMI)*, pages 476-484, 2019.

1.3 Summary

Overall, toward the improvement of deep models for skin lesion segmentation tasks, this thesis studies different approaches of encoding auxiliary information into deep networks. The thesis also discusses the limitations of ground truth pixels level annotations and proposes approaches to effectively leverage limited reliable annotations, reduce the requirement of careful labeling, handle inter-annotator disagreements while avoiding single annotator bias. We evaluated our contributions using five different publicly available skin lesion datasets (see table 1.1), all containing pixel-level annotations.

Chapter 2

Deep Learning for Skin Lesion Segmentation

2.1 Introduction

Segmentation is a challenging and critical operation in the automated skin-lesion analysis workflow. Shape information, such as size, symmetry, border definition, and regularity are important diagnostic criteria for skin cancers. Both the surgical excision and the radiation therapy of skin cancers require localization and delineation of lesions [3]. Manual delineation is a time-consuming, laborious task suffering from severe inter- and intra-observer variability. A fast and reliable segmentation is, thus, an integral part of the effective computer-aided diagnosis (CAD) for skin lesions. Recent studies show the utility of segmentation masks in improving the classification performance for certain diagnostic classes by allowing the dilated cropping of lesion images [232] and the removal of skin lesion imaging-related artifacts [233].

In this Chapter, we review several works on deep learning models for the segmentation of skin lesions from the surrounding healthy skin. We cross-examine 134 research papers for the automatic segmentation of skin lesions in both clinical and dermoscopic images and present a thorough survey of the studies leveraging the capability of deep learning models in skin lesion segmentation approaches. We review the contributions of existing literature and analyze the works from several dimensions, comprising input data (datasets, preprocessing and synthetic data generation), model design (architecture, modules, and losses), and evaluation aspects (data annotation, and evaluation metrics). We discuss those dimensions both from the viewpoint of selected seminal or influential works, and from a systemaic viewpoint, examining how those choices have dictated current trends, and how their limitations should be addressed in the future. We summarize all examined works into a comprehensive table encoding the analyzed dimensions.

2.1.1 Skin Cancer

Skin cancer and its associated expenses ($8.1 billion annually in U.S. [145]) have grown into a major publichealth issue in the past decades. In the USA alone, 99,780 new cases of melanoma are expected in 2022 siegel2022. Broadly speaking, there are two types of skin cancer: melanomas and non-melanomas, the former making up just 1% of the cases, but the majority of the deaths due to its aggressiveness.

Early diagnostic plays a critical role for a good prognosis: melanoma can be cured with a simple outpatient surgery, if detected early, but its five-year survival rate drops from 99% to 25% if it is diagnosed at an advanced stage [3].

2.1.2 Diagnosing Skin Diseases

Visual inspection by clinicians is the primary step of clinical screening for skin cancers. Two popular strategies for lesion analysis commonly used by experts are intra-patient comparative analysis (IPCA) and lesion-focused analysis (LFA). The "ugly duckling" sign is the strategy used in IPCA which compares the individual lesions to detect outliers as suspicious spots. On the other hand, LFA utilizes an algorithm to look into the morphological criteria of lesions [126]. ABCD rules (Asymmetry, Border, Color, Diameter of moles) [246], ABCDE rules (ABCD plus Evolution of moles) [5] and 7-point checklist [22] are widely used diagnostic algorithms.

In automatic skin images analysis, imaging tools provide two modalities of images: dermoscopic microscopic images and macroscopic clinical images. While dermoscopic images provide observable sub-surfaces of skin that cannot be seen by the naked eye, they are not always available even for dermatologists [112]. On the other hand, clinical images acquired by conventional cameras are easily accessible but suffer from lower quality.

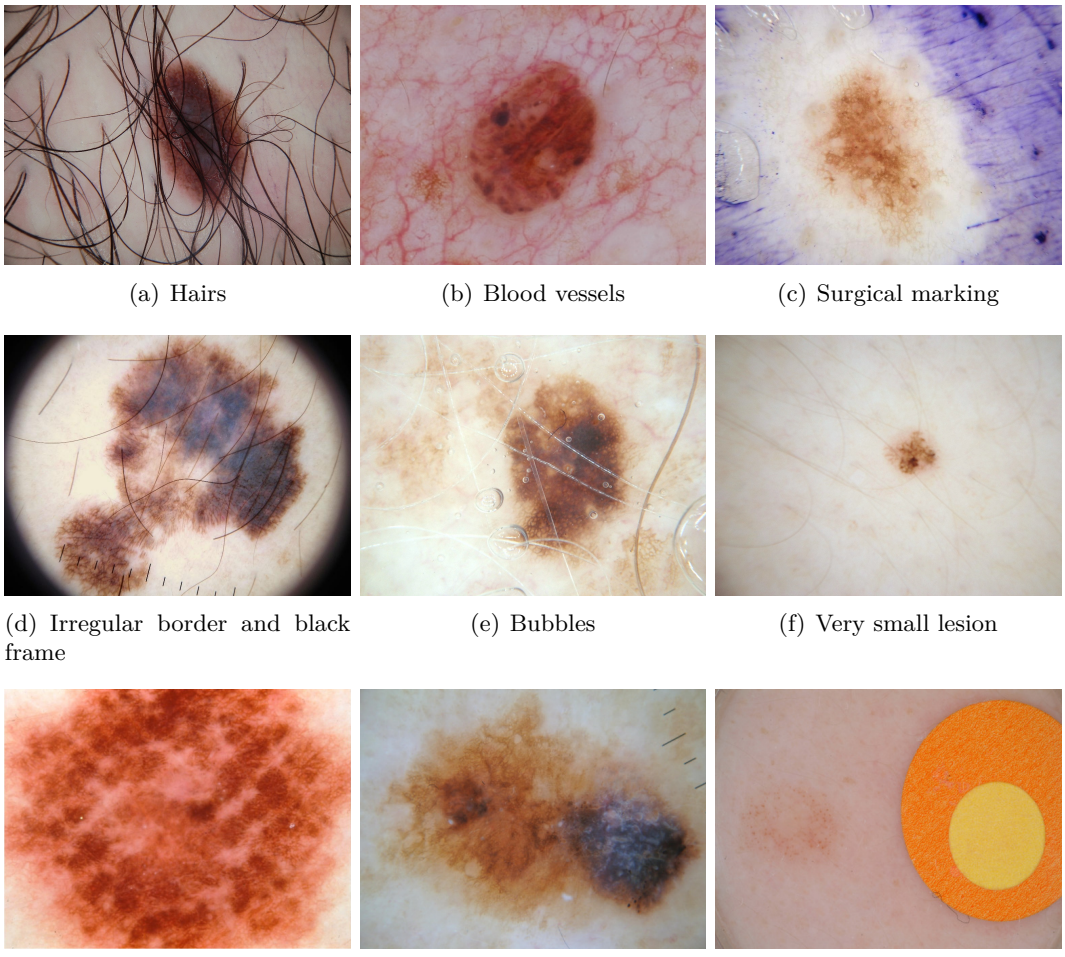
Dermoscopy is a non-invasive skin imaging technique that enables dermatologists to observe multiple features of skin surfaces and improve the diagnostic accuracy [196]. However, even while utilizing dermoscopy and different skin lesion analysis strategies, the diagnostic accuracy of skin condition varies from 24% to 77% depending on the clinicians' level of expertise [332]. Moreover, dermoscopy may actually lower the diagnostic accuracy in the hands of inexperienced dermatologists [54]. Therefore, to minimize the diagnostic errors that result from the difficulty and subjectivity of visual interpretation and to ameliorate the burden of skin disease and limited access to dermatologists, the development of CAD methods is crucial to provide faster and more accurate screening results.

2.1.3 Segmentation Challenges

Segmentation is the partition of an image into meaningful regions. Semantic segmentation, in addition, assigns appropriate class labels to each region. For skin lesions, the task is almost always binary, separating the lesion from the surrounding healthy skin. Skin-lesion

segmentation is hindered by illumination and contrast issues, intrinsic inter-class similarities and intra-class variability, occlusions, artifacts, and diversity of imaging tools and conditions, making automated segmentation challenging . The lack of large datasets with ground truth segmentation masks by experts compound the problem, hindering both the training of models and their reliable evaluation.

Skin lesion images, in particular, are occluded by natural artifacts like hair (Fig. 2.1(a)), blood vessels (Fig. 2.1(b)), and artificial ones like surgical marker annotations (Fig. 2.1(c)), lens artifacts (dark corners) (Fig. 2.1(d)), and air bubbles (Fig. 2.1(e)). Intrinsic factors like lesion size and shape variation (Fig. 2.1(f) and 2.1(g)), different skin colors (Fig. 2.1(h)), low contrast (Fig. 2.1(i)), and ambiguous boundaries (Fig. 2.1(h)) especially at the early stages of cancer, varying between different lesion instances is a critical issue.



(a) Hairs (b) Blood vessels (c) Surgical marking

(d) Irregular border and black frame (e) Bubbles (f) Very small lesion

(g) Very large lesion (h) Fuzzy border and variegated coloring (i) Low contrast and color calibration chart

Figure 2.1: Factors that complicate dermoscopy image segmentation (image source: ISIC 2016 image set [144]).

Before the deep learning revolution, segmentation was based on classical image processing and machine learning techniques such as adaptive thresholding [138], contour-based optimizations [113], region-growing [163], unsupervised clustering [130], and support vector machines [396]. Those approaches depended on hand-crafted features, which were difficult to engineer and often limited the technique invariance and discriminative power from the outset. As a result, performances dropped when extending those approaches to larger and more complex datasets. In contrast, deep learning integrates feature extraction and task-specific decision seamlessly, and not just cope with, but actually require larger datasets.

2.1.4 Survey of Surveys

Celebi et al. [70] reviewed 18 skin-lesion border detection algorithms in dermoscopic images, published between 1998 and 2008, with their required pre- and post-processing steps. Celebi et al. [74] extended that work with 32 additional techniques published between 2009 and 2014, discussing performance evaluation and computational requirements of each approach, and suggesting guidelines for future works. Both surveys appeared before deep learning was widely adopted for skin-lesion segmentation, but they broadly comprise all the important works based on classical machine learning.

Adeyinka et al. [15] analyzed comparatively 20 state-of-the-art skin-lesion segmentation approaches, highlighting their advantages and disadvantages. They benchmarked the performance of different algorithms on a skin dataset and concluded that deep learning surpasses other methodologies.

Baig et al. [31] reviewed deep learning approaches, and their preprocessing, for both skin-lesion segmentation and classification. They reviewed seven deep-learning methods for border detection, categorized them according to deep-learning architecture, and compared them to popular classical approaches to demonstrate their superior performance. Adegun et al. [13] reviewed the literature for deep-learning based skin-image analysis, with focus on the best-performing methods on ISIC (International Skin Imaging Collaboration) Skin Image Analysis Challenges 2018 [88] and 2019 [333, 89, 92].

2.1.5 Main Contributions

No existing survey approaches the present work in breadth or depth, as we cross-examine 134 research papers for the automatic segmentation of skin lesions in both clinical and dermoscopic images. We analyze the works from several dimensions, comprising input data (datasets, preprocessing, synthetic data generation), model design (architecture, modules, losses), and evaluation (data annotation, evaluation metrics). We discuss those dimensions both from the viewpoint of selected seminal or influential works, and from a systematic viewpoint, examining how those choices have dictated current trends, and how their limitations should be addressed in the future. We summarize all examined works into a comprehensive table encoding the analyzed dimensions.

2.1.6 Search Strategy

We searched DBLP and Arxiv Sanity Preserver for all scholarly publications: peer-reviewed journal articles, conference and workshop proceedings, and non-peer-reviewed preprints from 2014 to 2021. The DBLP search query was `(conv* | deep | neural | learn*) (skin | derm*) (segment* | delineat* | extract* | localiz*)`, thus restricting our search to deep learning-based works involving skin and segmentation. We chose DBLP for our literature search because (a) it allows for customized search queries and lists, and (b) we did not find any relevant publications on other platforms (Google Scholar and PubMed) that were not indexed by DBLP. For including unpublished preprints, we also searched on Arxiv Sanity Preserver using a similar query[1]. We filtered our search results to remove any false positives and included papers related only to skin lesion segmentation. We excluded papers that focused on general skin segmentation, general skin conditions (e.g., psoriasis, acne), or certain sub-types of skin lesions. We also included unpublished preprints on arXiv, which passed minimum quality checks levels and excluded those clearly of low quality. In particular, papers that had one or more of the following were excluded from this survey: (a) missing quantitative results, (b) missing important sections such as Abstract or Methods, (c) conspicuously poor writing quality, and (d) no methodological contribution.

The remaining text is organized as follows. In Section 2.2, we introduce publicly available datasets and discuss preprocessing and synthetic data generations. In Section 2.3, we discuss different categories of network architectures in deep segmentation models and discuss how deep models benefit from these networks. We also talk about different loss functions designed either generally or specifically for skin lesion segmentation task. In Section 2.4, we discuss segmentation evaluation techniques and measures. Finally, in Section 2.5, we discuss the open challenges and deficiencies in the deep skin lesion segmentation methods and conclude the surveyed studies.

2.2 Input Data

Obtaining data in sufficient quantity and quality is often the main challenge for obtaining effective models. State-of-the-art segmentation models have a huge number of adjustable parameters, allowing them to generalize well, provided they are trained on massive labeled datasets [64]. Unfortunately, skin-lesion datasets — like most medical image datasets — tend to be small [98] due to issues of copyright, patient privacy, acquisition/annotation cost and standardization, and scarcity of many pathologies of interest.

The two modalities of skin-lesion images used for training models are *clinical images*, which are close-ups of the lesions obtained by macrophotography with conventional cameras, and *dermoscopic images*, which are obtained by dermoscopy, a non-invasive skin imag-

[1] **Arxiv Sanity Preserver**: `http://www.arxiv-sanity.com/search?q=segmentation+skin+melanoma`

ing through optical magnification, and either liquid immersion and low angle-of-incidence lighting, or cross-polarized lighting. Dermoscopy eliminates skin surface reflections [196], reveals subsurface skin structures, and allows the identification of dozens of morphological features such as atypical pigment networks, dots/globules, streaks, blue-white areas, and blotches [239].

Annotation is often the greatest barrier for increasing data availability. Segmentation requires laborious *region-based annotation*, where an expert manually outlines the region where the lesion (or a clinical feature) appears in the image. That contrasts with more conventional *textual annotation*, which include diagnosis (e.g., melanoma, carcinoma, benign nevi), presence/absence/score of dermoscopic features (e.g., pigment networks, blue-white areas, streaks, globules), diagnostic strategy (e.g., pattern analysis, ABCD rule, 7-point checklist, 3-point checklist), clinical metadata (e.g., sex, age, anatomic site, familial history, etc.), and other details (e.g., timestamp, camera model) [65]. We discuss the issue of image annotation extensively in subsection 2.4.1.

2.2.1 Datasets

The appearance of larger, more diverse, better-annotated datasets is one of the main factors for the advances of dermatological image analysis in the past decade [69]. Works in dermatological images date back to the mid-1990s [53, 122], but until the mid-2000s, they overwhelmingly used small, private image sets, containing few hundred images.

The *Interactive Atlas of Dermoscopy* (sometimes called *Edra Atlas*, in reference to the publisher) by Argenziano et al. [23] mitigated the issue by providing a CD-ROM with $1,039$ dermoscopy images (26% melanomas, 4% carcinomas, 70% nevi) of $1,024 \times 683$ pixels, acquired by three European university hospitals (University of Graz, Austria, University of Naples, Italy, and University of Florence, Italy). The works of Celebi et al. [72, 71] popularized the dataset in the dermoscopy image analysis community, where it became a *de facto* evaluation standard for almost a decade, until the much larger ISIC Archive datasets (see below) were available. Recently, Kawahara et al. [187] placed this valuable dataset, along with additional textual annotations based on the 7-point checklist, publicly available, under the name *Derm7pt*. Shortly after the Interactive Atlas of Dermoscopy, Menzies et al. [239] published *An Atlas of Surface Microscopy of Pigmented Skin Lesions: Dermoscopy*, with a CD-ROM containing 217 dermoscopic images (39% melanomas, 7% carcinomas, 54% nevi) of 712×454 pixels, acquired by the Sydney Melanoma Unit, Australia.

The PH^2 dataset, released by Mendonca et al. [237] and detailed by their extended work [238], was the first public dataset to provide region-based annotations, with segmentation masks, and masks for clinically suggestive colors (white, red, light brown, dark brown, blue-gray, and black) present in the images. The dataset has 200 dermoscopic images (20% melanomas, 40% atypical nevi, and 40% common nevi) of 768×560 pixels, acquired at the Hospital Pedro Hispano, Portugal. The Edinburgh DermoFit Image Library [33] also

provides region-based annotations for 1300 clinical images (10 diagnostic classes including melanomas, seborrhoeic keratosis, and basal cell carcinoma) of sizes ranging from 177×189 to 2176×2549 pixels. The images were taken with a Canon EOS 350D SLR camera, in controlled lighting, and consistent distance from the lesions, resulting in a high level of quality, atypical for clinical images.

The ISIC Archive contains the world's largest curated repository of dermoscopic images. ISIC, an international academia-industry partnership sponsored by ISDIS (International Society for Digital Imaging of the Skin), aims to "facilitate the application of digital skin imaging to help reduce melanoma mortality" [1]. At the time of this writing, the archive contains more than $157,000$ images, of which almost $69,000$ are publicly available. Those images were acquired in leading worldwide clinical centers, using a variety of devices. Broad international participation intends to ensure a representative, clinically-relevant sample.

In addition to curating the datasets that collectively form the "ISIC Archive", ISIC has released standard archive subsets as part of its *Skin Lesion Analysis Towards Melanoma Detection* Challenge, organized annually since 2016. The 2016, 2017, and 2018 challenges comprised segmentation, feature extraction, and classification tasks, while the 2019 and 2020 challenges comprised only classification. Each subset is associated with a challenge (year), one or more tasks, and has two (train/test) or three (train/validation/test) splits.

The ISIC Challenge 2016 [144] (ISIC 2016, for short), is small in comparison to the following years, containing $1,279$ images split into 900 for training (19% melanomas, 81% nevi), and 379 for testing (20% melanomas, 80% nevi). There was a large variation in image size, from 0.5 until 12 megapixels. All tasks used the same images. The ISIC 2017 [89] dataset more than doubled, with $2,750$ images split into $2,000$ for training (18.7% melanomas, 12.7% seborrheic keratoses, 68.6% nevi), 150 for validation (20% melanomas, 28% seborrheic keratoses, 52% nevi), and 600 for testing (19.5% melanomas, 15% seborrheic keratoses, 65.5% nevi). Again, image size varied markedly, from 0.5 to 29 megapixels, and all tasks used the same images.

ISIC 2018 provided for the first time separate datasets for the tasks, with $2,594/100/1,000$ train/validation/test images (diagnostic distribution unknown), ranging from 0.5 to 29 megapixels, for the tasks of segmentation and feature extraction [88], and $10,015/1,512$ train/test images for the classification task, all with 600×450 pixels. The train dataset for classification was the HAM10000 dataset [333], acquired over a period of 20 years at the Medical University of Vienna, Austria and the private practice of Dr. Cliff Rosendahl, Australia. It allowed a five-fold increase in training images in comparison to 2017 and comprised seven diagnostic classes: melanoma (11.1%), nevus (66.9%), basal cell carcinoma (5.1%), actinic keratosis or Bowen's disease (3.3%), benign keratosis (solar lentigo, seborrheic keratosis, or lichen planus-like keratosis, 11%), dermatofibroma (1.1%), and vascular lesion (1.4%).

ISIC 2019 [89, 333, 92] contained 25,331 train images (18% melanomas, 51% nevi, 13% basal cell carcinomas, 3.5% actinic keratoses, 10% benign keratoses, 1% dermatofibromas, 1% vascular lesions, and 2.5% squamous cell carcinomas) and 8,238 test images (diagnostic distribution unknown). All image sizes range from 600×450 to $1,024 \times 1,024$ pixels.

ISIC 2020 [289] grew to 33,126 training images (1.8% melanomas, 97.6% nevi, 0.4% seborrheic keratoses, 0.1% lentigines simplex, 0.1% lichenoid keratoses, 0.02% solar lentigines, 0.003% cafe-au-lait macules, 0.003% atypical melanocytic proliferations) and 10,982 test images (diagnostic distribution unknown), ranging from 0.5 to 24 megapixels. Multiple centers, distributed worldwide, contributed to the dataset, including the Memorial Sloan Kettering Cancer Center, in USA, the Melanoma Institute, the Sydney Melanoma Diagnostic Centre, and the University of Queensland, in Australia, the Medical University of Vienna, in Austria, the University of Athens, in Greece, and the Hospital Clinic Barcelona, in Spain. An important novelty in this dataset is the presence of multiple lesions per patient, with the express motivation of exploiting intra- and inter-patient lesion patterns, e.g., the so-called "ugly-ducklings", lesions whose appearance is atypical for a given patient, and which present an increased risk of malignancy [120].

Biases in Computer Vision datasets are a constant source of issues [331], which is compounded for medical images due to the small of samples, the low resolution of images, lack of geographical or ethnic diversity, or statistics (including diagnostic statistics) unrepresentative of clinical practice. All existing skin-lesion datasets suffer in a higher or lesser degree from one or more of those issues, to which we add the specific issue of the availability and reliability of annotations. For lesion diagnostic, many samples lack histopathological diagnostic confirmation (the goldstandard), and the ground truth segmentation, even when available, is inherently noisy (subsection 2.4.2). The presence of artifacts (Fig. 2.1) may lead to spurious correlations, an issue that Bissotoet al. [55] attempted to quantify for diagnostic models. Table 2.1 shows a list of publicly available skin-lesion datasets with pixel-wise annotations, image modality, sample size, original split sizes, and diagnostic label distribution. Fig. 2.2 showcases how frequently those datasets appear in the literature.

2.2.2 Synthetic Data Generation

Data augmentation — synthesizing new samples from existing ones — is widespread for training deep-learning models. Augmented train samples serve as a regularizer, increase the amount and diversity of data [309], induce desirable invariances on the model, and may alleviate class imbalance.

Traditional data augmentation applies simple geometric, photometric, and colorimetric transformations on the samples, including mirroring, translation, scaling, rotation, cropping, random region erasing, affine or elastic deformation, modifications of hue, saturation, brightness, and contrast. Usually, several transformations are chosen at random and combined.

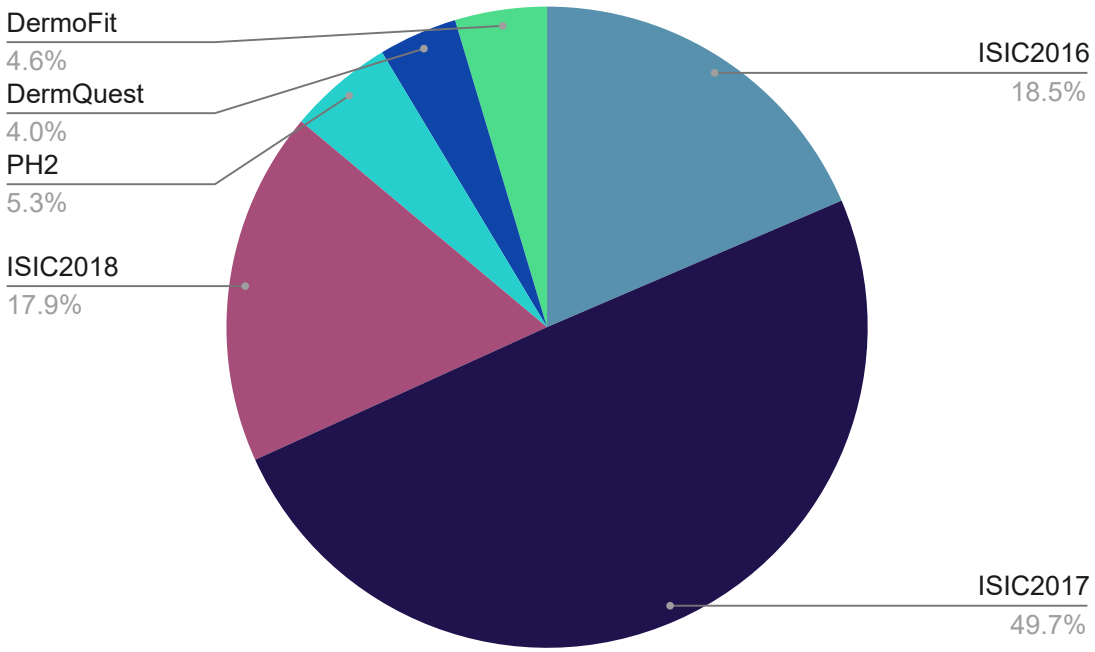


Figure 2.2: The frequency of different skin lesion segmentation datasets utilization in the evaluation of surveyed studies.

Fig. 2.3 exemplifies the procedure, as applied to a dermoscopic image with Albumentations [64], a state-of-the-art open-source library for image augmentation.

As mentioned, augmented train data induce invariance on the models: random translations/croppings, for example, help the model to be translation-invariant. That has implications for skin-lesion analysis, e.g., data augmentation for generalist datasets (such as ImageNet) forgo vertical mirroring and large-angle rotations, because natural scenes have a strong vertical anisotropy, while skin-lesion images are isotropic.

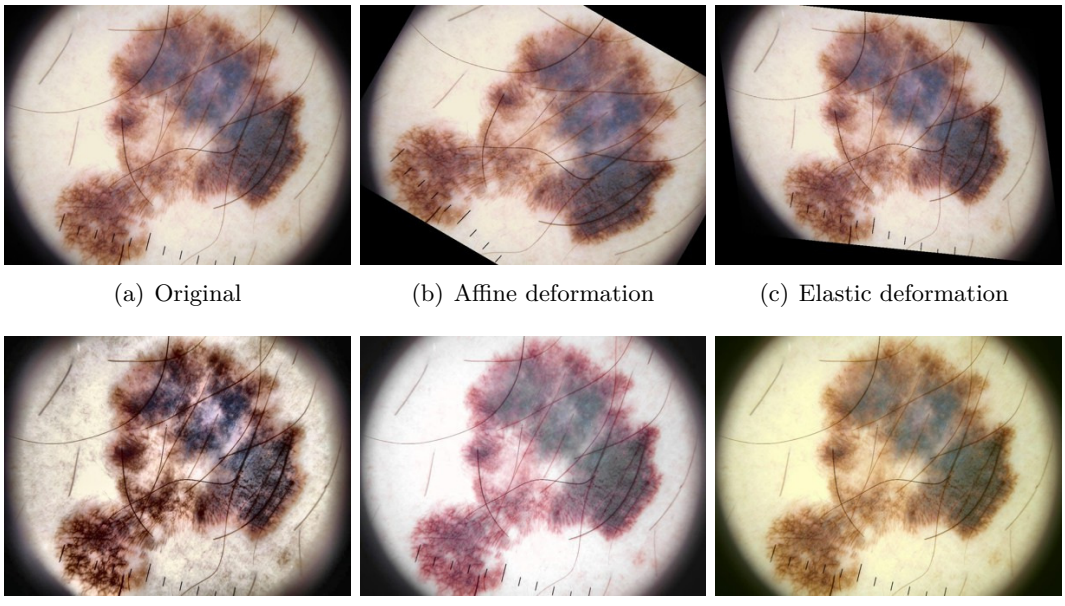
Augmented *test* data (test-time augmentation) also improves generalization by combining the predictions of several augmented samples through, e.g., average pooling or majority voting [309].

Perez et al. [265] have systematically evaluated the effect of several data-augmentation schemes for skin-lesion diagnostic, finding that the use of both train and test augmentation is critical for performance, surpassing at times increases of real data without augmentation. Valle et al. [340] found, in a very large-scale experiment, that test-time augmentation was the second most influential factor for diagnostic performance, after training set size. No systematic study of this kind exists for segmentation.

Although traditional data augmentation is crucial for deep-learning models, it falls short of providing samples at once diverse and plausibly from the same distribution as real data. Thus, modern data augmentation [322] employs generative modeling, learning the probabil-

Table 2.1: Public skin lesion datasets with segmentation annotations.

dataset	year	modality	size	train/val./test	class distribution	additional info
DermQuest [2]	2012	clinical	137	-	61 non-melanoma 76 melanoma	8-bit RGB images taken with different lighting and cameras
DermoFit [33]	2013	clinical	1300	-	1224 non-melanoma 76 melanoma	8-bit RGB images of sizes ranging from 177×189 to 2176×2549 pixels captured at a controlled lighting situation and the same distance
Pedro Hispano Hospital (PH^2) [237]	2013	Dermoscopy	200	-	160 benign nevi 40 melanoma	8-bit RGB images of sizes 553×763 to 577×769 pixels acquired at 20× magnification.
ISIC2016 [144]	2016	Dermoscopy	1279	900/-/379	Train: 727 non-melanoma 173 melanoma Test: 304 non-melanoma 75 melanoma	8-bit RGB images of sizes ranging from 566×679 to 2848×4288 pixels.
ISIC2017 [89]	2017	Dermoscopy	2750	2000/150/600	Train: 1626 non-melanoma 374 melanoma Test: 483 non-melanoma 117 melanoma	8-bit RGB images of sizes ranging from 540×722 to 4499×6748 pixels.
ISIC2018 [88]	2018	Dermoscopy	3694	2594/100/1000	-	8-bit RGB images of sizes ranging from 540×576 to 4499×6748 pixels.



(a) Original (b) Affine deformation (c) Elastic deformation

(d) Histogram equalization (e) HSV shift (f) RGB shift

Figure 2.3: Various data augmentation transformations (image source: ISIC 2016 image set [144])

ity distribution of the real data, and sampling from that distribution. Generative Adversarial

Networks (GANs) [132] are the most promising approach in that direction [309], especially for medical image analysis [371, 191, 306].

GANs employ an adversarial training between a generator, which attempts to generate realistic fake samples, and a discriminator, which attempts to differentiate real from fake samples. When the procedure converges, the generator output is surprisingly convincing, but GANs are computationally expensive and difficult to train [95].

Synthetic generation of skin lesions has received some recent interest, especially in the context of improving diagnostics. Works can be roughly divided into those that use GANs that create new images from a Gaussian latent variable [38, 269, 6], and those that implement GANs based on image-to-image translation [7, 56, 108].

Noise-based GANs, such as DCGAN [374], LAPGAN [104], and PGAN [182], learn to decode a Gaussian latent variable into an image that belongs to the distribution found in the training set. The main advantage of those techniques is the ability to create more, and more diverse images, as, in principle, any sample from a multivariate Gaussian distribution may become a different image. The disadvantage is that the images tend to be lower-quality, and, in the case of segmentation, there is the need to generate plausible pairs of images and segmentation masks.

Image-to-image translation GANs, such as pix2pix [161] and pix2pixHD [353], learn to create new samples from a semantic segmentation map. They have complementary advantages and disadvantages. Because the procedure is deterministic (one map creates one image), they have much less freedom in the number of samples available, but the images tend to be higher-quality, more "plausible". There is no need to generate separate segmentation maps because the generated image is intrinsically compatible with the input segmentation map.

Published simultaneously, the two seminal articles on GANs for skin lesions [38, 56] evaluate several models. Baur et al. [38] compare the noise-based DCGAN, LAPGAN, and PGAN for the generation of 256×256-pixel images using both qualitative and quantitative criteria, finding that the PGAN had considerably better results. They further examined the PGAN against a panel of human judges, composed by dermatologists and deep-learning experts, in a "visual Turing test", showing that both had difficulties in recognizing the fake from the true images. Bissoto et al. [56] adapt the PGAN to be class-conditioned on skin-lesion diagnostic, and the image-to-image pix2pixHD to employ the semantic annotation provided by the feature extraction task of the ISIC 2018 dataset (Section 1.1), comparing those to an unmodified DCGAN on 256×256-pixel images, and finding the adapted pix2pixHD qualitatively better. They use the improvement of a separate classification network as a quantitative metric, finding that the use of samples from both PGAN and pix2pixHD to bring the best improvements. They also showcase up to $1,024 \times 1,024$-pixel images on the pix2pixHD-derived model.

Pollastri et al. [269], comparing DCGAN and LAPGAN, extended both architectures to generate the segmentation masks (in the pairwise scheme explained above), making their work the only noise-based GANs usable for segmentation of which we are aware. Bi et al. [44] introduced stacked adversarial learning to learn class-specific GANs generating skin images given the ground truth segmentations. Abhishek et al. [7] employ pix2pix to translate a binary segmentation mask into a dermoscopic image.

Ding et al. [108] feed to the generator a segmentation mask and an instance mask stating the diagnostic to be synthesized. In both cases, the discriminator receives different resolutions of the generated image, being required to make a decision for each of them. Abdelhalim et al. [6] is a recent work that also conditions PGAN on the class label.

Recently, Bissoto et al. [57] cast doubt on the power of GAN-synthesized data augmentation to reliably improve lesion diagnostic results. Their evaluation, which included four GAN models, four datasets, and several augmentation scenarios, showed improvement only on a severe cross-modality scenario (training on dermoscopic and testing on clinical images). As far as we know, no corresponding systematic evaluation exists for image segmentation.

2.2.3 Supervised, Semi-supervised, Weakly Supervised, Self-supervised Learning

Although supervised deep learning has achieved striking performance for medical images, its strict dependency on high-quality annotations limits its applicability, as well as its generalization to unseen, out-of-distribution data.

Since the pixel-level annotation of skin images is costly, there is a trade-off between annotation precision/accuracy and efficiency. In practice, the annotations are intrinsically noisy, which can be modeled explicitly to avoid over-fitting. (We discuss the issue of annotation variability in detail in subsection 2.4.2.)

Semi-supervised techniques attempt to learn from both labeled and unlabeled samples. Weakly supervised techniques attempt to exploit partial annotations like image-level labels or bounding boxes, often in association with a subset of pixel-level fully-annotated samples.

To remove the dependency on having a set of perfectly clean annotations, Redekop et al. [279] propose to alter noisy ground truth masks during training by considering the quantification of aleatoric uncertainty [205] to obtain a map of regions of high and low uncertainty. Pixels of ground truth masks in highly uncertain regions are flipped, progressively increasing the model's robustness to label noise. Ribeiro et al. [284] deal with noise by discarding inconsistent samples and annotation detail during training time, showing that the model generalizes better, even when detailed annotations are required in test time.

When no labeled images are available for training, Kamalakannan et al. [180] propose to generate ground truth masks that enable the training of a deep neural network. Their method cluster similar pixels in the image and require a manual inspection to verify if foreground and background had been assigned coherent labels across images. When there

is a labeled set, even though the unlabeled one greatly outmatched it, semi- and self-supervision techniques can be applied. Li et al. [218] propose a semi-supervised approach, using a transformation-consistent self-ensemble to leverage unlabeled data in addition to labeled data and regularize the model. They minimized the difference between the network predictions of different transformations (random perturbations, flipping, and rotation) applied to the input image and the transformation of the model prediction for the input image. Self-supervision attempts to exploit intrinsic labels by solving proxy tasks, enabling the use of a large unlabeled corpus of data to pretrain a model before fine-tuning it to the target task. An example is to artificially apply random rotations in the input images, and train the model to predict the exact degree of rotation [199]. Note that the degree of rotation of each image is known (since it was artificially applied), and thus, can be used as a label during training. Similarly, for skin lesion segmentation, Li et al. [220] propose to exploit the color distribution information, the proxy task being to predict values from blue and red color channels while having the green one as input. They also include a task to estimate the red and blue color distributions to improve the model's ability to extract global features. After the pretraining, they use a smaller set of labeled data to fine-tune the model.

2.2.4 Image Preprocessing

Preprocessing may facilitate the segmentation of skin-lesion images, including:

- **Downsampling**: Dermoscopy is typically a high-resolution technique, resulting in large image sizes, while many CNN architectures (e.g., LeNet, AlexNet, VGG, GoogLeNet, ResNet) require fixed-size input images, usually 224×224 or 299×299 pixels, and even those CNNs that can handle arbitrary-sized images (e.g., fully-convolutional networks, FCN) may benefit from downsampling for computational reasons. Downsampling is commonplace in segmentation literature [90, 373, 377, 16, 385, 269].

- **Color space transformations**: RGB images are expected by most models, but some works [90, 16, 378, 269, 271] employ transformed color spaces [63], such as CIELAB, CIELUV, and HSV. Often, one or more channels of the transformed space are combined to the RGB channels, is the hope of increasing the class separability, decoupling luminance and chromaticity, ensuring (approximate) perceptual uniformity, achieving invariance to illumination or viewpoint, or eliminating highlights.

- **Additional inputs**: Apart from color space transformations, recent works have incorporated more focused and domain-specific inputs to the segmentation models, such as Fourier domain representation using discrete Fourier transform [328] and inputs based on the physics of skin illumination and imaging [9].

- **Contrast enhancement**: contrast deficit (Fig. 2.1(i)) is a prime reason for segmentation failures [58], leading some works [291] to enhance it prior to processing.

- **Color normalization**: varying illumination [34, 35] may lead to inconsistencies that some studies [135] attempt to eliminate with color normalization.
- **Artifact removal**: dermoscopic images often present artifacts, among which hair (Fig. 2.1(g)) is the most distracting [4], leading some studies [338, 379, 216] to attempt to pre-filter it out.

Classical machine-learning models (e.g., nearest neighbors, decision trees, support vector machines [72, 71, 162, 37, 308]), which rely on hand-crafted features [36], tend to benefit more from preprocessing than deep learning models, which, when properly trained, tend to from the data how to bypass input issues [68, 340]. Preprocessing may still be helpful when dealing with noisy small image sets.

2.3 Model Design and Training

Multi-Layer Perceptrons (MLPs) for pixel-level classification [127, 184] appeared soon after the dissemination of backpropagation [290], but those shallow feed-forward networks had many drawbacks [211], including an excess of parameters, lack of invariance, and disregard for the inherent structure present in images.

Convolutional Neural Networks (CNNs) addressed those issues by dramatically reducing the number of parameters through an aggressive, implicit parameter sharing provided by the convolution operation, which also fosters translation invariance and respects the image neighborhood structure. The resulting models require minimal preprocessing and automate feature engineering [41], transforming the raw pixels into progressively abstract features [210]. CNNs became the preferred architecture for most medical image tasks [224].

Semantic segmentation may be understood as the attempt to answer the parallel, complementary questions "what" and "where". The former is better answered by translation-invariant global features, while the latter requires well-localized features, posing a challenge to deep models. CNNs for pixel-level classification appeared since the mid-2000s [250], but their use accelerated after the seminal article on FCNs by Long et al. [228], which became the basis for many state-of-the-art segmentation models. In contrast to classification CNNs (e.g., LeNet, AlexNet, VGG, GoogLeNet, ResNet), FCNs easily cope with arbitrary-sized input images.

2.3.1 Architecture

The ideal skin-lesion segmentation is accurate, computationally cheap, frugal for training data, invariant to noise and input transformations, and easy to implement and train. Unfortunately, no actual technique has, so far, conciliated those conflicting goals. Deep learning segmentation tends towards accuracy and invariance at the cost of computation and data. Ease of implementation is debatable: on the one hand, those techniques often forgo costly,

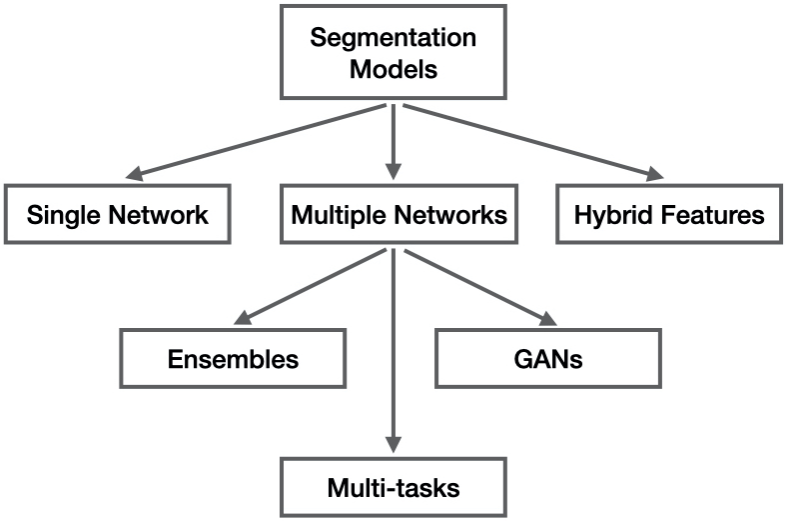


Figure 2.4: Taxonomoic organization of skin lesion DL segmentation methods.

delicate preprocessing, post-processing, and feature engineering. On the other hand, tuning and optimizing them is often a painstaking task,

We have classified existing literature into single-network models, multiple-network models, hybrid-feature models, which we discuss separately next. The first and second groups are somewhat self-descriptive, but notice that the latter is further divided into ensembles of models, multi-task methods (often simultaneous classification and segmentation), and GANs. Hybrid-feature models combine deep learning with hand-crafted interventions (Fig. 2.4). We classified works according to their most relevant feature, but the architectural improvements discussed in subsection 2.3.1 also appear in the models listed in the other sections.

Table 2.3 summarizes all works surveyed, separated by group.

Single Network Models

The approaches in this section employ a single DL model, usually a fully convolutional network, following an *encoder-decoder* structure, where the encoder extracts increasingly abstract features, and the decoder outputs the segmentation mask. In this section, we discuss those architectural choices for designing deep learning for skin-lesion segmentation.

Earlier works adopted either FCN [228] or U-Net [287]. FCN originally comprised a backbone of VGG16 [311] CNN layers in the encoder, and a single deconvolution layer in the encoder. The original paper proposes three versions, two with skip connections (FCN-8 and FCN-16), and one without them (FCN-32). U-Net [287], originally proposed for segmenting electron microscopy images, was rapidly adopted for medical applications. As its name suggests, it is a U-shaped model, with an encoder stacking convolutional layers that double

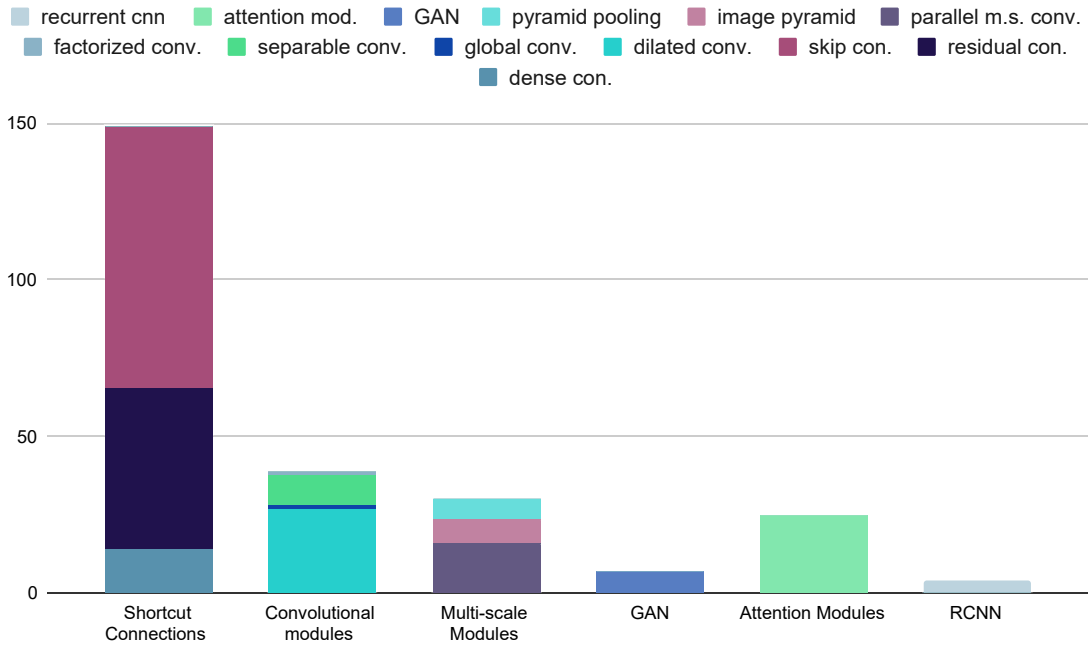


Figure 2.5: The frequency of utilization of architectural modules in surveyed studies.

in size filterwise, intercalated by pooling layers, and a symmetric decoder with pooling layers replaced by up-convolutions. Skip connections between corresponding encoder-decoder blocks improve the flow of information between layers, preserving low-level features lost during pooling and allowing detailed segmentation boundaries.

U-Net appears frequently in skin-lesion segmentation both in its original form [90, 269, 276], and in adapted models [326, 21, 149], discussed below. Some works introduce their own models [377, 16]). Fig. 2.5 plots how frequently different architectures appeared in our survey.

Shortcut Connections Connections between early and late layers in FCNs have been widely explored to improve both the forward and backward (gradient) information flow in the models, the latter easing the training. The three most popular types of connections are described below.

Residual connections: creating non-linear blocks that add their unmodified inputs to their outputs [150] alleviates gradient degradation in very deep networks. It provides direct path flow of the gradient to the early layers of the network, while still allowing for very deep models. The technique appears often in skin-lesion segmentation, in the implementation of the encoder [300, 30, 373] or both encoder and decoder [151, 344, 214, 335, 382, 152, 368]. Residual connections have also appeared in recurrent units [21, 20], dense blocks [314], chained pooling [151, 214, 152], and 1-D factorized convolutions [312].

Skip connections: appear in encoder-decoder architectures, connecting high-resolution features from the encoder's contracting path into semantic features on the decoder's expanding path [287]. Those connections allow to preserve localization, especially near region boundaries, and to combine multi-scale features, resulting in sharper boundaries in the predicted segmentation. Being at once effective and easy to implement makes skip connections very popular in skin-lesion segmentation [382, 30, 314, 362, 344, 26, 151, 21, 300, 380, 214, 335, 373, 312, 152, 368, 20, 346, 226].

Dense connections: expand the convolutional layers by connecting each layer to all its subsequent layers, concatenating their features [157]. Iterative reuse of features in dense connections maximizes information flow forward and backward, while avoiding additional parameters or computation. Similar to deep supervision (subsection 2.3.2), the gradient backpropagates directly through all previous layers. Several works [380, 314, 218, 335, 346] integrated dense blocks in both the encoder and the decoder. Baghersalimi et al. [30], Hasan et al. [149] and Wei et al. [362] used multiple dense blocks iteratively in just the encoder, while Li et al. [214] proposed dense deconvolutional blocks to reuse features from the previous layers. Azad et al. [26] encoded densely connected convolutions into the bottleneck of their encoder-decoder to obtain better features.

Convolutional Modules As mentioned, the convolution not only provides a structural advantage, respecting the local connectivity structure of images in the output futures, but also dramatically improves parameter sharing since the parameters of a relatively small convolutional kernel are shared by all patches of a large image.

Convolution is a critical element of deep segmentation models. In this section, we discuss some new variants, which have enhanced and diversified this operation, appearing in the skin-lesion segmentation literature.

Dilated convolution: In contrast to requiring full-resolution outputs in dense prediction networks, pooling and striding operations adopted in deep convolutional neural networks (DCNN) to increase the receptive field, diminish the spatial resolution of feature maps. Dilated or atrous convolutions are designed specifically for the semantic segmentation task to exponentially expand the receptive fields while keeping the number of parameters constant [372]. Dilated convolutions are convolutional modules with upsampled filters containing zeros between consecutive filter values. Sarker et al. [300] and Jiang et al. [174] utilized dilated residual blocks in the encoder to control the image field-of-view explicitly and incorporated multi-scale contextual information into the segmentation network. SkinNet [346] used dilated convolutions at the lower level of the network to enlarge the field of view and capture non-local information. Liu et al. [226] introduced dilated convolutions to the U-Net architecture, which significantly improved the segmentation performance. Also, different versions of the DeepLab architecture [79, 80, 81] which replace standard convolutions with dilated ones have been utilized in skin lesion segmentation tasks [135, 134, 97, 83, 66].

Separable convolution: Separable convolution or depth-wise separable convolution [86] is a spatial convolution operation that assigns a kernel to each input channel and convolves each input channel with its corresponding kernel. It is followed by a 1×1 standard convolution to capture the channel-wise dependencies in the output of depth-wise convolution. Depth-wise convolutions are designed to reduce the number of parameters and the computation of standard convolutions while keeping the same accuracy. DSNet [149] and separable-Unet [326] utilized depth-wise separable convolutions in the model to have a lightweight network with a reduced number of parameters. Adopted from the DeepLab architecture, Goyal et al. [135], Cui et al. [97] and, Canalini et al. [66] incorporated depth-wise separable convolutions in conjunction with dilated convolution to improve the speed and the accuracy of dense predictions.

Global convolution: State-of-the-art segmentation models remove densely connected and global pooling layers to preserve spatial information required for full resolution output recovery. As a result, segmentation models become optimal for localization and in contrast, sub-optimal for per-pixel classification which needs transformation invariant features. To increase the connectivity between feature maps and classifiers, large convolutional kernels should be adopted. However, they suffer from high computational costs and the number of parameters. To tackle this, global convolutional networks (GCN) modules adopt a combination of symmetric parallel convolutions in the form of $1 \times k + k \times 1$ and $k \times 1 + 1 \times k$ to cover a $k \times k$ area of feature maps [263]. SeGAN [368] employed GCN modules with large kernel size in the generator's decoder to reconstruct segmentation masks and in the discriminator architecture to optimally capture a larger receptive field.

Factorized convolution: Factorized convolutions [350] are designed to decrease the number of convolution filter parameters as well as its computation time through kernel decomposition when a high dimensional kernel is substituted with a sequence of lower-dimensional convolutions. Also, by adding non-linearity between composited kernels, the network's capacity may improve. FCA-Net [312] and MobileGAN [299] utilized residual 1-D factorized convolutions (a sequence of $k \times 1$ and $1 \times k$ convolutions with ReLU non-linearity) in their segmentation architecture.

Multi-scale Modules In FCNs, taking semantic context into account when assigning per-pixel labels leads to a more accurate prediction. Exploiting multi-scale contextual information, effectively combining them as well as encoding them in deep semantic segmentation have been widely explored.

Image Pyramid: RefineNet [151] and its extension [152], MSFCDN [380], FCA-Net [312], and Abraham et al. [11] adopted an image pyramid of multi-resolution skin images as inputs to their deep segmentation network architectures to extract multi-scale discriminative features. RefineNet [151, 152], FCA-Net [312] and Abraham et al. [11] applied convolutional blocks to different image resolutions in parallel to generate features which are then

up-sampled in order to fuse multi-scale feature maps. MSFCDN [380] gradually integrated multi-scale features extracted from the image pyramid into the encoder's down-sampling path. Also, Jafari et al. [168, 166] extracted multi-scale patches from clinical images to predict semantic labels and refine lesion boundaries by deploying local and global information. While aggregating the feature maps computed on various image scales improves the segmentation performance, it also increases the computational cost of the network.
Parallel multi-scale convolutions: Alternatively, given a single image resolution, multiple convolutional filters with different kernel sizes [382, 349, 169] or multiple dilated convolutions with different dilation rates [135, 134, 97, 83, 66] are adopted in parallel paths to extract multi-scale contextual features from images. DSM [382] integrated multi-scale convolutional blocks into the skip connections of an encoder-decoder structure to handle different lesion sizes. Wang et al. [349] utilized multi-scale convolutional branches in the bottleneck of an encoder-decoder architecture followed by attention modules to selectively aggregate extracted multi-scale features.
Pyramid pooling: Another way of incorporating multi-scale information into deep segmentation models is to integrate the pyramid pooling (PP) module in the network architecture [389]. PP fuses a hierarchy of features extracted from different sub-regions by adopting different sizes of parallel pooling kernels followed by up-sampling and concatenation to create the final feature maps. Sarker et al. [300] and Jahanifar et al. [169] utilized PP in the decoder to benefit from coarse to fine features extracted by different receptive fields from skin images.

Dilated convolutions and skip connections are also two other types of multi-scale information extraction, which are explained in subsections 2.3.1 and 2.3.1, respectively.

Attention Modules An explicit approach of exploiting contextual dependencies in the pixel-wise labeling task is the self-attention mechanism [155, 118]. Two types of attention modules capture global dependencies in spatial and channel dimensions by integrating features among all positions and channels, respectively. Wang et al. [349] and Sarker et al. [299] leveraged both spatial and channel attention modules to recalibrate the feature maps by looking into the features' similarity between pairs of positions or channels and updating each feature value by a weighted sum of all other features. Also, Singh et al. [312] utilized a channel attention block in the proposed factorized channel attention (FCA) blocks, which was used to investigate the correlation of different channel maps for extraction of relevant patterns. Inspired by attention U-Net [255], Wei et al. [362], Song et al. [314] and Abraham et al. [11] integrated a spatial attention gate in an encoder-decoder architecture to combine coarse semantic feature maps and fine localization feature maps. Kaul et al. [185] proposed FocusNet which utilizes squeeze and excitation (SE) blocks into a hybrid encoder-decoder architecture. SE blocks model the channel-wise inter-dependencies to re-weight feature maps and improve their representation power. Experimental results demonstrate that attention

modules improve the network focus on the lesions and suppress irrelevant feature responses in the background.

Recurrent Convolutional Neural Networks Recurrent convolutional neural networks (RCNN) integrate recurrent connections into convolutional layers by evolving the recurrent input over time [267]. Stacking recurrent convolutional layers (RCL) on top of the convolutional layer feature extractors ensures capturing spatial and contextual dependencies in images while limiting the network capacity by sharing the same set of parameters in RCL blocks. In the application of skin lesion segmentation, Attia et al. [25] utilized recurrent layers in the decoder to capture spatial dependencies between deep encoded features and recover segmentation maps at the original resolution. ∇^N-Net [20], RU-Net, and R2U-Net [21] incorporated RCL blocks into the FCN architecture to accumulate features across time in a computationally efficient way and boosted the skin lesion boundary detection. Azad et al. [26] deployed a non-linear combination of the encoder feature and decoder feature maps by adding a bi-convolutional LSTM (BConvLSTM) in skip connections. BConvLSTM consists of two independent ConvLSTMs which take the feature maps and process the data sequence in two backward and forward directions and make the final output based on the concatenation of their outputs. Modifications to the traditional pooling layers were also proposed, with the use of a dense pooling strategy [247].

Multiple Network Models

Motivations for models comprising more than one DL sub-model are diverse, ranging from alleviating the noise of the training procedure, exploiting a diversity of features learned by different models, and exploring synergies between multi-task learners. After assessing the literature (Fig. 2.4), we further classified the works in this section into standard ensembles and multi-task models. We also discuss generative adversarial models, which are intrinsically multi-network, in a separate category in this section.

Standard Ensembles Ensemble models are widely used in machine learning, motivated by the hope that the complementarity of different models may lead to more stable combined predictions [293]. Ensemble performance is contingent on the quality and diversity of component models, which can be combined at the featurelevel (early fusion) or the prediction level (late fusion). The former combines the features extracted by the components and learns a meta-model on them, while the latter pools or combines the models' predictions, with or without a meta-model.

All methods discussed in this section employ late fusion, except for an approach loosely related to early fusion [326], exploring various learning-rate decay schemes, and building a single model by averaging the weights learned at different epochs, to bypass poor local min-

imum during training. Since the weights correspond to features learned by the convolution filters, the approach can be interpreted as feature fusion.

Most works employ a single DL architecture with multiple trainings, varying configurations more or less during training [66]. The changes between component models may involve network hyperparameters (number of filters per block, and their size [90]), optimization/regularization hyperparameters (learning rate, weight decay [325]), the training set (multiple splits of a training set [377, 378], separate models per diagnostic label [49]), the preprocessing (different color spaces [269]), different pretraining strategy to initialize features extractors [66], or different approaches of network parameter initialization [97]. Test-time augmentation also may be seen as a form of inference-time ensembling [83, 226, 169], by combining the outputs of multiple augmented images to generate a more reliable prediction.

Bi et al. [49] trained a separate DL model for each diagnostic label, as well as a separate diagnostic classification model. For inference, the classification model output is used to weight the outputs of the category-specific segmentation networks. In contrast, Soudaniet al. [315] trained meta "recommender" model to dynamically choose, for each input, a segmentation technique from the top five scorers in the ISIC 2017 challenge, although their proposition was validated in a very small test set (10% of ISIC 2017 test set).

A couple of works ensemble different architectures [135, 360]. Goyal et al. [135] investigate multiple fusion approaches to bypass severe errors from individual models, comparing the average-, maximum- and minimum-pooling of their outputs. A usual assumption on ensemble is that the component models are trained independently, but Bi et al. [50] cascaded the component models, i.e., used the output of one model as the input of the next (in association with the actual image input). Thus, each model attempts to refine the segmentation already obtained by the previous one. They consider not only the final model output, but all the outputs in the cascade, making the technique a legitimate ensemble.

Multi-task Models Multi-task models jointly address more than one goal, in the hope that synergies among the tasks will improve overall performance [387], especially in the case of medical images, in which aggregating tasks may alleviate the issue of insufficient data or annotations. For skin lesions, few multi-task models exist [83, 222, 366, 177, 370], always exploiting the tasks of segmentation and (diagnostic) classification.

The synergy between tasks may appear when their models share common relevant features. Li et al. [222] assume that all features are shareable between the tasks and trains a single fully convolutional residual network to assign diagnostic category probabilities at a pixel-level. They use probability maps to estimate both lesion region and diagnostic category by weighted averaging of probabilities for different categories inside the lesion area. Yang et al. [370] learn an end-to-end model formed by a shared convolutional feature extractor followed by three task-specific branches to segment skin lesions, classify them as melanoma vs. non-melanoma, and classify them as seborrheic keratosis vs. non-seborrheic keratosis.

Chen et al. [83] do the same, but introduce a common latent layer between the feature extractor and the task heads, and a gate function that controls the flow of information between the tasks. The gate function work like...

Instead of using a single architecture for classification and segmentation, Xie et al. [366] and Jin et al. [177] use three CNNs in sequence to perform a coarse segmentation, followed by classification and, finally, a fine segmentation. Instead of shared features, those works exploit sequential guidance, in which the results of each task improve the learning of the next. While Xie et al. [366] feed the outputs of each network to the next, assuming that the classification network is a diagnostic category and a class activation map [392], Jin et al. [177] introduce feature entanglement modules, which establish relationships between features learned between subsequent networks.

All the multi-task models discussed so far have results suggesting complementarity between classification and segmentation, but there is no clear advantage among them. The segmentation of dermoscopic features (e.g., networks, globules, regression zones) combined with the other tasks is an explored avenue of improvement, which could bridge classification and segmentation, by fostering the extraction of features that "see" the lesion as human specialists do.

We do not consider in the hybrid group, two-stage models in which segmentation is used as ancillary preprocessing to classification [373, 90, 131, 18], since without mutual influence (sharing of losses or features) or feedback between the two tasks, there is no opportunity for synergy.

Vesal et al. [345] stressed the importance of object localization as an ancillary task for lesion delineation, in particular deploying Faster-RCNN to regress a bounding box to crop the lesions before training a SkinNet segmentation model. While that two-stage approach considerably improves results, it is computationally expensive. Goyal et al. [134] employed ROI detection with a deep extreme cut to extract the extreme points of lesions (leftmost, rightmost, topmost, bottommost pixels) and feed them (in a new auxiliary channel) to a segmentation model.

Generative Adversarial Models We discussed GANs for synthesizing new samples, their main use in skin-lesion analysis, in subsection 2.2.2, which also briefly explains their working principles. In this section, we are interested in GANs not for creating extra training samples, but for directly providing enhanced segmentation models. Adversarial training encourages high-order consistency in predicted segmentation by implicitly looking into the joint distribution of diagnostic labels and ground truth segmentation masks.

Peng et al. [264], Tu et al. [335], Lei et al. [213], and Izadi et al. [164] use a U-Net-like generator that takes a dermoscopic image as input, and outputs the corresponding segmentation, while the discriminator is a traditional CNN which attempts to discriminate pairs of image and generated segmentation from pairs of image and ground truth. The

generator has to learn to correctly segment the lesion in order to fool the discriminator. Jiang et al. [174] use the same scheme, with a dual discriminator. Lei et al. [213] also employ a second discriminator, but receiving only segmentations (unpaired from input images).

The discriminator may trivially learn to recognize the generated masks due to the presence of continuous probabilities, instead of the sharp discrete boundaries of the ground truths. Wei et al. [362] and Tu et al. [335] address this by pre-multiplying both generated and real segmentations by the (normalized) input images before feeding them to the discriminator.

We will further discuss adversarial loss functions in subsection 2.3.2.

Hybrid Feature Models

Although the major strength of CNNs is their ability to learn meaningful image features without human intervention, a few works tried to conciliate both worlds, with strategies ranging from employing pre- or post-processing to enforce prior knowledge until adding explicitly hand-crafted features.

CRFs use pixel-level color information models to refine the segmentation masks output by the CNN. While both Tschandl et al. [334] and Adegun et al. [14] consider a single CNN, Qiu et al. [273] combine the output of multiple CNNs into a single mask, before feeding it together with the input image to the CRFs. GrabCut [338] obtains the segmentation mask given the dermoscopy image and a region proposal obtained by the YOLO [280] network. Those methods regularize the CNN segmentation, which is mainly based on textural patterns, with expected priors based on the color of the pixels.

Works that combine hand-crafted with CNNs follow two distinct approaches. The first consists of pre-filtering the input images, in the hope to better contrast lesion from surrounding skin. Techniques explored include local-binary patterns (LBPs) [288, 171], wavelets [288], Laplacian pyramids [271], and Laplacian filtering [291]. The second consists of predicting an additional segmentation mask to combine with the one generated by the CNN. Zhang et al. [385], for example, use LBPs to consider the textural patterns of skin lesions and guide the networks towards more refined segmentations. Bozorgtabar et al. [62] also employ LBPs combined with pixel-level color information to divide the dermoscopic image into superpixels, which are then scored as part of the lesion or the background. That score mask is combined with the CNN output mask to compute the final segmentation mask.

Despite the limited number of works devoted to integrating deep features with hand-crafted ones, the results achieved so far indicate that this may be a promising direction of research.

2.3.2 Loss Functions

A segmentation model may be formalized as a function $\hat{y} = f_\theta(x)$, which maps an input image x into an estimated segmentation map $\hat{y}$, parameterized by a (large) set of parameters θ. For skin lesions, $\hat{y}$ is a binary mask separating lesion from surrounding healthy skin.

Training the model, given a training set of images x_i and their ground truth masks y_i $\{(x_i, y_i); i = 1, ..., N\}$, consists of finding the model parameters θ that maximizes the likelihood of observing those data:

$$\theta^* = \arg\max_\theta \sum_{i=1}^{N} \log \mathrm{P}(y_i|x_i;\theta), \tag{2.1}$$

which is performed indirectly, via the minimization of a loss function between the estimated and true segmentation masks:

$$\theta^* = \arg\min_\theta \sum_{i=1}^{N} \mathcal{L}(\hat{y}_i|y_i) = \arg\min_\theta \sum_{i=1}^{N} \mathcal{L}(f_\theta(x_i)|y_i). \tag{2.2}$$

The choice of the loss function is, thus, critical, as it encodes not only the main optimization objective, but much of the prior information needed to guide the learning and constraint the search space. Most skin-lesion segmentation models employ multiple losses to enhance generalization.

Losses based on p-norms

Losses based on p-norms are the simplest losses, and comprise the Mean Squared Error (MSE) (for $p = 2$) and the Mean Absolute Error (ℓ_1) (for $p = 1$).

$$MSE(X, Y; \theta) = -\sum_{i=1}^{N} \|y_i - \hat{y}_i\|_2, \tag{2.3}$$

$$l_1(X, Y; \theta) = -\sum_{i=1}^{N} \|y_i - \hat{y}_i\|_1. \tag{2.4}$$

In GANs to regularize the segmentations produced by generator, it is common to utilize hybrid losses containing MSE (ℓ_2 loss) [264] or the ℓ_1 loss [264, 335, 213]. The MSE has also been used as a regularizer to match attention and ground truth maps [365].

Cross entropy Loss

Semantic segmentation may be approached as classification at the pixel-level, i.e., as assigning a class label to each pixel. With that hypothesis, minimizing the negative log-likelihoods of pixel-wise predictions (i.e., maximizing their likelihoood) may be achieves by minimizing

a cross entropy loss $\mathcal{L}_{ce}$:

$$\mathcal{L}_{ce}(X, Y; \theta) = -\sum_{i=1}^{N} \sum_{p \in \Omega_i} y_{ip} \log \hat{y}_{ip} + (1 - y_{ip}) \log(1 - \hat{y}_{ip}), \quad \hat{y}_{ip} = P(y_{ip} = 1 | X(i); \theta), \quad (2.5)$$

where Ω_i is the set of all image i pixels, P is the probability, x_{ip} is p^{th} image pixel in i^{th} image and, $y_{ip} \in \{0, 1\}$ and $\hat{y}_{ip} \in [0, 1]$ are, in order, the true and predicted label of pixel p in image i.

Cross entropy loss appears in the majority of deep skin-lesion segmentation works (e.g., [314, 312, 382]). Since the gradient of the cross entropy loss function is inversely proportional to the predicted probabilities, hard-to-predict samples are weighted more to update the parameters, leading to faster convergence. A variant, weighted cross entropy, penalizes pixels and class labels differently. Nasr et al. [247] used pixel weights inversely proportional to their distance to lesion boundaries to enforce sharper boundaries. Class weighting may also mitigate the lesion/background pixel imbalance, which, left uncorrected, tend to bias models towards the latter, since lesions tend to occupy a relatively small portion of images. Goyal et al. [134], Chen et al. [83], and Wang et al. [354] apply that correction, using class weights inversely proportional to class pixel frequency.

All those losses, however, are independent pixel-wise, enforcing no spatial coherence, which motivates their combination with other, consistency-seeking losses.

Dice and Jaccard Loss

The Dice score and the Jaccard index are two popular metrics for segmentation evaluation (subsection 2.4.3), measuring the overlap between predicted segmentation and ground truth.

Models may employ differentiable approximations of those metrics known as soft Dice [151, 185, 152, 349] and soft Jaccard [344, 149, 299] to optimize an objective directly related to the evaluation metric.

For two classes, those losses are defined as follows:

$$\mathcal{L}_{dice}(X, Y; \theta) = 1 - \frac{1}{N} \sum_{i=1}^{N} \frac{2 \sum_{p \in \Omega} y_{ip} \hat{y}_{ip}}{\sum_{p \in \Omega} y_{ip} + \hat{y}_{ip}}, \quad (2.6)$$

$$\mathcal{L}_{jacc}(X, Y; \theta) = 1 - \frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{p \in \Omega} y_{ip} \hat{y}_{ip}}{\sum_{p \in \Omega} y_{ip} + \hat{y}_{ip} - y_{ip} \hat{y}_{ip}}. \quad (2.7)$$

Different variations of overlap-based loss functions account for class imbalance problem in medical image segmentation tasks. The Tanimoto distance loss, $\mathcal{L}_{td}$ is a modified Jaccard loss optimized in some models [66, 30, 377]:

$$\mathcal{L}_{td}(X, Y; \theta) = 1 - \frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{p \in \Omega} y_{ip} \hat{y}_{ip}}{\sum_{p \in \Omega} y_{ip}^2 + \hat{y}_{ip}^2 - y_{ip} \hat{y}_{ip}}, \quad (2.8)$$

which it is equivalent to the Jaccard loss when both y_{ip} and $\hat{y}_{ip}$ are binary.

The Tversky loss [11], inspired by the Tversky index, is another Jaccard variant, penalizing false positives and false negatives differently, to address the imbalance between the lesion and background pixels:

$$\mathcal{L}_{tv}(X, Y; \theta) = 1 - \frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{p \in \Omega} y_{ip} \hat{y}_{ip}}{\sum_{p \in \Omega} y_{ip} \hat{y}_{ip} + \alpha y_{ip}(1 - \hat{y}_{ip}) + \beta(1 - y_{ip})\hat{y}_{ip}}, \tag{2.9}$$

where α and β tune the contribution of false negative and false positive in $\mathcal{L}_{tv}$ and $\alpha+\beta = 1$.

Abraham. et al. [11] combined the Tvserky and the focal losses [223], the latter enforcing a focus on the most difficult pixels:

$$\mathcal{L}_{ftv} = \mathcal{L}_{tv}^{\frac{1}{\gamma}}, \tag{2.10}$$

where γ controls the relative importance of difficult samples.

Matthews Correlation Coefficient Loss

Matthews correlation coefficient (MCC) loss is a metric-based loss function based on the correlation between predicted and ground truth labels [8]. In contrast to overlap-based losses in subsection 2.3.2, MCC considers misclassifying the background pixels by penalizing false negative labels, making it more effective in the presence of skewed class distribution. MCC loss is defined as:

$$\mathcal{L}_{MCC}(X, Y; \theta) = 1 - \frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{p \in \Omega} \hat{y}_{ip} y_{ip} \frac{\sum_{p \in \Omega} \hat{y}_{ip} \sum_{p \in \Omega} y_{ip}}{M_i}}{f(\hat{y}_i y_i)}, \tag{2.11}$$

$$f(\hat{y}_i, y_i) = \sqrt{\sum_{p \in \Omega} \hat{y}_{ip} \sum_{p \in \Omega} y_{ip} - \frac{\sum_{p \in \Omega} \hat{y}_{ip} (\sum_{p \in \Omega} y_{ip})^2}{M_i} - \frac{(\sum_{p \in \Omega} \hat{y}_{ip})^2 \sum_{p \in \Omega} y_{ip}}{M_i} + (\frac{\sum_{p \in \Omega} \hat{y}_{ip} \sum_{p \in \Omega} y_{ip}}{M_i})^2}, \tag{2.12}$$

where M_i is the total number of pixels in image i.

Deep Supervision Loss

In deep learning models, the loss may apply not only to the final decision layer, but also to intermediate hidden layers. That supervision of hidden layers, known as deep supervision, guides the learning of intermediate features. Deep supervision also addresses the vanishing gradient problem, leading to faster convergence.

Deep supervision loss appears in several skin-lesion segmentation works [151, 380, 329, 214, 219, 382, 152], where it is computed in multiple layers, at different scales. The loss has the general form of a weighted summation of multi-scale segmentation losses:

$$\mathcal{L}_{ds}(X, Y; \theta) = \sum_{l=1}^{m} \gamma_l \mathcal{L}_l(X, Y; \theta), \tag{2.13}$$

where m is the number of scales, $\mathcal{L}_l$ is the loss at l^{th} scale, and γ_l adjusts the contribution of different losses.

Deep supervision improves segmentation by constraining the feature space. Notice that the auxiliary outputs are computed only during the training and discarded at inference time.

End-Point Error Loss

Most authors consider the lesion boundary the most challenging region segment. The end-point error loss [300, 312] underscores borders by using the first derivative of the segmentation masks instead of their raw values:

$$\mathcal{L}_{epe}(X, Y; \theta) = \sum_{i=1}^{N} \sum_{p \in \Omega} \sqrt{(\hat{y}_{ip}^0 - y_{ip}^0)^2 + (\hat{y}_{ip}^1 - y_{ip}^1)^2)}, \tag{2.14}$$

where $\hat{y}_{ip}^0$ and $\hat{y}_{ip}^1$ are the directional first derivatives of the estimated segmentation map in the x and y spatial directions, and, similarly, y_{ip}^0 and y_{ip}^1 for the ground truth derivatives. Thus, the loss function encourages the magnitude and orientation of edges of estimation and ground truth to match. The end-point loss attempts to use the relationship between adjacent pixel to solve vague boundaries in skin lesion segmentation.

Adversarial Loss

Another way to add high-order class-label consistency, adversarial training — with a discriminator attempting to distinguish estimated segmentation from ground truths — may be employed along with traditional supervised training. The objective will weight a pixel-wise loss $\mathcal{L}_s$ matching prediction to ground truth, and an adversarial loss, as follows:

$$\mathcal{L}_{adv}(X, Y; \theta, \theta_a) = \mathcal{L}_s(X, Y; \theta) - \lambda[\mathcal{L}_{ce}(Y, 1; \theta_a) + \mathcal{L}_{ce}(\hat{Y}, 0; \theta, \theta_a)], \tag{2.15}$$

where θ_a are the adversarial model parameters. The adversarial loss deploys a binary cross entropy loss to encourage the segmentation model to produce indistinguishable prediction maps from ground truth maps. Optimizing 2.15 is performed simultaneously in a mini-max game by minimizing 2.15 with respect to θ and maximizing it with respect to θ_a.

Pixel-wise losses, such as cross entropy [164, 312, 174], soft Jaccard [299, 335, 362], end-point error [335, 312], mean square error [264] and ℓ_1 loss [299, 312, 174] losses all have been incorporated in adversarial learning of skin-lesion segmentation. In addition, Tu et al. [335] and Xue et al. [368] presented a multi-scale adversarial term to match a hierarchy of local and global contextual features in the ground truth and predicted maps. In particular, they

minimize the mean absolute error of multi-scale features extracted from different layers of the adversarial model.

Rank Loss

Assuming that hard pixels makes larger prediction errors while training the model, rank loss [366] is proposed to encourage learning more discriminative information for harder pixels. The image pixels are ranked based on their prediction error, and the top K pixels with the largest prediction error from lesion or background areas are selected. Let $\hat{y}_{ij}^0$ and $\hat{y}_{il}^1$ are selected j^{th} hard pixel of background and l^{th} hard pixel of lesion in image i, we have:

$$\mathcal{L}_{rank}(X, Y; \theta) = \sum_{i=1}^{N} \sum_{j=1}^{K} \sum_{l=1}^{K} \max\{0, \hat{y}_{ij}^0 - \hat{y}_{il}^1 + margin\}, \tag{2.16}$$

which encourages $\hat{y}_{il}^1$ to be greater than $\hat{y}_{ij}^0$ plus margin.

Similar to rank loss, narrowband suppression loss [102] also adds a constraint between hard pixels of background and lesion. Different from rank loss, narrowband suppression loss collects pixels in a narrowband along the ground truth lesion boundary with radius r instead of all image pixels and then selects the top K pixels with a larger prediction error.

2.4 Evaluation

Evaluation is one of the main challenges for any image segmentation, skin-lesions included [74]. Segmentation evaluation may be subjective or objective [383], the former involving the visual assessment of results by a panel of human experts, and the latter involving the comparison of results with a ground truth segmentation using quantitative evaluation metrics. The subjective evaluation may provide a nuanced assessment of results, but because experts must grade each batch of results, it is usually too laborious to be applied, except in a very limited setting.

In objective assessment, experts are consulted once to provide the ground truth segmentation, and that knowledge can then be reused indefinitely, but due to intra- and inter-annotator variations, it raises the question of whether any individual ground truth segmentation reflects the ideal "true" segmentation, which we address in subsection 2.4.2. It also raises the issue of choosing one or more evaluation metrics (subsection 2.4.3).

2.4.1 Segmentation Annotation

Obtaining ground truth segmentations is paramount for the objective evaluation of methods using metrics. For synthetically generated images (subsection 2.2.2), ground truth augmentations may be known by construction, either by applying parallel transformations to the original ground truth masks in the case of traditional data augmentation, or by training generative models to synthesize images paired to their segmentation masks. For images

obtained from actual patients, however, human experts have to provide the segmentation. Different workflows have been proposed to conciliate the conflicting goals of ease of learning, speed, accuracy, and flexibility of annotation.

On one end of the spectrum, the expert traces the lesion by hand, on images of the skin lesion printed on photographic paper, which are then scanned [58]. The technique is easy to learn and fast, but the printing and scanning procedure limits the accuracy, and the physical nature of the annotations makes corrections burdensome.

On the other end of the spectrum, the annotation is performed on the computer, by a semi-automated procedure [88], with an initial border proposed by a segmentation algorithm, which is then refined by the expert using the annotation software, by adjusting the parameters of the segmentation algorithm manually. The technique is fast and easy to correct, but there might be a learning curve, and accuracy may depend on which algorithm is employed and how much the experts understand it.

By far, the commonest annotation method in the literature is somewhere in the middle, with fully manual annotations performed in a computer. The skin-lesion image file may be opened either in a raster graphics editor (e.g., GIMP or Adobe Photoshop), or in a dedicated annotation software [114], where the expert traces the borders of the lesion using a mouse or stylus, with continuous freehand drawing, or with discrete control points connecting line segments (resulting in a polygon [88]) or smooth curve segments (e.g., cubic B-splines [67]). That technique provides a good compromise, easy to implement, fast and accurate to perform, after an acceptable learning period for the annotator.

2.4.2 Inter-Annotator Agreement

Formally, dataset ground truths must be approached as samples of an estimator about the true label, which can never be directly observed [313]. That distinction is often immaterial for classification, when annotation noise is small. However, in medical image segmentation, ground truths suffer from both biases (systematic deviations from the "ideal") and significant noise [395, 76, 178, 137, 58, 207], the latter appearing as inter-annotator (different experts) and intra-annotator (same expert at different times) ground truth variability.

In the largest study of its kind to date, Fortina et al. [116] measured the inter-annotator variability among 12 dermatologists with varying levels of experience on a set of 77 dermoscopic images, showing that the average pairwise XOR dissimilarity (subsection 2.4.3) between annotators was $\sim 15\%$, and that in 10% of cases, that value was $> 28\%$. They found more agreement among more experienced dermatologists than less experienced ones. Also, more experienced dermatologists tend to outline tighter borders than less experienced ones. They suggest that the level of agreement among experienced dermatologists could serve as an upper bound for the accuracy achievable by a segmentation algorithm, i.e., if even highly experienced dermatologists disagree on how to classify 10% of an image, it might be

unreasonable to expect a segmentation algorithm to agree on more than 90% of any given ground truth on the same image [116].

Due to those issues, whenever possible, skin-lesion segmentation should be evaluated against multiple expert ground truths, a good algorithm being one that agrees with the ground truths at least as well as the expert agree among themselves [76]. Due to the cost of annotation, however, algorithms are often evaluated against a single ground truth.

When multiple ground truths are available, the critical issue is how to employ them. Several approaches have been proposed:

- Preferring one of the annotations (e.g., the one by the most experienced expert) and ignoring the others [67].
- Measuring and reporting the results for each annotator separately [71], which might require non-trivial multivariate analyses if the aim is to rank algorithms.
- Measuring each automated segmentation against all corresponding ground truths and reporting the average result [302].
- Measuring each automated segmentation against an *ensemble ground truth* formed by combining the corresponding ground truths pixel-wise using a bitwise OR [125, 123], bitwise AND [124], or a majority voting‘[163, 162, 251].

The ground truth ensembling operations can be generalized using a *thresholded probability map* [51]. First, all ground truths for a sample are averaged pixel-wise into a *probability map*. Then the map is binarized, with the lesion corresponding to pixels greater than or equal to a chosen threshold. The operations of OR, AND, and majority voting, correspond, respectively to thresholds of $1/n$, 1, and $(n - \varepsilon)/2n$, with n being the number of ground truths, and ε being a small positive constant. AND and OR correspond, respectively, to the tightest and loosest possible contours, with other thresholds leading to intermediate results. While the optimal threshold value is data-dependent, large thresholds focus the evaluation on unambiguous regions, leading to overly optimistic evaluations of segmentation quality [313, 207].

All approaches so far fail to consider the differences of expertise, experience, or performance of the annotators [359].

More elaborate ground truth fusion alternatives include shape averaging [286], border averaging [82, 76], binary label fusion algorithms such as STAPLE [359], TESD [51], and SIMPLE [208], as well as other more recent algorithms [260, 261, 262].

STAPLE (Simultaneous Truth And Performance Level Estimation) was very influential for medical image segmentation, inspiring many variants. For each image and its ground truth segmentations, STAPLE estimates a probabilistic true segmentation through the optimal combination of individual ground truths, weighting each one by the estimated sensitivity and specificity of its annotator. STAPLE may fail when there are few annotators or when

their performances vary too much [208, 207], a situation addressed by SIMPLE (Selective and Iterative Method for Performance Level Estimation) [208] by iteratively discarding poor quality ground truths.

Instead of attempting to conciliate multiple ground truths into a single one before employing conventional evaluation metrics, the latter may be adapted to take into account annotation variability.

Celebi et al. [73] proposed the *normalized probabilistic rand index* (NPRI) [337], a generalization of the *rand index* [277]. It penalizes segmentation results more (less) in regions where the ground truths agree (disagree). Fig. 2.6 illustrates the idea: ground truths outlined by three experienced dermatologists appear in red, green, and blue, while the automated result appears in black. NPRI does *not* penalize the automated segmentation in the upper part of the image, where the blue border seriously disagrees with the other two [73].

Despite many desirable qualities, NPRI has a subtle flaw: it is non-monotonic on the fraction of misclassified pixels [266]. Consequently, it measure might be unsuitable for comparing poor segmentation algorithms.

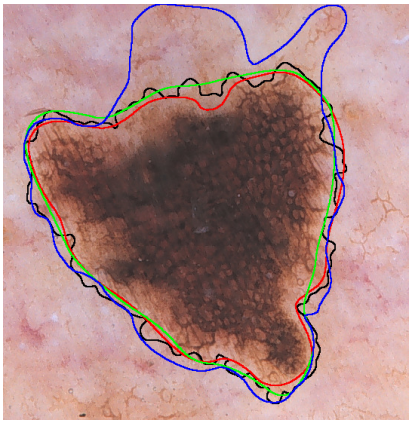


Figure 2.6: Sample border detection result.

2.4.3 Evaluation Metrics

We can frame the dermatological image segmentation problem as a binary pixel-wise classification task, where the positive and negative classes correspond to the lesion and the background skin, respectively.

Suppose that we have an input image and its corresponding segmentations: an *automated segmentation* (AS) produced by a segmentation algorithm and a *manual segmentation* (MS) outlined by a human expert. We can formulate a number of quantitative segmentation evaluation measures based on the concepts of *true positive*, *false negative*, *false positive*, and *true negative*, whose definitions are given in Table 2.2. In this table, actual and detected pixels refer to any given pixel in the MS and the corresponding pixel in the AS, respectively.

Table 2.2: Definitions of true positive, false negative, false positive, and true negative.

		Detected Pixel	
		Lesion (+)	Background (−)
Actual Pixel	Lesion (+)	True Positive	False Negative
	Background (−)	False Positive	True Negative

For a given pair of automated and manual segmentations, we can construct a 2×2 confusion matrix $\mathsf{C} = TPFN$
$FPTN$, where TP, FN, FP, and TN denote the numbers of true positives, false negatives, false positives, and true negatives, respectively. Clearly, we have $N = TP+FN+FP+TN$, where N is the number of pixels in either image. Based on these quantities, we can define a variety of scalar similarity measures to quantify the accuracy of segmentation [32, 170, 321]:

- Sensitivity (SE) $= \dfrac{TP}{TP+FN}$ & Specificity (SP) $= \dfrac{TN}{TN+FP}$
- Precision (PR) $= \dfrac{TP}{TP+FP}$ & Recall (RE) $= \dfrac{TP}{TP+FN}$
- Accuracy (AC) $= \dfrac{TP+TN}{TP+FN+FP+TN}$
- F-measure (F) $= \dfrac{2|AS \cap MS|}{|AS|+|MS|} = \dfrac{2 \cdot PR \cdot RE}{PR+RE} = \dfrac{2TP}{2TP+FP+FN}$ [341]
- G-mean (GM) $= \sqrt{SE \cdot SP}$ [203]
- Balanced Accuracy (BA) $= \dfrac{SE+SP}{2}$ [87]
- Jaccard index (J) $= \dfrac{|AS \cap MS|}{|AS \cup MS|} = \dfrac{TP}{TP+FN+FP}$ [165]
- Matthews Correlation Coefficient (MCC) $= \dfrac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$ [234]

For each similarity measure, the higher the value, the better the segmentation. Except for MCC, all of these measures have a unit range, that is, $[0, 1]$. The $[-1, 1]$ range of MCC can be mapped to $[0, 1]$ by adding one to it and then dividing by two. Each of these unit-range similarity measures can then be converted to a unit-range dissimilarity measure by subtracting it from one. Note that there are also dissimilarity measures with no corresponding similarity formulation. A prime example is the well-known XOR measure [147] defined as follows:

$$\text{XOR} = \frac{|AS \oplus MS|}{|MS|} = \frac{|\,(AS \cup MS) - (AS \cap MS)\,|}{|MS|} = \frac{FP+FN}{TP+FN}. \tag{2.17}$$

It is essential to notice that different evaluation measures capture different aspects of a segmentation algorithm's performance on a given image set, and thus there is no universally applicable evaluation measure [170].

This is why most studies employ multiple evaluation measures in an effort to perform a comprehensive performance evaluation. Such a strategy, however, complicates algorithm comparisons, unless one algorithm completely dominates the others with respect to all adopted evaluation measures.

Based on their observation that experts tend to avoid missing parts of the lesion in their manual borders, [125] argue that true positives have the highest importance in the segmentation of dermatological images. The authors also assert that false positives (background pixels incorrectly identified as part of the lesion) are less important than false negatives (lesion pixels incorrectly identified as part of the background). Accordingly, they assign a weight of 1.5 to TP to signify its overall importance. Furthermore, in measures that involve both FN and FP (e.g., AC, F, and XOR), they assign a weight of 0.5 to FP to emphasize its importance over FN. Using these weights, they construct a *weighted performance index*, which is an arithmetic average of six commonly used measures, namely SE, SP, PR, AC, F, and (unit complement of) XOR. This scalar evaluation measure facilitates comparisons among algorithms.

In a follow up study, Garnavi et al. [123] parameterize the weights of TP, FN, FP, and TN in their weighted performance index and then use a constrained nonlinear program to determine the optimal weights. They conduct experiments with five segmentation algorithms on 55 dermoscopic images. They conclude that the optimized weights not only lead to automated algorithms that are more accurate against manual segmentations, but also diminish the differences among those algorithms.

- Historically, AC has been the most popular evaluation measure owing to its simple and intuitive formulation. However, this measure tends to favor the majority class, leading to overly optimistic performance estimates in class-imbalanced domains. This drawback prompted the development of more elaborate performance evaluation measures, including GM, BA, and MCC.

- SE and SP are especially popular in medical domains. SE (aka *True Positive Rate*) quantifies the accuracy on the positive class, whereas SP (aka *True Negative Rate*) quantifies the accuracy on the negative class. These measures are generally used together because it is otherwise trivial to maximize one at the expense of the other (an automated border enclosing the corresponding manual border will attain a perfect SE, whereas in the opposite case, we will have a perfect SP). Unlike AC, they are suitable for class-imbalanced domains. BA and GM combine these measures into a single evaluation measure through arithmetic and geometric averaging, respectively. Unlike AC, these composite measures are suitable for class-imbalanced domains [231].

- PR is the proportion of examples assigned to the positive class that actually belongs to the positive class. RE is equivalent to SE. PR and RE are typically used in information retrieval applications, where the focus is solely on relevant documents (positive class). F combines these measures into a single evaluation measure through harmonic averaging. This composite measure, however, is unsuitable for class-imbalanced domains [397, 84, 231].

- MCC is equivalent to the *phi coefficient*, which is simply the *Pearson correlation coefficient* applied to binary data [84]. MCC values fall within the range of $[-1, 1]$ with -1 and 1 indicating perfect misclassification and perfect classification, respectively, while 0 indicating a classification no better than random [234]. Although it is biased to a certain extent [231, 394], this measure appears to be suitable for class-imbalanced domains [60, 84, 231].

- J (aka *Intersection over Union*) and F (aka *Dice coefficient* [106]) are highly popular in medical image segmentation [96]. These measures are monotonically related as follows: $J = F/(2 - F)$ and $F = 2J/(1 + J)$. Thus, it makes little sense to use them together. There are two major differences between these measures: [(i)]

- $(1 - J)$ is a proper distance metric, whereas $(1 - F)$ is *not* (it violates the triangle inequality).

- It can be shown [395] that if TN is sufficiently large compared to TP, FN, and FP, which is common in dermatological image segmentation, F becomes equivalent to *Cohen's kappa* [91], which is a chance-corrected measure of inter-observer agreement.

- Among the seven composite evaluation measures given above, AC, GM, BA, and MCC are symmetric, that is, invariant to class swapping, while F, J, and XOR are asymmetric.

- XOR is similar to *False Negative Rate*, that is, the unit complement of SE, with the exception that XOR has an extra additive TN term in its numerator. While XOR values are guaranteed to be nonnegative, they do *not* have a fixed upper bound, which makes aggregations of this measure difficult. XOR is also biased against small lesions [73]. Nevertheless, owing to its intuitive formulation, XOR was popular in dermatological image segmentation until about 2015 [74].

- The 2016 ISIC Challenge [144] adopted five measures: AC, SE, SP, F, and J, with the participants ranked based on the last measure. The 2018 ISIC Challenge [89] featured a *thresholded Jaccard index*, which returns the same value as the original J if the value is greater than or equal to a predefined threshold and zero otherwise. Essentially, this modified index considers automated segmentations yielding J values below the threshold as complete failures. The organizers of the challenge set the threshold equal to

0.65 based on an earlier study [90] that determined the average pairwise J similarities among the manual segmentations outlined by three expert dermatologists.

- Some of the aforementioned measures (*i.e.*, GM and BA) have *not* been used in a dermatologica image segmentation study yet.

- The evaluation measures discussed above are all region-based and thus fairly insensitive to border irregularities [212], that is, indentations and protrusions along the border. Boundary-based evaluation measures [321] have *not* been used in the dermatological image analysis literature much except the symmetric Hausdorff metric [310], which is known to be sensitive to noise [159] and biased in favor of small lesions [58].

2.5 Discussion and Future Research

Although various techniques like regularizing the parameters search space through multiple loss functions, multi-task learning, adversarial training, and synthetic data generation are integrated into training deep neural networks to overcome the problem of insufficient annotated training data, the small size of test datasets questions the actual generalization capability of deep models. Leveraging large-scale weakly and non-annotated skin images acquired from various imaging devices and environments alleviate the problem of over-fitting to limited training data and convergence of parameters to local minima but approximately all the current state-of-the-art deep skin lesion models heavily depend on densely supervised data.

Further, the laborious nature of pixel-wise annotations as well as ambiguous boundaries affect the quality of ground truth annotations. On one side looking into the largest skin lesion images dataset in the ISIC archive confirms that even experts may disagree substantially in delineating a common skin lesion. On the other side, most of the deep skin lesion segmentation models are designed based on the assumption that perfect annotations are available (see Fig. 2.7). Working on the deep models which are capable to aggregate multiple image annotations and handle inconsistent ground truth pixel labels is a valuable research direction toward the real-life problem of leveraging imperfect annotations in learning.

Another limitation is insufficient benchmark clinical skin lesion dataset with expert pixel-level annotations. Fig. 2.8 shows while the number of dermoscopy images with ground truth segmentation masks is increasing over the last few years, a few clinical data are available. In contrast to dermoscopy images requiring a special tool that is not always utilized even by dermathologists [112], clinical images captured by a digital camera and smartphones have the advantage of easy accessibility which can be utilized to evaluate the priority of patients by their lesion severity level. Most of the deep skin lesion segmentation models are

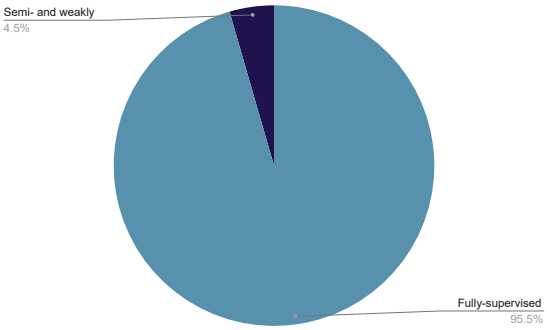


Figure 2.7: Percentage of supervised studies vs. semi-supervised studies.

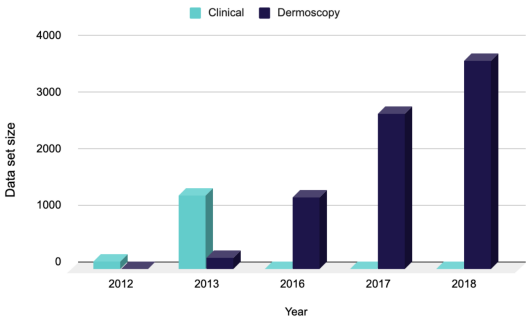


Figure 2.8: Number of skin lesion images with ground truth segmentation maps per year categorized based on modalities.

performed on dermoscopy images, leaving the need for the development of automatic tools for non-specialists unattended.

Table 2.3: Deep learning models for skin lesion segmentation task. Performance is the Jaccard index reported on the bold dataset. The score is asterisked if it is computed based on the reported Dice index. The following abbreviations are used: Ref.: reference, Arch.: architecture, Seg.: segmentation, Perf.: Jaccard performance, C.D. : cross-data evaluation. the highlighted dataset and P.P.: post-processing, con.: connection and conv.: convolution, CE: cross entropy, WCE: weighted cross entropy, DS: deep supervision, EPE: end point error, L_1: L_1 norm, L_2: L_2 norm and ADV: adversarial loss.

Ref.	Venue	Data	Arch. modules	Seg. loss	Perf.	C.D.	Augmentation	P.P.	code
[345]	peer-reviewed conference	**ISIC2017** PH^2	dilated conv. dense con. skip con.	Dice	88.00%	✓	-	✗	✗
[135]	peer-reviewed journal	**ISIC2017** PH^2	dilated conv. parallel m.s. conv. separable conv.	-	79.34%	✓	-	✓	✗
[151]	peer-reviewed conference	ISIC2016 **ISIC2017**	residual con. skip con. image pyramid	Dice CE DS	75.80%	✗	rotation	✓	✗
[344]	peer-reviewed conference	ISIC2017	residual con. skip con.	Jaccard	76.40%	✗	rotation,flipping translation, scaling	✓	✗
[26]	peer-reviewed conference	ISIC2018	skip con. dense con. recurrent cnn	CE	74.00%	✗	-	✗	✓
[21]	peer-reviewed journal	ISIC2017	skip con. residual con. recurrent cnn	CE	75.68%	✗	-	✗	✗
[378]	peer-reviewed journal	ISIC2017	-	Tanimoto	76.50%	✗	rotation,flipping shifting, scaling random normaliz.	✓	✗

Table 2.3

Ref.	Venue	Data	Arch. modules	Seg. loss	Perf.	C.D.	Augmentation	P.P.	code
[134]	peer-reviewed conference	**ISIC2017** PH^2	dilated conv. parallel m.s. conv.	WCE	82.20%	✓	-	✗	✗
[370]	non peer-reviewed technical report	ISIC2017	skip con. parallel m.s. conv.	-	74.10%	✗	rotation,flipping	✗	✗
[300]	peer-reviewed conference	ISIC2016 **ISIC2017**	skip con. residual con. dilated conv. pyramid pooling	CE EPE	78.20%	✗	rotation,scaling	✗	✓
[16]	peer-reviewed journal	**ISIC2017** PH^2	-	CE	77.10%	✓	rotation	✗	✗
[219]	peer-reviewed conference	ISIC2017	skip con. residual con.	DS	77.23%	✗	flipping, rotation	✗	✓
[49]	peer-reviewed journal	ISIC2016 **ISIC2017** PH^2	skip con. residual con.	CE	77.73%	✓	flipping, cropping	✓	✗
[62]	peer-reviewed journal	ISIC2016	-	-	80.60%	✗	rotation	✗	✗
[334]	peer-reviewed journal	ISIC2017	skip con.	CE Jaccard	76.80%	✗	flipping, rotation	✓	✗
[380]	peer-reviewed conference	ISIC2017	dense con. skip con. image pyramid	CE ℓ_2 DS	78.50%	✗	flipping, rotation	✓	✗

Table 2.3

Ref.	Venue	Data	Arch. modules	Seg. loss	Perf.	C.D.	Augmentation	P.P.	code
[218]	peer-reviewed journal	ISIC2017	skip con. dense con.	CE ℓ_1	79.80%	✗	flipping,rotating scaling	✓	✗
[105]	non peer-reviewed technical report	ISIC2017	skip con.	CE	73.00%	✗	flipping, rotation	✗	✗
[385]	peer-reviewed journal	ISIC2016 **ISIC2017**	skip con.	CE	72.94%	✗	-	✗	✗
[30]	peer-reviewed journal	ISIC2016 **ISIC2017** PH^2	skip con. residual con. dense con.	Tanimoto	78.30%	✓	flipping,cropping	✗	✗
[149]	peer-reviewed journal	**ISIC2017** PH^2	skip con. dense con. separable conv.	CE Jaccard	77.50%	✓	rotation, zooming shifting, flipping	✗	✓
[164]	peer-reviewed conference	DermoFit	skip con.	CE ADV	81.20%	✗	flipping, rotation elastic deformation	✗	✓
[174]	peer-reviewed conference	ISIC2017	residual con. dilated conv. GAN	ADV ℓ_2	76.90%	✗	rotation,flipping	✗	✗
[329]	peer-reviewed conference	ISIC2016	skip con.	Tanimoto DS	85.34%	✗	rotation,flipping	✗	✗
[44]	peer-reviewed conference	ISIC2017	residual con.	CE	77.14%	✗	GAN	✗	✗

Table 2.3

Ref.	Venue	Data	Arch. modules	Seg. loss	Perf.	C.D.	Augmentation	P.P.	code
[214]	peer-reviewed journal	ISIC2016 **ISIC2017**	skip con. residual con. dense con.	Jaccard DS	76.50%	✗	-	✗	✗
[241]	peer-reviewed conference	ISIC2017	residual con.	CE Star shape	77.30%	✗	-	✗	✗
[11]	peer-reviewed conference	ISIC2018	skip con. image pyramid attention	TV Focal	74.80%	✗	-	✗	✓
[268]	peer-reviewed conference	ISIC2017	-	Jaccard ℓ_1	78.10%	✗	GAN	✓	✗
[97]	peer-reviewed conference	ISIC2018	dilated conv. parallel m.s. conv. separable conv.	-	83.00%	✗	-	✗	✗
[19]	peer-reviewed conference	**ISIC2018** PH^2	skip con.	Dice	80.00%	✓	rotation, zooming flipping,elastic dist. Gaussian dist. histogram equal. color jittering	✗	✓
[314]	peer-reviewed conference	ISIC2017	skip con. residual con. dense con. attention mod.	CE Jaccard	76.50%	✗	-	✗	✗
[312]	peer-reviewed journal	ISIC2016 **ISIC2017** ISIC2018	skip con. residual con. factorized conv. attention mod. GAN	CE ℓ_1 EPE	78.65%	✗	-	✗	✓

Table 2.3

Ref.	Venue	Data	Arch. modules	Seg. loss	Perf.	C.D.	Augmentation	P.P.	code
[325]	peer-reviewed journal	**ISIC2017** DermoFit PH^2	dilated conv.	Dice	62.29%*	✓	-	✓	✗
[185]	peer-reviewed conference	ISIC2017	skip con. residual con. attention mod.	Dice	75.60%	✗	channel shift	✗	✗
[101]	peer-reviewed conference	**ISIC2017** Private	skip con.	CE Dice	76.07%	✗	flipping, shifting rotation color jittering	✓	✗
[382]	peer-reviewed journal	**ISIC2017** PH^2	skip con. residual con. parallel m.s. conv.	CE Dice DS	78.50%	✓	flipping, rotation whitening contrast enhance.	✓	✗
[346]	abstract	ISIC2017	dilated conv. dense con. skip con.	Dice	76.67%	✗	rotation, flipping, translation, scaling, color shift	✗	✗
[315]	peer-reviewed journal	ISIC2017	residual con.	CE	78.60%	✗	rotation, flipping	✗	✗
[243]	peer-reviewed conference	ISIC2017	skip con.	WCE	68.91%*	✗	-	✗	✗
[83]	peer-reviewed conference	ISIC2017	residual con. dilated conv. parallel m.s. conv.	WCE	78.70%	✗	rotation, flipping cropping, zooming Gaussian noise	✓	✗
[247]	peer-reviewed journal	DermQuest	dense con.	WCE	85.20%	✗	rotation,flipping cropping	✗	✗

Table 2.3

Ref.	Venue	Data	Arch. modules	Seg. loss	Perf.	C.D.	Augmentation	P.P.	code
[169]	non peer-reviewed technical report	ISIC2016 **ISIC2017** ISIC2018	skip con. pyramid pooling parallel m.s. conv.	Tanimoto	80.60%	✓	flipping, rotation zooming,translation shearing,color shift intensity scaling adding noises contrast adjust. sharpness adjust. disturb illumination hair occlusion	✓	✗
[349]	peer-reviewed conference	**ISIC2017** ISIC2018	skip con. residual con. parallel m.s. conv. attention mod.	WDice	77.60%	✗	copping, flipping	✗	✗
[299]	non peer-reviewed technical report	**ISIC2017** ISIC2018	factrized conv. attention mod. GAN	CE Jaccard ℓ_1,ADV	77.98%	✗	flipping gamma reconst. contrast adjust.	✗	✗
[242]	peer-reviewed conference	ISIC2016	skip con.	CE	83.30%	✗	flipping,rottaion	✗	✗
[335]	peer-reviewed journal	**ISIC2017** PH^2	skip con. residual con. dense con. GAN	Jaccard EPE,ℓ_1 DS,ADV	76.80%	✓	flipping	✗	✗
[362]	peer-reviewed journal	ISIC2016 **ISIC2017** PH^2	skip con. residual con. attention mod. GAN	Jaccard ℓ_1 ADV	80.45%	✓	rotation,flipping color jittering	✗	✗
[338]	peer-reviewed journal	**ISIC2017** PH^2	-	ℓ_2	74.81%	✓	-	✓	✗

Table 2.3

Ref.	Venue	Data	Arch. modules	Seg. loss	Perf.	C.D.	Augmentation	P.P.	code
[17]	peer-reviewed conference	ISIC2017	-	-	77.11%	✗	rotation,flipping	✗	✗
[102]	peer-reviewed conference	**ISIC2017** PH^2	dilated conv. parallel m.s. conv. separable conv.	Dice Narrowband suppression	83.9%	✓	rotation	✓	✗
[66]	peer-reviewed conference	ISIC2017	dilated conv. parallel m.s. conv. separable conv.	CE Tanimoto	85.00%	✗	rotating, flipping shifting, shearing scaling color jittering	✓	✗
[354]	peer-reviewed conference	ISIC2017	residual con.	WCE	78.10%	✗	flipping, scaling	✗	✗
[20]	non peer-reviewed technical report	ISIC2018	skip con. residual con. recurrent cnn	CE	88.83%	✗	flipping	✗	✗
[269]	peer-reviewed journal	ISIC2017	-	Tanimoto	78.90%	✗	GAN flipping,rotation shifting, scaling color jittering	✗	✗
[226]	peer-reviewed conference	ISIC2017	skip con. dilated conv.	CE	75.20%	✗	scaling, cropping rotation, flipping image deformation	✗	✗
[275]	peer-reviewed journal	ISIC2017	-	CE	79.20%	✗	rotation, flipping color jittering	✗	✗
[45]	non peer-reviewed technical report	ISIC2018	residual con.	CE	83.12%	✗	GAN	✗	✗

Table 2.3

Ref.	Venue	Data	Arch. modules	Seg. loss	Perf.	C.D.	Augmentation	P.P.	code
[7]	peer-reviewed conference	**ISIC2017** PH^2	skip con.	-	68.69%*	✓	rotation,flipping GAN	✗	✗
[366]	peer-reviewed journal	**ISIC2017** PH^2	dilated conv. parallel m.s. conv. separable conv.	Dice Rank	80.4%	✓	cropping,scaling rotation, shearing shifting,zooming whitening, flipping	✗	✓
[152]	peer-reviewed journal	ISIC2016 **ISIC2017**	skip con. residual con. image pyramid	CE Dice DS	76.10%	✗	rotation	✓	✗
[373]	peer-reviewed journal	ISIC2016	skip con. residual con.	-	82.90%	✗	rotation,translation random noise cropping	✗	✓
[50]	peer-reviewed journal	**ISIC2016** PH^2	-	CE	84.64%	✓	flipping,cropping	✓	✗
[168]	peer-reviewed conference	DermQuest	image pyramid	-	-	✗	-	✓	✗
[166]	peer-reviewed journal	DermQuest	image pyramid	-	-	✗	-	✓	✗
[368]	peer-reviewed conference	ISIC2017	skip con. residual con. global conv. GAN	ℓ_1 DS ADV	78.50%	✗	cropping color jittering	✗	✗
[377]	peer-reviewed journal	**ISIC2016** PH^2	-	Tanimoto	84.7%	✓	flipping, rotation scaling,shifting contrast norm.	✓	✗

Table 2.3

Ref.	Venue	Data	Arch. modules	Seg. loss	Perf.	C.D.	Augmentation	P.P.	code
[384]	peer-reviewed conference	SCD ISIC2016 **ISIC2017** ISIC2018	skip con.	Kappa Loss	84.00%*	✗	rotation,shifting shearing,zooming flipping	✗	✓
[294]	peer-reviewed conference	ISIC2017 **ISIC2018**	skip con. dense con.	CE	81.9%	✗	color jittering rotation flipping translation	✗	✗
[153]	peer-reviewed conference	**ISIC2018**	skip con. parallel m. s. conv. attention mod.	-	78.04%	✗	color jittering rotation,cropping flipping,shift	✗	✓
[167]	peer-reviewed conference	**ISIC2018**	skip con. residual con. dense con.	CE	75.5%	✗	-	✗	✓
[215]	peer-reviewed conference	**ISIC2018**	skip con. residual con. ensemble semi-supervised	CE Dice	75.5%	✗	-	✗	✗
[142]	peer-reviewed conference	**ISIC2018**	skip con. dilated conv. parallel m. s. conv.	Focal Jaccard	77.60%	✗	-	✗	✓
[221]	peer-reviewed conference	**ISIC2018**	skip con. residual con. self-supervised	MSE KLD	87.74%*	✗	-	✗	✗
[8]	peer-reviewed conference	**ISIC2017** DermoFit PH^2	skip con.	MCC	75.18%	✗	rotation flipping	✗	✓

Table 2.3

Ref.	Venue	Data	Arch. modules	Seg. loss	Perf.	C.D.	Augmentation	P.P.	code
[274]	non peer-reviewed technical report	**ISIC2017**	dilated conv.	CE	64.20%	✗	rotation flipping	✓	✗
[61]	peer-reviewed conference	**ISIC2016**	-	CE	82.90%	✗	rotations	✓	✗
[48]	peer-reviewed conference	**ISIC2016**	parallel m. s.	-	86.36%	✗	crops,flipping	✓	✗
[25]	peer-reviewed conference	**ISIC2016**	recurrent net.	-	93.00%	✗	-	✗	✗
[103]	peer-reviewed conference	**ISIC2016**	parallel m. s.	-	84.1%	✗	-	✗	✗
[244]	peer-reviewed conference	**ISIC2017**	skip con.	Dice	84.2%	✗	rotation flipping	✓	✗
[136]	peer-reviewed conference	**ISIC2017**	-	CE Dice	-	✗	-	✗	✗
[176]	peer-reviewed journal	**ISIC2017**, PH^2	skip con. residual con. attention mod.	CE	73.35%	✗	flipping	✗	✗
[273]	peer-reviewed journal	**ISIC2017**, PH^2	ensemble	-	80.02%	✗	translation rotation shearing	✓	✗
[365]	peer-reviewed journal	ISIC2016 **ISIC2017** PH^2	attention mod.	CE	78.3%	✗	rotation flipping	✗	✗

Table 2.3

Ref.	Venue	Data	Arch. modules	Seg. loss	Perf.	C.D.	Augmentation	P.P.	code
[379]	peer-reviewed journal	**ISIC2017**, PH^2	skip con. residual con.	CE	77.2%	✗	rotation	✗	✗
[305]	peer-reviewed conference	**ISIC2018**	skip con. image pyramid	Generalized Dice	73.8%	✗	rotation flipping zooming	✗	✗
[12]	peer-reviewed conference	**ISIC2017**	-	Dice	83.0%	✗	elastic	✗	✗
[27]	peer-reviewed conference	**ISIC 2017** ISIC 2018 PH2	dilated conv. attention mod.	-	96.98%	✗	-	✗	✓
[248]	non peer-reviewed technical report	ISIC 2016 **ISIC 2017** ISIC 2018 PH2	skip con. residual con.	CE Dice	78.28%	✗	rotation, flipping shearing, zoom	✗	✗
[240]	non peer-reviewed technical report	**ISIC Archive** PH2 DermoFit	skip con. residual con. ensemble	CE	72.11%	✗	-	✗	✗
[24]	peer-reviewed journal	ISIC 2018	skip con. attention mod.	Dice Tversky Focal Tversky	83%	✗	flipping	✓	✗
[319]	peer-reviewed conference	ISIC 2017	skip con.	Dice ℓ_1 SSIM	69.35%*	✗	rotation flipping gradient-based perturbation	✗	✗
[257]	peer-reviewed journal	**ISIC 2017** PH2	residual con.	-	78.34%	✓	-	✗	✗

Table 2.3

Ref.	Venue	Data	Arch. modules	Seg. loss	Perf.	C.D.	Augmentation	P.P.	code
[9]	peer-reviewed conference	**ISIC 2017** DermoFit PH2	skip con.	Dice	75.70%	✓	rotation flipping	✗	✓
[296]	peer-reviewed conference	**ISIC 2017** ISIC 2018 PH2	skip con. multi-task	Dice	84.9%	✗	rotation, flipping shearing, stretch crop, contrast	✗	✗
[190]	peer-reviewed journal	ISIC 2017	-	-	72.5%	✗	-	✗	✗
[171]	peer-reviewed journal	ISIC2016	skip con. parallel m.s. conv.	-	92.42%	✗	-	✗	✗
[352]	non peer-reviewed technical report	ISIC2016 **ISIC2017** PH^2	residual con. dilated conv. attention mod.	CE DICE DS	80.30%	✓	flipping, rotation cropping	✗	✗
[355]	peer-reviewed journal	ISIC2016 **ISIC2017**	skip con. residual con. dilated conv.	WCE	81.47%	✗	flipping,scaling	✗	✗
[177]	peer-reviewed journal	**ISIC2017** ISIC2018	skip con. residual con. attention mod.	DICE Focal	80.00%	✗	flipping, rotation affine trans. scaling, cropping	✗	✓
[148]	non peer-reviewed technical report	ISIC-2016 **ISIC-2017**	skip con. residual con. separable conv.	DICE CE	66.66*	✗	flipping, rotation shifting, zooming intensity adjust.	✗	✗
[200]	peer-reviewed journal	**ISIC 2017** PH^2	-	MSE CE	90.25%	✗	-	✗	✗

Table 2.3

Ref.	Venue	Data	Arch. modules	Seg. loss	Perf.	C.D.	Augmentation	P.P.	code
[357]	peer-reviewed journal	**ISIC2018** PH^2	attention mod. skip con. parallel m.s. conv. recurrent CNN	Dice Focal tversky	80.60%	✗	rotation flipping cropping	✗	✗
[284]	peer-reviewed conference	ISIC Archive PH^2 DermoFit	skip con. residual con. dilated conv.	Soft Jaccard CE	-	✓	Gaussian noise color jittering	✓	✓
[110]	non peer-reviewed technical report	**ISIC-2018**	skip con.	Dice	75.6%	✗	rotation flipping zooming	✓	✓
[28]	peer-reviewed journal	ISIC2016 ISIC2017 **ISIC2018** PH^2 DermQuest	parallel m.s. conv. dilated conv.	Dice CE	85.04%	✓	rotation flipping color jittering	✗	✗
[297]	peer-reviewed conference	**ISIC2017** ISIC2018 PH^2	pyramid pooling residual con. skip con. dilated conv. attention mod.	Dice	85.00%	✓	rotation,shearing color jittering	✗	✗
[330]	peer-reviewed journal	ISIC2016 ISIC2017 **PH^2**	skip con. attention mod.	CE	84.2%	✓	flipping	✗	✗
[29]	peer-reviewed journal	**Dermquest** ISIC2017 PH^2	ensemble	CE Focal	86.53%	✓	rotation flipping color jittering	✓	✗
[282]	peer-reviewed journal	ISIC2017	dense con. dilated conv. separable conv. attention mod.	DICE CE	76.92%	✗	flipping, rotation	✗	✗

Table 2.3

Ref.	Venue	Data	Arch. modules	Seg. loss	Perf.	C.D.	Augmentation	P.P.	code
[227]	peer-reviewed journal	ISIC2017	residual con. dilated conv. pyramid pooling	WCE	79.46%	✗	flipping, cropping rotation image deformation	✗	✗
[194]	peer-reviewed journal	ISIC2018	skip con. image pyramid	Dice	85.10%	✗	-	✗	✓
[393]	peer-reviewed conference	ISIC2018	skip con. residual con. dilated conv. attention mod.	CE DICE	82.15%	✗	flipping	✗	✗
[279]	peer-reviewed conference	ISIC2017	-	-	68.77%*	✗	-	✗	✗
[186]	peer-reviewed conference	ISIC2018	skip con. residual con. attention mod.	CE Tversky adaptive logarithmic	82.71%	✗	-	✗	✓
[8]	peer-reviewed conference	**ISIC2017** PH^2 DermoFit	skip con.	MCC	75.18%	✗	flipping, rotation	✗	✓
[328]	peer-reviewed journal	ISIC2018	skip con.	CE	78.25%	✗	-	✗	✗
[367]	peer-reviewed conference	ISIC2018	dilated conv.	CE KL div.	82.37%	✗	scaling,rotation elastic transformation	✗	✗
[270]	peer-reviewed journal	ISIC2017	skip con. attention mod.	CE	87.44%	✗	scaling, flipping rotation Gaussian noise median blur	✗	✗

Table 2.3

Ref.	Venue	Data	Arch. modules	Seg. loss	Perf.	C.D.	Augmentation	P.P.	code
[295]	peer-reviewed journal	ISIC2016 **ISIC 2017**	skip con. Gaussian process	-	74.51%	✗	resize rotation reflection	✓	✗
[301]	peer-reviewed journal	**ISIC 2017** ISIC 2018	parallel m.s. conv. attention mod. GAN	ℓ_1 Jaccard	81.98%	✗	flipping, contrast gamma reconstruction	✗	✗
[356]	peer-reviewed journal	ISIC 2016 **ISIC 2017**	residual con. skip con. lesion-based pooling feature fusion	CE	82.4%	✗	flipping, scaling cropping	✗	✗
[292]	book chapter	ISIC 2018	residual con. skip con.	-	75.96%	✗	flipping, scaling color jitter	✗	✗
[363]	peer-reviewed journal	**ISIC 2017** ISIC 2018 PH2	BConvLSTM separable conv. residual con. skip con.	Jaccard	80.25%	✗	distortion, blur color jitter contrast gamma sharpen	✓	✓
[141]	peer-reviewed journal	ISIC 2018	dilated conv. residual con. skip con.	CE	91%	✗	flipping, scaling shearing, color jitter Gaussian blur Gaussian noise	✗	✓
[195]	peer-reviewed journal	ISIC 2016 **ISIC 2017** ISIC 2018 PH2	feature pyramid residual con. skip con. attention mod.	-	86.92%*	✗	-	✗	✗
[140]	peer-reviewed journal	ISIC 2018	residual con. skip con. attention mod.	Dice	85.32%*	✗	cropping, flipping rotation	✗	✓

Table 2.3

Ref.	Venue	Data	Arch. modules	Seg. loss	Perf.	C.D.	Augmentation	P.P.	code
[139]	peer-reviewed conference	ISIC 2017	asymmetric conv. skip con.	DS	79.4%	✗	cropping, flipping rotation	✗	✗
[388]	peer-reviewed journal	ISIC 2018	pyramid pooling attention mod. residual con. skip con.	CE Dice	86.84%	✗	cropping	✗	✗
[327]	peer-reviewed journal	ISIC 2016 **ISIC 2017** ISIC 2018	attention mod. residual con. skip con. ensemble	Focal	80.7%	✗	copying	✗	✗
[398]	peer-reviewed journal	ISIC 2018	pyramid pooling sharpening kernel residual con.	CE	79.78%	✗	-	✗	✓
[99]	peer-reviewed journal	**ISIC2018** PH2	residual con. skip con. dilated conv. image pyramid	CE Dice SoftDice	83.45%	✓	cropping, flipping rotation	✗	✗
[46]	peer-reviewed journal	ISIC2016 **ISIC2017** PH2	attention mod. residual con. skip con. attention mod.	CE	83.70%	✓	cropping, flipping	✗	✗

Chapter 3

Deep Auto-context Fully Convolutional Neural Network for Skin Lesion Segmentation

3.1 Introduction

Over the last three decades, the prevalence of skin cancer in the United States (U.S.) has been higher than all other cancers combined [285]. The most lethal type of skin cancer is melanoma with the mortality rate of one person per hour in the U.S. [3]. Early detection of melanoma plays an essential role in increasing the skin cancer survival rate. Even when utilizing dermoscopy with skin reflection suppression and widely used diagnostic criteria, like the 7-point checklist, diagnostic accuracy is still not perfect. Development of automatic approaches for skin lesion analysis has the potential to accelerate and improve skin cancer detection and improve survival prognosis.

In this Chapter, we propose a deep auto-context architecture that incorporates image appearance information as well as contextual information to predict the pixel-wise probability of a skin lesion. A sequence of fully convolutional networks is trained in a consecutive manner, where the input of each classifier is the original image concatenated with a degraded a posteriori probability estimated by the previous classifier. In contrast to common approaches that use morphological operations or thresholds to correct irregularities in the predicted lesion segmentation mask, our auto-context architecture efficiently refines the skin segmentation without any post-processing.

3.1.1 Auto-context

Auto-context is an iterative learning algorithm for structural refinement, which uses contextual information in addition to appearance information for image understanding models [336]. Auto-context takes as input appearance information as well as features from the predicted probability maps of the previous iteration into the current iteration. By iterating

this process, classifiers are able to gradually correct earlier mistakes by using new contextual features. The original auto-context algorithm was originally proposed for patch-based segmentation with handcrafted features [336]. Salehi et al. recently proposed Auto-Net, an auto-context CNN to extract the fetal brain from 3D MRI [298].

3.1.2 Contributions

In this work, we applied an auto-context deep framework that sequentially learns improved skin lesion segmentation maps given RGB skin images. We train a sequence of FCNs so that each take as input the original images as well as the degraded a posteriori probability map estimated by the previous early-stopped FCN. Compared to earlier patch-based auto-context approaches, feeding the whole contextual information into a CNN, leads to automatic learning of deep multi-scale contextual features. Also, in comparison to Auto-Net, during training we use the probability maps generated by early-stopped FCNs to prevent overfitting in the subsequent models, and use the fully converged FCNs for testing. The goal of this work is to show the advantage of applying deep architectures to the skin lesion segmentation task in an auto-context fashion. Our experimental results illustrate how deep auto-context framework and the early stopping technique refine the predicted probabilities when compared to a single FCN.

3.2 Methodology

3.2.1 Deep Auto-context

Given a set of N images and their corresponding ground truth segmentations $\{(X(n), Y(n)); n = 1, 2, \ldots, N\}$, our goal is to learn the segmentation model parameters θ that generalize well on unseen samples. For an image with m-pixels $X = (x_1, x_2, \ldots, x_m)$ and a corresponding ground truth labelling $Y = (y_1, y_2, \ldots, y_m)$ such that $y_i \in \{0, 1\}$, we seek a dense prediction configuration Y^* which maximizes the a posteriori probability given an observed image, $p(Y|X; \theta)$.

For the binary segmentation task in fully convolutional networks (FCNs), the a posteriori probability $p(\cdot)$ is modeled by applying a sigmoid function on the activation maps in the last layer as follows:

$$p(y_i = 1|X(N_i); \theta) = \frac{1}{1 + exp(-a(X(N_i))} \tag{3.1}$$

where $X(N_i)$ is a neighboring window around pixel x_i, and a is the output activation. The a posteriori probability gets updated iteratively by measuring the compatibility of Y^* and Y in a loss function and back propagating the error to update the set of model parameters θ. We note that although we write this equation in terms of individual pixels, we predict

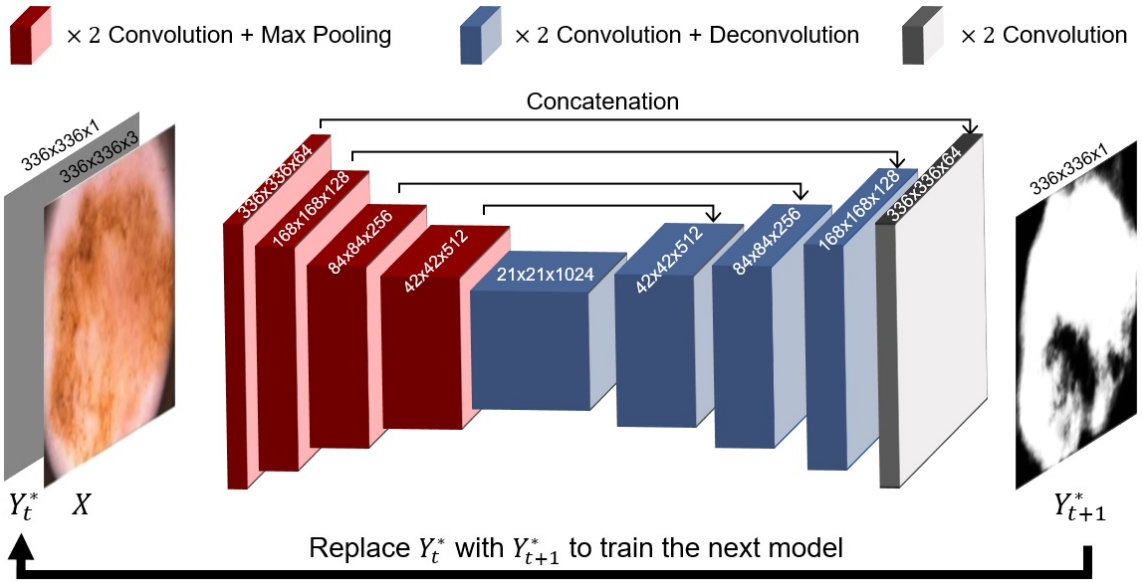


Figure 3.1: Deep auto-context architecture schematic. The model $t+1$ is trained on the concatenation of the original image and the a posteriori probability from model t. Sizes on red and blue blocks show the feature map sizes before max pooling and deconvolution, respectively.

the entire dense segmentation in a single forward pass. In addition, the neighboring pixels N_i in equation 3.1, highlight how the output value y_i is dependent on the network receptive field, and not all the pixels in the input X.

As illustrated in equation 3.1, in conventional classification approaches like FCN, the likelihood of assigning the class label to a pixel, only depends on the deep features extracted from image appearance. However, considering the class labels of surrounding pixels is informative. Conditional random field [204] has been widely used to explicitly formulate the dependency of each class label to the neighboring pixels class labels. The common drawback of these approaches is the explosion of computational complexity if long-range contextual information from a large neighborhood is considered in the model.

To include a large scale of context information in the predicted segmentation, we adopt the auto-context architecture, which is composed of multiple FCN models. The idea is to design an iterative framework that predicts pixel-wise classification not only based on the image appearance but also considers the a posteriori probabilities estimated by the previous classifier. In the proposed approach, we have a sequence of T FCN models learned in a consecutive manner. The $t+1$-th model is trained given the training data $(X(n), Y(n), Y_t^*(n)); n = 1, 2, ..., N$, where $Y_t^*(n)$ is the a posteriori probability provided by:

$$Y_t^* = p(Y|X, Y_{t-1}^*; \theta_t). \tag{3.2}$$

For the first model ($t = 1$), the segmentation probability map, $Y_1^*(n)$, is a uniform distribution map. Since the uniform distribution does not contain actual contextual information, in the first iteration, the network gains no additional information from it. Fig. 3.1 shows the proposed deep auto-context architecture. We start by training a fully convolutional network with an architecture similar to U-Net to segment the skin lesion. Once the first FCN is trained, it is applied to all training and validation data and a posteriori probability map is generated for each image. We concatenate the original RGB image channel with the a posteriori probability map and train a new FCN to refine the a posteriori probability estimated by previous network. The same procedure is repeated until the algorithm converges. At the test time, given a new image, FCNs are sequentially applied.

3.2.2 Overfitting Avoidance

When training using the auto-context architecture, passing the training data sequentially to the subsequent FCN models may not ensure effective fine-tuning because the data and their ground truth were already shown to the previous models. One way to prevent overfitting, when training patch-based auto-context models, is to split the data in such a way that $t + 1$-th model is not trained on data used in the first t models [189]. Splitting the data in this way may not be the best approach to deal with this overfitting. An alternative approach for dealing with this problem is degrading the a posteriori segmentation maps generated by the FCN to produce new maps that look more like the segmentation probability map encountered with novel test images. In this work, we hypothesize that using the parameters of the t-th auto-context model *before convergence* will result in degraded segmentation maps that in turn cause the $t+1$-th model to be trained on more realistic and challenging a posteriori probability maps (rather than ones already overfit to the training data). Thus, during training, we train each FCN until convergence but generate the a posteriori probabilities of training data using the network parameters before convergence. We feed the concatenation of these probability maps and the original image to the next deep model. At test time, we applied the sequence of fully converged FCNs to a new test image.

3.3 Experiments

3.3.1 Data Description

We validated the proposed method on ISBI 2016, *Skin Lesion Analysis Towards Melanoma Detection Challenge*, dataset [144]. The dataset is composed of 900 training images. We used 20% of the training data for validation, and to set model hyper-parameters. Another separate set of 379 test images and their ground truth, provided by the challenge organizers, is used to evaluate the model. We re-sized all images to 336×336 and normalized them using the mean and standard deviation of RGB pixels values computed over all training

Table 3.1: Segmentation quantitative performance comparison in U-Net and different auto-context iterations. T is the number of FCNs in the auto-context model. Bold numbers indicate the best performance. All values are in percentages.

	Method	DICE	J	AC	SP	SE
A	U-Net [287]	86.86	78.29	93.63	93.51	93.05
B	Ours (T=2)	88.42	80.16	95.09	95.07	93.51
C	Ours (T=3)	88.98	80.76	**95.12**	92.04	**96.11**
D	Ours degraded (T=2)	**90.11**	**83.30**	95.02	**97.00**	90.15

data. To increase training data and make the model more robust, we augment the training images with the rotations of 90, 180 and 270 degrees, and horizontal and vertical flipping without any replication.

3.3.2 Implementation

We implemented and trained our deep auto-context networks using the PyTorch framework. All fully convolutional networks are initialized by a random Gaussian distribution and learned from scratch. We used stochastic gradient descent (SGD) as a solver and a mini-batch of size 2, restricted by our GPU memory. A momentum of 0.99 and a weight decay of 0.0005 is used for all fully convolutional networks. The learning rate was tuned for each FCN on our validation set. The network trained for the first step of the auto-context architecture converges after approximately $90,000$ iterations while the second and third networks in the auto-context architecture take approximately $34,000$ and $11,000$, respectively. Training the whole deep auto-context architecture takes 2 days on our single 12 GB GPU memory.

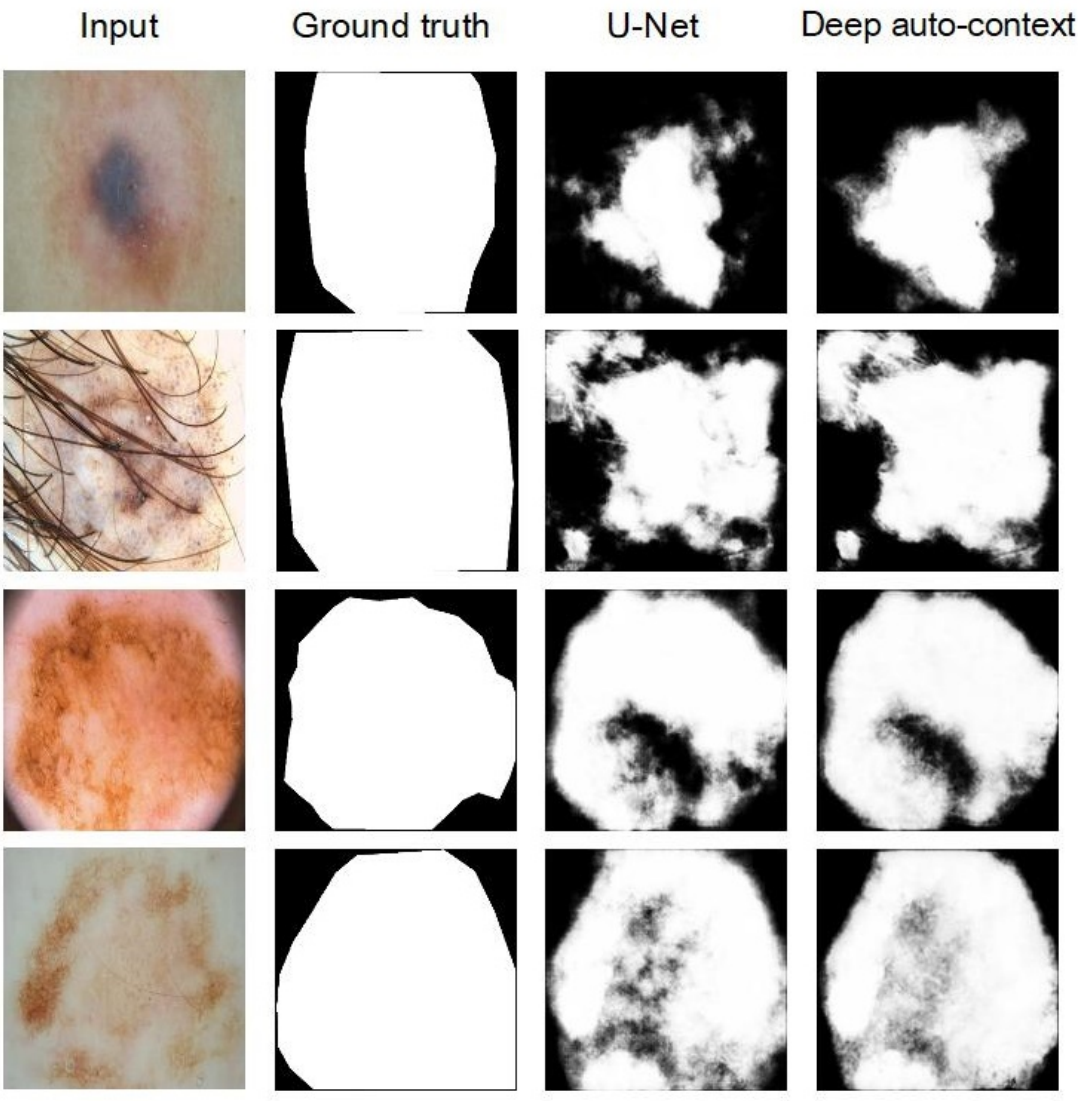


Figure 3.2: Resulting segmentation masks over challenging cases.

3.3.3 Results

We evaluate the contribution of stacking the FCNs in an auto-context by using the following pixel-level metrics used in the ISBI 2016 segmentation challenge: Sensitivity, Specificity, Accuracy, Jaccard and Dice. We calculated each of these metrics for each test image and reported the average value over all test samples. We use U-Net [287], as our baseline network architecture. For a fair comparison, we used the same architecture for sequential FCNs in the auto-context architecture. Each model within the auto-context architecture is trained individually by optimizing a binary cross entropy loss function. Table 3.1 indicates the performance of auto-context in different iterations in comparison to U-Net. Rows B and C confirm the advantage of using FCNs in an auto-context fashion. In comparison to U-Net, after one iteration of auto-context, Dice similarity coefficient increases by approximately

1.5% (Row B). Iterating the auto-context for the second iteration further improve the Dice similarity coefficient (0.5% as shown in Row C). In our experiments, we found that iterating the auto-context model beyond the second iteration did not improve results. To degrade the a posteriori probability maps of training data and make them more similar to the probability maps of unseen data at test time, we used the FCN parameters at the auto-context iteration 0 before convergence (after $34,000$ iterations) and generate the training data probability maps. Row D shows the result of training the FCN at iteration 1 using contextual information generated by the degraded probability maps. Iterating the new auto-context model after the first iteration didn't improve the predictions. We observe that degrading the a posteriori probability maps helps avoid overfitting and improves results, and thus recommend the degraded (T=2) as the best option. 28 teams have participated in the skin lesion segmentation part of the 2016 challenge. Based on these reported numbers[1], our performance ranked the second among the challenge participants.

Fig. 3.2 presents qualitative results of our proposed approach over some challenging cases. Comparing the results of our degraded deep auto-context with U-Net illustrates that by iterative training a fully convolutional network using the a posteriori probability segmentation map in addition to the original image, FCNs are able to gradually correct earlier mistakes by using new contextual features. While many previously proposed deep architectures for skin segmentation apply post-processing approaches (e.g., multi-thresholding, morphological operations) to filter false negative and positive gaps inside and outside the lesions [376], these post-processing operations are disconnected from the training step, require additional hyper-parameters, and are computationally expensive at test time.

3.4 Conclusions

We proposed to use a sequence of fully convolutional networks in an auto-context manner to sequentially refine the predicted skin lesion segmentation map of the previous network. The key contribution of this work is to incorporate contextual information into deep feature extraction models. Our proposed deep auto-context approach is a general, easy to implement framework, that is applicable regardless of the deep architecture, and is used to further refine segmentations.

[1] `https://challenge.kitware.com/#challenge/560d7856cad3a57cfde481ba`

Chapter 4

Generative Adversarial Networks to Segment Skin Lesions

4.1 Introduction

The task of Melanoma segmentation is not trivial as melanoma is subject to many challenging variability in appearance such as size, shape, and texture. Furthermore, melanoma potentially has fuzzy boundaries such that the contrast between the lesion and its surroundings may be unclear. Furthermore, within the image, irrelevant distracting artifacts may be present, such as hairs, vessels, air blobs, medical gauze, and light reflection, over the lesion surface, which makes the segmentation task more difficult and error-prone.

In this Chapter, we present a novel approach for skin lesion segmentation through leveraging generative adversarial networks. Our approach consists of two models: a fully convolutional neural network designed to synthesize an accurate skin lesion segmentation mask (the segmenter), and a convolutional neural network that distinguishes between synthetic and real segmentation masks (the critic). Our experimental results on 1300 images from the DermoFit dataset show that incorporating a critic network to complement a fully convolutional segmenter, like UNet, increases segmentation accuracy.

4.1.1 Generative Adversarial Networks

Deep learning techniques based on generative models, known as generative adversarial networks (GANs) [133], have further pushed the state of the art in some domains. Generally, GANs perform a minmax game between two players, namely a *generator* and a *discriminator* network, referred to as a *segmenter* and *critic*, respectively, in our work. Given a training dataset, the generator/segmenter attempts to synthesize outputs that match the ground truth segmentations, while the discriminator/critic is responsible for distinguishing between synthetic and real outputs. Training these two networks in an adversarial fashion results in two strong models after stabilization.

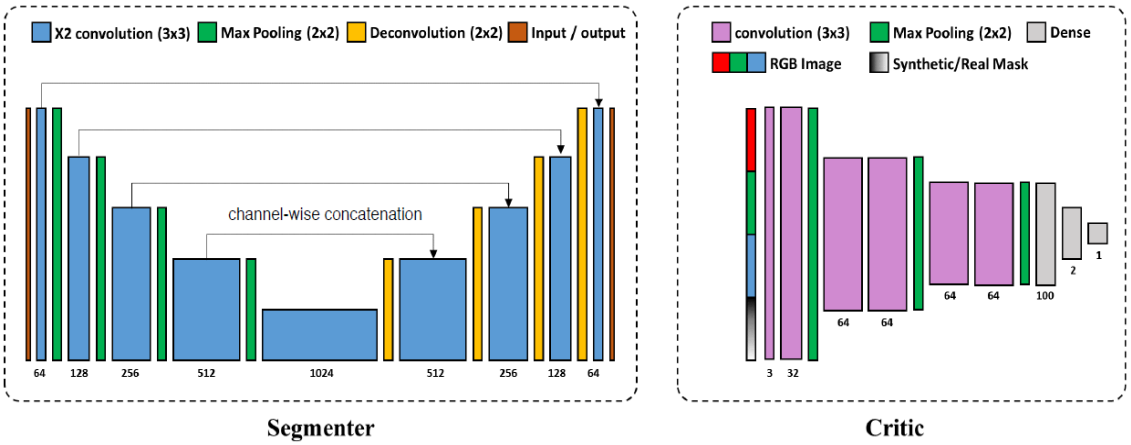


Figure 4.1: The schematic of the proposed UNet-Critic model for skin lesion segmentation. The error in the critic is backpropagated through the segmenter to make it produce more realistic segmentation masks.

4.1.2 Contributions

Inspired by Pan et al. [258] and Luc et al. [230], we propose to use a generative adversarial network to segment skin lesions. In the domain of medical images, other works have used GANs to segment aggressive prostate cancer [197] and brain regions [245] from MRI images. In this paper, we aim to practically examine the role of critic network in improving the performance of an existing model. To this end, we use a fully convolutional segmentation model and augment it with a critic neural network model. Once trained appropriately, we show that including the critic into our network increases the quality of segmentations produced by the segmenter model, compared to the case of a critic-free network architecture. We evaluate our model on the DermoFit skin lesion dataset.

4.2 Method

Our goal is to accurately segment the skin lesions from their surroundings, independent of the diversity in their appearance and without any manual intervention. To do so, we frame the problem as a binary dense labeling task: Given a dermoscopic image, we aim to predict either “lesion" or “background" labels for each pixel.

Given an existing fully convolutional segmentation model, i.e. segmenter, which synthesizes probabilistic segmentation masks, we propose to design and employ a DCNN with a single output node, i.e. critic, to distinguish between the synthetic segmentation masks and real ground truth. By importing the feedback from the critic into the segmenter, the latter learns to produce more plausible lesion segmentations. A stabilization state occurs when the segmenter synthesizes segmentation masks that the critic is unable to differentiate from ground truth lesion segmentations. We hypothesize that by training these two networks

adversarially, the competitive atmosphere will lead to a segmenter that produces more accurate lesion segmentations. Our experimental results demonstrate that adding the critic network to the segmenter model leads to improvements compared to increasing the complexity of the architecture and/or the design. Fig. 4.1 depicts the schematic of our proposed model.

4.2.1 Segmenter

We use UNet [287] as our base segmenter model. This model has an encoder-decoder architecture to transform an RGB image into a segmentation mask, while connecting the feature maps from earlier to later layers. These so called skip connections deliberately leverage the precise localization cues captured in earlier layers to produce finer boundaries in the resultant segmentation mask.

Let $I_{rgb} \in \mathbf{I}$ denote an input image, $\mathcal{T} \in \mathbf{T}$ be the ground truth mask, and $\mathcal{M} \in \mathbf{M}$ be the synthetic segmentation mask. Each pixel i in the segmentation mask $\mathcal{M} = \{m_i, i = 1, \ldots, N\}$ takes a value in the range $L = [0, 1]$, and each pixel in $\mathcal{T} = \{t_i, i = 1, \ldots, N\}$ takes a value from the set $\{0, 1\}$. Given an input image, I_{rgb}, and a set of learned parameters, θ_s, the conditional probability of a label assignment $\mathcal{M}$ is:

$$P(\mathcal{M}|I_{rgb}; \theta_s) = \sigma\left(\psi_{\theta_s}(I_{rgb})\right) \tag{4.1}$$

where $\sigma(\cdot)$ is the *sigmoid* activation function applied to the neural network model's output layer $\psi_{\theta_s}(.)$. We use binary cross entropy as the loss function to train the segmenter network, which is computed as follows:

$$\mathcal{L}_{\theta_s} = -\frac{1}{N}\sum_{i=1}^{N}\left[t_i \log(m_i) + (1 - t_i)\log(1 - m_i)\right] \tag{4.2}$$

where t_i and m_i are the predicted and true labels for each pixel, respectively.

4.2.2 Critic

We augment the segmenter network with a DCNN that receives a dermoscopic image and either a synthetic or real lesion segmentation mask as inputs, and attempts to distinguish between the two cases. In particular, synthetic or real segmentation masks are concatenated to the RGB image along the channel dimension, and are assigned a label of 0 (indicating a synthetic image) or 1 (indicating a real image). The new 4-channel image is fed into a set of convolutional and max-pooling operations (Fig. 4.1), and the final single node learns to predict the true binary labels. The critic network contains six 3×3 convolutional layers, three max-pooling, and three linear layers, all using ReLU activation functions, except for the final layer which uses the sigmoid function. Batch normalization [160] is also used after every convolution operation.

Figure 4.2: Results of elastic deformation on skin lesions and their corresponding segmentation masks.

As above, let $I_{rgb} \in \mathbf{I}$ be an input image and $\mathcal{S} \in \{\mathcal{M}, \mathcal{T}\}$ be either the synthetic or real segmentation mask. After both inputs (I_{rgb} and $\mathcal{S}$) are concatenated along the channel dimension, it can take the label value of $L = \{0, 1\}$. Once fed into the network, the conditional probability of a label assignment y is:

$$P(y|I_{rgb}, \mathcal{S}; \theta_c) = \sigma\left(\psi_{\theta_c}(I_{rgb}, \mathcal{S})\right) \tag{4.3}$$

where θ_c denotes the parameters in the critic network, and $\psi_{\theta_c}(\cdot)$ refers to the output of the critic network. Similar to the segmenter model, we use the binary cross entropy as the loss function for training the critic network, denoted as $\mathcal{L}_{\theta_c}$.

4.2.3 Training

Optimizing the proposed framework proceeds by alternating between training the segmenter to produce synthetic segmentation masks while keeping the critic fixed, and training the critic using the synthetic and real segmentation masks while the segmenter is fixed. The error in the critic must be backpropagated through the segmenter in order for the segmenter to learn how to produce segmentations that can fool the critic. This is performed through adding the pixel-wise binary cross entropy error in the segmenter with that of critic. Thus, the final loss function for updating the segmenter is as follows:

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\theta_s} + \lambda \mathcal{L}_{\theta_c} \tag{4.4}$$

where $\lambda = 0.2$ is the coefficient to balance the effect of critic error value. In other words, the coefficient is set to encourage the learning rate of both networks to be comparable in value, in order to reduce training instability.

Table 4.1: Quantitative Results. Bold numbers indicate the best performance.

	Method	DICE	J	SE	SP	AC
A	U-Net [287]	0.887	0.781	**0.906**	0.955	0.936
B	UNet-Critic	**0.898**	**0.812**	0.891	**0.971**	**0.942**

4.3 Experimental Results

We validate our proposed model on the DermoFit dataset [33], which contains 1300 high quality focal skin lesions. The dataset contains lesions from ten different disease categories and encompasses various potential challenges in lesion appearances, which complicates the segmentation task. In addition to category annotation, there exists a binary segmentation mask for each image. To the best of our knowledge, we are the first to benchmark an automatic lesion segmentation approach on this challenging dataset.

4.3.1 Data Augmentation

In addition to horizontal flipping, vertical flipping, and rotations, we apply a set of elastic deformations to each image to generate synthetic lesions with different geometric shapes. Fig. 4.2 shows a set of newly generated lesions using DeformIt [146]. We enlarge the size of the training set by a factor of $\sim$60.

4.3.2 Implementation Details

We divide the original dataset into a training (80%) and test set (20%), and augment the training examples with the aforementioned mask-consistent deformations. The segmenter network is trained for 10 epochs ($\sim$60K iterations). Since the segmenter and critic networks are trained alternately in the adversarial setting, we double the number of epochs in the *UNet-Critic* model for fair comparison. We use SGD with momentum and weight decay regularization to train both networks. All hyper-parameters (e.g., learning rate, λ) of our model are selected on 20 percent of the unaugmented training set via grid search. Training takes $\sim$35 hours on a machine with one Titan X (Pascal) GPU using Lasagne [107].

4.3.3 Quantitative Results

Table 4.1 presents the resulting quantitative metrics from our proposed model and the competing approach. We see that by incorporating a CNN critic, the segmentation performance of UNet is improved. We also highlight this work as a substitution approach to other works that use more complicated architectures, where this additional training may improve overall model performance.

4.3.4 Qualitative Results

As shown in Fig. 4.3, our approach leads to more uniform and compact segmentation masks. We note that our proposed model succeeds in filling the holes in the foreground and eliminating the island regions in the background area. For lesions with clearly defined boundaries, both approaches perform similar; however, our proposed model yields higher quality segmentations than the simple UNet model for cases with low contrast, unclear and irregular lesion boundaries. Our experiments verify that the boundary improvements are a result of adding the critic, which explicitly compares the synthetic segmentation mask with the ground truth during training.

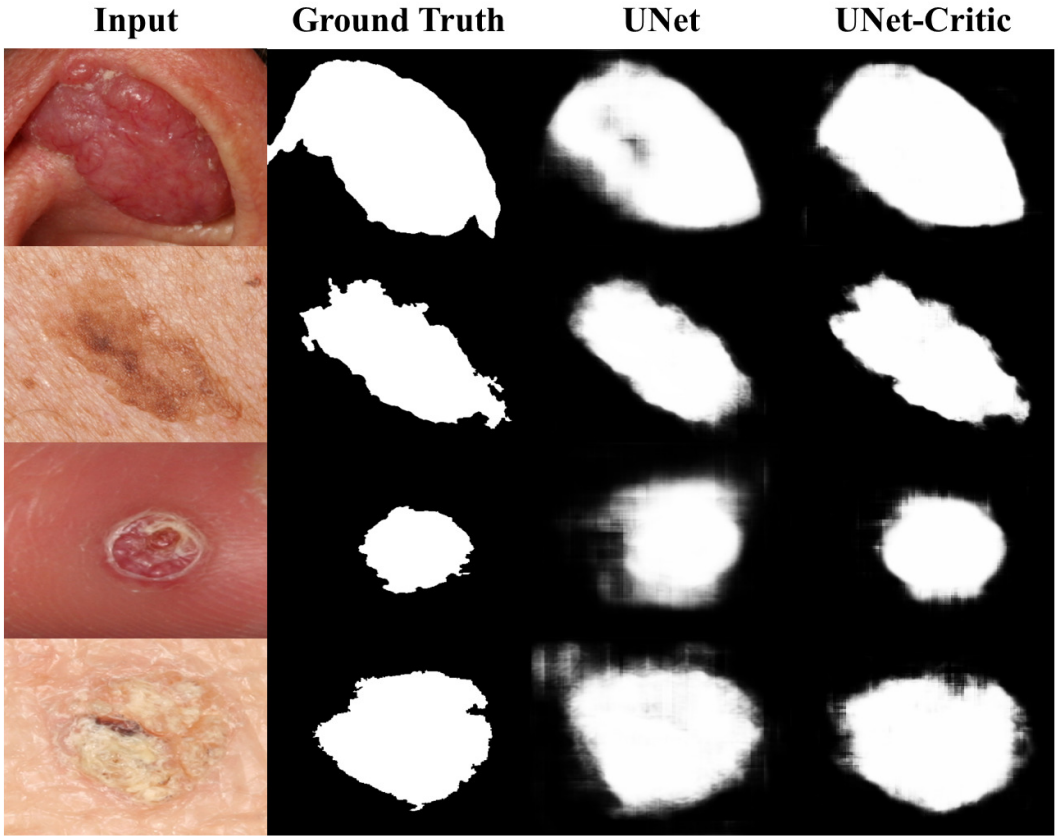


Figure 4.3: Qualitative results of UNet-Critic vs. UNet.

4.4 Conclusion

We examined the effect of adding a critic network to an existing skin lesion segmenter model. The critic network receives the synthetic or real segmentation mask along with the input dermoscopy image and learns to distinguish between these two cases. We then backpropagate the error of the critic into the segmenter training procedure to encourage more realistic segmentation masks. Quantitatively, our proposed approach shows a relative

improvement to a state-of-the-art model. Our qualitative results also reveals that including the critic module helps the segmenter to uniformly highlight the interior regions of the lesion and produce fine predictions around boundaries. Our work is also the first to benchmark lesion segmentation over the DermoFit dataset. Future work would evaluate our proposed approach over other skin datasets such as the ISIC dataset [89].

Chapter 5

Star Shape Prior in Fully Convolutional Networks for Skin Lesion Segmentation

5.1 Introduction

While incorporating prior knowledge about the structure of target objects has proven effective in traditional energy-based segmentation approaches, there has not been a clear way for encoding prior knowledge into deep learning frameworks. In this Chapter, we propose a new loss term that encodes the star shape prior into the loss function of an end-to-end trainable fully convolutional network (FCN) framework. We penalize non-star shape segments in FCN prediction maps to guarantee a global structure in segmentation results.

5.1.1 Prior Knowledge Incorporation in Objective Functions

For decades, since the seminal work of Kass et al. [183], energy functional minimization techniques were the most popular approaches to solve image segmentation problems [235]. Imaging artifacts and variability in the appearance of image regions make the data fidelity term insufficient to achieve robust segmentation results. Therein, the segmentation that minimizes a weighted sum of unary (data) and regularization energy functional terms is sought. Incorporating prior knowledge about the structure of target object in the objective function to regularize plausible solutions with anatomically meaningful constraints have been widely leveraged to obtain more reliable delineations [94, 253]. Active shape models (ASM) was one of the pioneering works to incorporate shape priors into deformable models [93]. To effectuate the shape prior, ASM and many other shape-encoding segmentation methods required an estimate of the object pose (i.e., the orientation, scale, and location of the target object in the image) [117, 348]. Some examples of priors which have been utilized in energy optimization based segmentation methods are shape models, topology

preservation, moment constraints and geometrical and distance interaction between image regions.

In the context of fully convolutional networks, leveraging prior information about the target object structure in the segmentation model has not been widely studied. By optimizing individual pixel-level class predictions in the FCNs loss function, independent class labels are assigned to image pixels without considering high-level label dependencies. There have been some efforts towards structured prediction and leveraging meaningful priors into deep learning frameworks. Deeplab-CRF and CRF-RNN employ probabilistic graphical modeling either as a post-processing step or by implementing recurrent layers in FCNs to enforce assigning similar labels to pixels with similar color and position and further improve the object boundaries [79]. Recently BenTaieb et al. proposed a new loss function to encode the geometrical and topological priors of containment and detachment in an end-to-end FCN framework [42, 391]. To leverage the shape prior in segmentation models, Chen et al. learn a shape constraint by a deep Boltzmann machine and then employ the learned prior in a variational segmentation method [78]. In addition, training convolutional auto-encoder networks to learn anatomical shape variations has demonstrated improvements in the robustness of FCN segmentation models [254, 278].

To the best of our knowledge, none of the existing works incorporates a star shape prior as a regularization term in the loss function of FCNs trained in an end-to-end fashion. The star shape prior was first introduced in the context of image segmentation by Veksler, where it was encoded as a regularization term into the cost function formulation of a graph-based (discrete) image segmentation approach [343]. Later, Chittajallu et al. incorporated three types of shape constraints including star shape prior into a Markov random field based segmentation model and applied their method to non-contrast cardiac computed tomography scans [85]. Yuan et al. extended the star shape prior to 3D objects and applied it to prostate magnetic resonance images [375]. Nosrati et al. derived a star shape prior in a continuous variational formulation and applied it to segmenting overlapping cervical cells [252]. Although the star shape prior clearly improved results for a variety of target objects, one limiting requirement of Veksler's approach and its variants, however, is the assumption that the center of foreground objects is known (e.g. provided by user interaction).

5.1.2 Contributions

We aim to harness the powerful proven capabilities of deep learning in automatically extracting learnt (i.e., not hand-crafted) pixel-driven image features (i.e., likelihood) and augment it with demonstrably useful shape priors without requiring the knowledge of the target object pose. We propose to encode the star shape prior into the training of fully convolutional networks to improve segmentation of skin lesions from their surrounding healthy skin. Our idea is to formulate the star shape prior in the loss function of FCN frameworks to penalize non-star shape segments in prediction maps and preserve global structures in the output

space. Integration of the star shape prior in the loss function makes it possible to train the whole FCN framework in an end-to-end manner. In contrast to Veksler's work and its variants, our approach to star shape prior in a deep learning setting not only eliminates the need for manually setting object centers, but also alleviates, at inference time, the computationally intensive optimization associated with the energy minimizing approaches. Our experimental results illustrate how imposing the shape prior constraint in deep networks refines skin lesion segmentation in comparison to using a single pixel-level loss in FCNs.

5.2 Methodology

Our goal is to leverage the star shape prior into the learning process of an FCN to generate plausible segmentation maps (e.g. skin lesions) from their surrounding background without requiring additional training, user interaction, pre- or post-processing.

5.2.1 FCN's Pixel-wise Loss

In FCNs, given a set of N training images and their corresponding ground truth segmentations, $\{(X(i), Y(i)); i = 1, 2, \ldots, N\}$, the deep network learns to take unseen image samples and generate a segmentation probability map, the same size as the input images that assigns a semantic label to each pixel. Learning the deep network parameters θ, is performed by maximizing the a posteriori probability of giving the true label to each image pixel given the input image. Maximizing the a posteriori probability is usually replaced by minimizing its negative log-likelihood function as a cost function L:

$$\theta^* = \arg\min_{\theta} L(X, Y; \theta). \tag{5.1}$$

For binary dense class prediction, a binary cross entropy loss L_{ce} is generally deployed:

$$L_{ce}(X, Y; \theta) = -\sum_{i=1}^{N} \sum_{p \in \Omega} y_{ip} \log \hat{y}_{ip}(\theta) + (1 - y_{ip}) \log(1 - \hat{y}_{ip}(\theta)) \tag{5.2}$$

where Ω is the pixel space, y_{ip} is the ground truth label of pixel p in image i and $\hat{y}_{ip}$ is the FCN sigmoid function output indicating the predicted probability of the p^{th} pixel of the i^{th} image being a skin lesion. The pixel-wise binary logistic loss L_{ce} penalizes the deviation of the predicted label for each pixel from its true label.

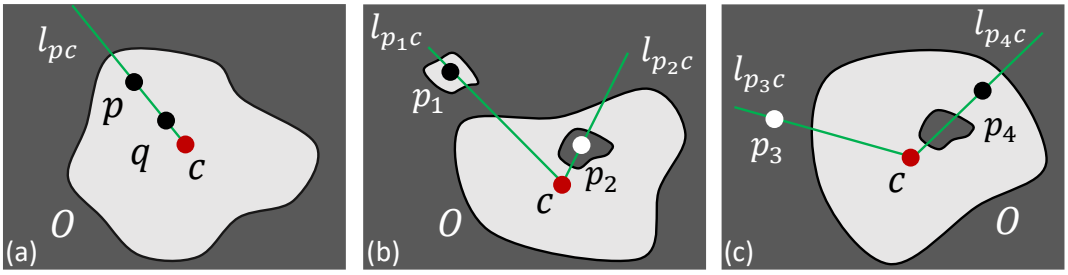


Figure 5.1: (a) Star shape object O w.r.t. the supplied object center c (*red dot*). (b) Examples of the star shape constraint violation. (c) Examples of cases where conditions (i) and (ii) in (5.7) are required.

5.2.2 Star Shape Regularized Loss

Assuming c is the center of object O, object O is a star shape object if, for any point p interior to the object, all the pixels q lying on the straight line segment connecting p to the object (e.g. lesion mask) center c are inside the object (Fig. 5.1-(a)). This definition of star shape prior holds for a large group of object shapes including convex ones. To incorporate the star shape prior as a new regularization term, we augment the loss function in (5.2) with a new loss term to penalize line segments that violate the prior (e.g. Fig. 5.1-(b)) in the prediction maps:

$$L(X, Y; \theta) = \alpha L_{ce} + \beta L_{sh} \tag{5.3}$$

where α and β are hyper-parameters setting the contribution of each term in the optimization function, L_{ce} is the binary cross entropy loss and L_{sh} is our star shape prior:

$$L_{sh}(X, Y; \theta) = \sum_{i=1}^{N} \sum_{p \in \Omega} \sum_{q \in l_{pc}} D_{pq}^{i} \times B_{pq}^{i} \times C_{p}^{i} \tag{5.4}$$

where l_{pc} is the line segment connecting pixel p to the object center c, q is any pixel incident on line l_{pc}.

D_{pq}^{i} is given by

$$D_{pq}^{i} = |\hat{y}_{ip}(\theta) - \hat{y}_{iq}(\theta)| \tag{5.5}$$

and determines how labels of pixels internal to the lesion are penalized to ensure star shapes, i.e. assigns all pixels q a label identical to the label of pixel p (Fig 5.1-a).

B_{pq}^{i} is given by

$$B_{pq}^{i} = \begin{cases} 1, & \text{if } y_{ip} = y_{iq} \\ 0, & \text{otherwise} \end{cases} \tag{5.6}$$

and is designed to allow discontinuities of pixel labels across the ground truth boundary of the lesion, i.e. $B^i_{pq} = 1$ only when p and q have the same ground truth labels.

C^i_p is given by

$$C^i_p = |y_{ip} - \hat{y}_{ip}(\theta)| \tag{5.7}$$

and is used to weighs down D^i_{pq} as the predicted probability of p approaches its ground truth label (i.e., the star shape prior is not applied when the predictions are correct.)

In Fig. 5.1-(c), $p = p_3$ and $p = p_4$ are examples where the value of $\sum_{q \in l_{pc}} D^i_{pq}$ is positive while their assigned labels should not be penalized. B^i_{pq} allows discontinuities between the background (p_3) and foreground assigned labels to pixels q along l_{p_3c}, and C^i_p enforces the loss function not to penalize the label assigned to p_4.

In our implementation of 5.7, instead of penalizing the difference between the predicted probabilities and ground truth labels for all the points on the straight line l_{pc}, we only examine the m closest pixels to p on l_{pc} and compute the loss value per pixel p based on those m predicted probabilities. In more detail, before starting the training process, for each ground truth map in the training data, we find the center of the lesion for image i by averaging the positions of the set of all skin lesion pixels, U:

$$c_i = 1/|U| \sum_{p \in U} (y_{ip_x}, y_{ip_y}) \tag{5.8}$$

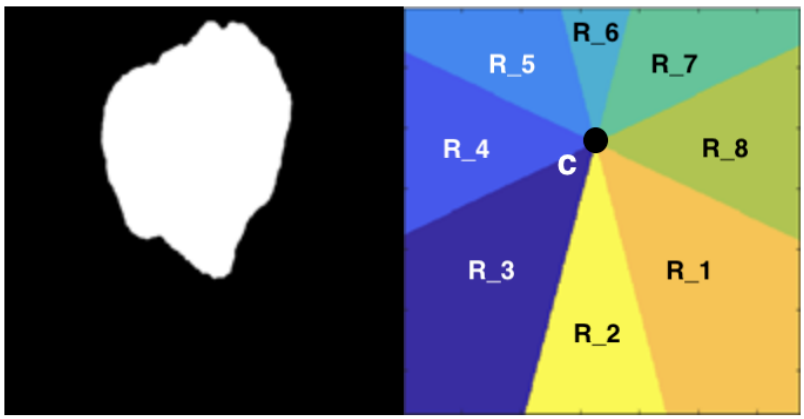


Figure 5.2: A sample of a skin lesion segmentation mask and its corresponding regional map.

In the next step, for each ground truth mask y_i in the training data, we generate a mask by quantizing the possible angles of all lines passing through c_i to a set of d directions and splitting the image domain into d regions. Fig. 5.2 illustrates quantization of the image space into $d = 8$ regions, $\{R_1, ..., R_8\}$, for a sample skin lesion ground truth mask. We design d kernels of size $2m + 1 \times 2m + 1$ to examine the m closest pixels to p on l_{pc} and compute the

loss value per pixel p based on those m predicted probabilities. For example, for $m = 2$ and $d = 8$, the following $K_1, K_2, ..., K_8$ are used to compare the predicted label on pixel p with two closest points along the direction of non-zero values in the matrices. K_k corresponds to R_k in Fig. 5.2. Using convolution of a predicted segmentation probability maps $\hat{y}$ with K_k kernels, we compute $\sum_{q \in l_{pc}} D^i_{pq}$. In more detail, we convolve a predicted segmentation mask $\hat{y}$ with each kernel K_k and concatenate the 8 resulting maps along the channel dimension and perform element-wise multiplication between the resulting tensor and its corresponding one-hot coded regional mask (8-channel mask where channel k takes value one at region k and zero otherwise.)

$$K_1 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 \end{bmatrix}, K_2 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \end{bmatrix}, K_3 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 \end{bmatrix}, K_4 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$K_5 = \begin{bmatrix} -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, K_6 = \begin{bmatrix} 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, K_7 = \begin{bmatrix} 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, K_8 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad (5.9)$$

In order to optimize the model using SGD, the star shape loss needs to be differentiable with respect to model parameters θ. In eq. 5.7, B^i_{pq}, is only based on the ground truth maps and is independent of θ. C^i_p is the mean absolute error between the predicted and true labels that is differentiable at all points except when $y_{ip} = \hat{y}_{ip}$. D^i_{pq} ensuring the labels similairty inside the lesion is implemented using differentiable convolution operations using kernels K_k. So we have:

$$\frac{\partial L_{sh}}{\partial \theta} = \frac{\partial L_{sh}}{\partial \hat{y}_{ip}} \times \frac{\partial \hat{y}_{ip}}{\partial \theta}, \quad (5.10)$$

$$\frac{\partial L_{sh}}{\partial \hat{y}_{ip}} = \sum_i \sum_p \sum_q B^i_{pq} \times [\frac{\partial C^i_p}{\partial \hat{y}_{ip}} \times D^i_{pq} + \frac{\partial D^i_{pq}}{\partial \hat{y}_{ip}} \times C^i_p] \quad (5.11)$$

where

$$\frac{\partial C^i_p}{\partial \hat{y}_{ip}} = (-1)\frac{y_{ip} - \hat{y}_{ip}(\theta)}{|y_{ip} - \hat{y}_{ip}(\theta)|} \quad (5.12)$$

and:

$$\frac{\partial D_{pq}^{i}}{\partial \hat{y}_{ip}} = \frac{\partial |\hat{y}_{ip}(\theta) - \hat{y}_{iq}(\theta)|}{\partial \hat{y}_{ip}} \propto \frac{\partial \hat{y}_{ip}(\theta) \circledast K_k}{\partial \hat{y}_{ip}(\theta)} \tag{5.13}$$

In kernel K_k of size $2m+1 \times 2m+1$, $K[m][m]$ (element in the center) is equal to m and all other non-zero elements are -1. So by expanding the convolution operation we have:

$$\hat{y}_{ip}(\theta) \circledast K_k = m\hat{y}_{ip} - \sum_{q} \hat{y}_{iq}, \tag{5.14}$$

and then:

$$\sum_{q} \frac{\partial |\hat{y}_{ip}(\theta) - \hat{y}_{iq}(\theta)|}{\partial \hat{y}_{ip}} \propto m - \sum_{q} \frac{\partial \hat{y}_{iq}(\theta)}{\partial \hat{y}_{ip}(\theta)}. \tag{5.15}$$

With the above gradients calculated, SGD is used to update the model parameters. Note that in the training of our deep network, we automatically find the star object center from binary ground truth maps. When training is completed, at inference time, the model parameters are known and an input image is fed forward through the network to calculate the dense predicted segmentation map, a procedure that does not require knowing the center of star object.

5.3 Experiments

5.3.1 Data Description

We validated our proposed segmentation approach on dermoscopy data provided by the International Skin Imaging Collaboration (ISIC) at ISBI 2017 *Skin Lesion Analysis Towards Melanoma Detection Challenge* [89]. The dataset contains 2000 training, 150 validation, and 600 test images. We first re-scaled all images to 192×192 pixels and normalized each RGB channel by the mean and standard deviation of the training data. To confirm the suitability of adopting the star-shape prior for this task, we calculated the percentage of segmentation mask pixels that violate the star shape definition to be only 0.14% over the whole dataset (0.05% of training, 0.3% of validation, and 0.38% of test image pixels). Fig. 5.3 shows examples of rare pixels where the star shape constraint is violated.

5.3.2 Network Architecture

We exploited two state-of-the-art fully convolutional network architectures to evaluate our proposed new loss: 1)U-Net[287] 2)ResNet-DUC. ResNet-DUC deploys the FCN version of ResNet-152, pretrained on ImageNet as an encoder [150]. Instead of using multiple deconvolutional layers to decode low resolution feature maps into the original image size prediction

Figure 5.3: Examples of skin lesion pixels violating the star shape constraint.

maps, single Dense Upsampling Convolution (DUC) layer is used to reconstruct fine-detailed information from coarse feature maps [351]. Furthermore, dilated convolutions are used in the encoder to benefit from multi-scale contextual information from previous layers activations [372].

We trained deep networks implemented with the PyTorch library, over mini-batches of size 12. We tuned all hyper-parameters on the validation set. Loss functions are optimized using the SGD algorithm with an initial learning rate of 10^{-4}. The learning rate was divided by 10 when the performance of model on validation dataset stopped improving. Momentum and weight decay were set to 0.99 and $5 \times 10^{-}5$, respectively. For the implementation of the star shape regularized loss function, $\alpha = 1$, $\beta = 5$ and $m = 6$. We first trained the deep network with binary cross entropy function for 5 epochs and then fine-tuned the network parameters with the proposed loss function. Training takes 2 days and test takes 1 sec/image on our 12 GB GPU.

5.3.3 Results

We evaluated the performance of U-Net and ResNet-DUC trained with and without the star shape prior. As shown in Table 5.1, using our shape regularized loss function in the training of U-Net and ResNet-DUC, the Jaccard index is improved by more than 3% (row A vs. B and row C vs. D). We measured the statistical significance of our results by exploring the Jaccard index over the test data. We used the non-parametric Wilcoxon signed rank sum test and found that the results of U-Net and ResNet-DUC with and without incorporation of star shape prior are statistically significantly different at $p < 0.05$.

We compared our proposed method with 21 competing methods participating in the challenge. The ResNet-DUC architecture trained with our star shape regularized loss achieved the first rank based on the challenge ranking metric, Jaccard index. Table 5.1, rows E, F and G, show results of the first three ranked teams. Although all top three teams used FCNs to perform image segmentation, in contrast to our work, they employed various additional steps like averaging over multiple model results, multi-scale image input as well as pre- and post-processing approaches like inclusion of different color spaces in the input and multi-thresholding. Qualitative results of our proposed approach are presented in Fig. 5.4.

Table 5.1: Segmentation quantitative performance. Bold numbers indicate the best performance. All values are in percentages.

	Method	Jaccard	Dice	Accuracy	Specificity	Sensitivity
A	U-Net [287]	70.5	79.7	91.8	97.8	77.0
B	U-Net + Star Shape	73.3	82.4	92.4	95.3	85.4
C	ResNet-DUC [351]	74.0	83.3	93.00	98.2	80.0
D	ResNet-DUC + Star Shape	**77.3**	**85.7**	**93.8**	97.3	**85.5**
E	Yuan et al. [376]	76.5	84.9	93.4	97.5	82.5
F	Berseth et al. [43]	76.2	84.7	93.2	97.8	82.0
G	Bi et al. [47]	76.0	84.4	93.4	**98.5**	80.2

Encoding star shape prior into the loss function results in smoother prediction maps with a single connected component as lesion for most cases.

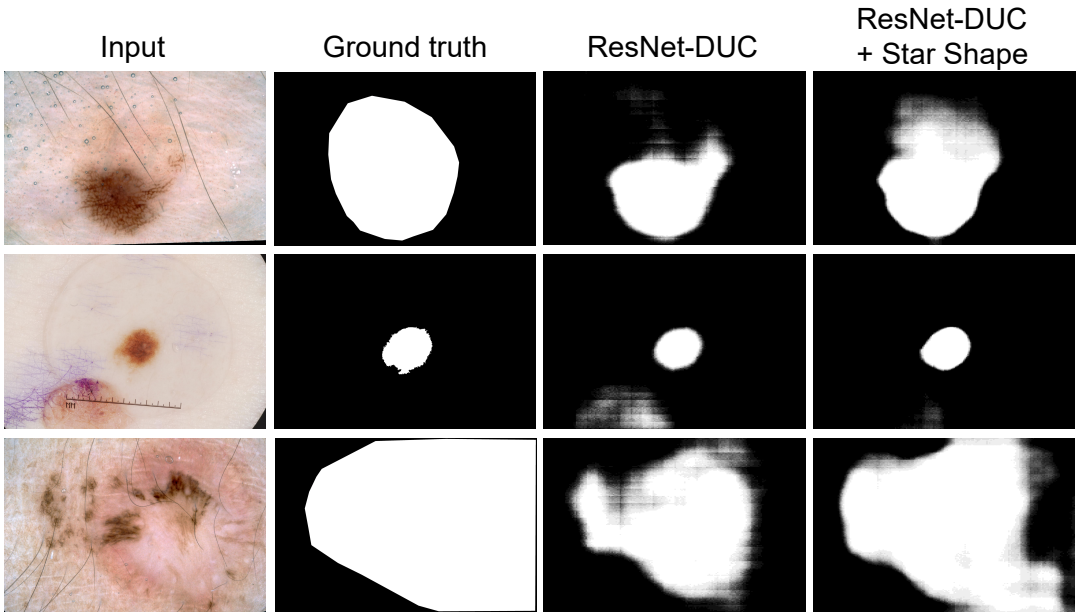


Figure 5.4: Qualitative comparison of ResNet-DUC architecture results with and without star shape prior.

5.4 Conclusion

We encoded the star shape prior in the loss function of an end-to-end trainable fully convolutional network to generate more accurate and plausible skin lesion segmentations. In contrast to energy minimization approaches, our proposed framework does not require computationally expensive optimization at inference time nor a user-defined object centre. Our experiments indicated that leveraging the prior knowledge in fully convolutional networks

yield convergence to an improved output space. In future works, we will extend to other prior information including but not limited to anatomically meaningful priors in fully convolutional networks trained for other 2D and 3D medical imaging applications. It is also interesting to study whether the violation of star shape prior in the skin lesions is a biomarker affecting the diagnosis of skin lesions.

Chapter 6

Learning to Segment Skin Lesions from Noisy Annotations

6.1 Introduction

Deep convolutional neural networks have driven substantial advancements in the automatic understanding of images. Requiring a large collection of images and their associated annotations is one of the main bottlenecks limiting the adoption of deep networks. In the task of medical image segmentation, requiring pixel-level semantic annotations performed by human experts exacerbate this difficulty. This Chapter proposes a new framework to train a fully convolutional segmentation network from a large set of cheap unreliable annotations and a small set of expert-level clean annotations. We declare a pixel-level annotation as 'noisy' when either:
(i) a healthy skin pixel is erroneously labeled as lesion i.e. 'false positive' or,
(ii) a skin lesion pixel is erroneously labeled as healthy skin i.e. 'false negative'.
We propose a spatially adaptive reweighting approach to treat clean and noisy pixel-level annotations commensurately in the loss function. We deploy a meta-learning approach to assign higher importance to pixels whose loss gradient direction is closer to those of clean data.

6.1.1 Robust to Noise Models

Despite the success of the FCN-based methods, they all assume that reliable ground truth annotations are abundant, which is not always the case in practice, not only because collecting pixel-level annotation is time-consuming, but also since human-annotations are inherently noisy. Further, annotations suffer from inter/intra-observer variation even among experts as the boundary of the lesion is often ambiguous. On the other hand, as the high capacity of deep neural networks (DNN) enable them to memorize a random labeling of training data [381], DNNs are potentially exposed to overfitting to noisy labels. Therefore, treating the annotations as completely accurate and reliable may lead to biased models with

weak generalization ability. This motivates the need for constructing models that are more robust to label noise.

Previous works on learning a deep classification model from noisy labels can be categorized into two groups. Firstly, various methods were proposed to model the label noise, together with learning a discriminative neural network. For example, probabilistic graphical models were used to discover the relation between data, clean labels and noisy labels, with the clean labels treated as latent variables related to the observed noisy label [364, 339]. Sukhbaatar et al. [316] and Goldberger et al. [129] incorporated an additional layer in the network dedicated to learning the noise distribution. Veit et al. [342] proposed a multi-task network to learn a mapping from noisy to clean annotations as well as learning a classifier fine-tuned on the clean set and the full dataset with reduced noise.

Instead of learning the noise model, the second group of methods concentrates on reweighting the loss function. Jiang et al. [175] utilized a long short-term memory (LSTM) to predict sample weights given a sequence of their cost values. Wang et al. [358] designed an iterative learning approach composed of a noisy label detection module and a discriminative feature learning module, combined with a reweighting module on the softmax loss to emphasize the learning from clean labels and reduce the influence of noisy labels. Recently, a more elaborate reweighting method based on a meta-learning algorithm was proposed to assign weights to classification samples based on their gradient direction [281]. A small set of clean data is leveraged in this reweighting strategy to evaluate the noisy samples gradient direction and assign more weights to sample whose gradient is closer to that of the clean dataset.

6.1.2 Contributions

In this work, we aim to extend the idea of example reweighting [281] explored previously for the classification problem to the task of pixel-level segmentation. We propose the first deep robust network to target the segmentation task by considering the spatial variations in the quality of pixel-level annotations. We learn spatially adaptive weight maps associated with training images and adjust the contribution of each pixel in the optimization of deep network. The importance weights are assigned to pixels based on the pixel-wise loss gradient directions. A meta-learning approach is integrated at every training iteration to approximate the optimal weight maps of the current batch based on the CE loss on a small set of skin lesion images annotated by experts. Learning the deep skin lesion segmentation network and spatially adaptive weight maps are performed in an end-to-end manner. Our experiments show how efficient leveraging of a small clean dataset makes a deep segmentation network robust to annotation noise.

6.2 Methodology

Our goal is to leverage a combination of a small set of expensive expert-level annotations as well as a large set of unreliable noisy annotations, acquired from, e.g., novice dermatologists or crowdsourcing platforms, into the learning of a fully convolutional segmentation network.

6.2.1 FCN's Average Loss

In the setting of supervised learning, with the assumption of the availability of high-quality clean annotations for a large dataset of N images and their corresponding pixel-wise segmentation maps, $\mathcal{D} : \{(X(i), Y(i));$
$i = 1, 2, \ldots, N\}$, parameters θ of a fully convolutional segmentation network are learned by minimizing the negative log-likelihood of the generated segmentation probability maps in the cost function $\mathcal{L}$:

$$\mathcal{L}(X, Y; \theta) = -\frac{1}{N}\Sigma_{i=1}^{N}\frac{1}{P}\Sigma_{p\in\Omega_i} y_p \log Pr(y_p|x_p; \theta) \tag{6.1}$$

where P is the number of pixels in an image, Ω_i is the pixel space of image i, x_p and y_p refer, in order, to the image pixel p and its ground truth label, and Pr is the predicted probability. As the same level of trust in the pixel-level annotations of this clean training data annotations is assumed, the final value of the loss function is averaged equally over all pixels of the training images.

6.2.2 FCN's Weighted Loss

As opposed to the fully supervised setting, when the presence of noise in most training data annotations is inevitable while only a limited amount of data can be verified by human experts, our training data comprises of two sets: $\mathcal{D}^c : \{(X^c(i), Y^c(i)); i = 1, 2, \ldots, K\}$ with verified clean labels and $\mathcal{D}^n : \{(X^n(i), Y^n(i)); i = 1, 2, \ldots, M \gg K\}$ with unverified noisy labels. We also assume that $\mathcal{D}^c \subset \mathcal{D}^n$. Correspondingly, we have two losses, $\mathcal{L}^c$ and $\mathcal{L}^n$. Whereas $\mathcal{L}^c$ has equal weighting, $\mathcal{L}^n$ penalizes a log-likelihood of the predicted pixel probabilities but *weighted* based on the amount of noise:

$$\mathcal{L}^c(X^c(i), Y^c(i); \theta) = -\frac{1}{P}\Sigma_{p\in\Omega_i} y_p^c \log Pr(y_p^c|x_p^c; \theta), \tag{6.2}$$

$$\mathcal{L}^n(X^n(i), Y^n(i); \theta, W(i)) = -\Sigma_{p\in\Omega_i} y_p^n w_{ip} \log Pr(y_p^n|x_p^n; \theta) \tag{6.3}$$

where w_{ip} is the weight associated with pixel p of image i. All the weights of the P pixels of image i are collected in a spatially adaptive weight map $W(i) = \{w_{i1}, \ldots, w_{ip}, \ldots, w_{iP}\}$, and weight maps associated with all M noisy training images X^n are collected in $W = \{W(1), \ldots, W(M)\}$.

6.2.3 Model Optimization

The deep noise-robust network parameters θ are now found by optimizing the weighted objective function $\mathcal{L}^n$ (as opposed to equal weighting in (6.1)) on the noisy annotated data $\mathcal{D}^n$, as follows:

$$\theta^* = \arg\min_{\theta} \Sigma_{i=1}^{M} \mathcal{L}^n(X^n(i), Y^n(i); \theta, W(i)). \tag{6.4}$$

6.2.4 Optimal Spatially Adaptive Weights

The optimal value of unknown parameters W is achieved by minimizing the expectation of negative log-likelihoods in the meta-objective function $\mathcal{L}^c$ over the clean training data $\mathcal{D}^c$:

$$W^* = \arg\min_{W,\ W \geqslant 0} \frac{1}{K} \Sigma_{i=1}^{K} \mathcal{L}^c(X^c(i), Y^c(i); \theta^*(W)). \tag{6.5}$$

6.2.5 Efficient Meta-training

Solving (7.1) to optimize the spatially adaptive weight maps W for each update step of the network parameter θ in (6.4) is inefficient. Instead, an online meta-learning approach is utilized to approximate W for every gradient descent step involved in optimizing θ (6.4). At every update step t of θ (6.4), we pass a mini-batch b_n of noisy data forward through the network and then compute one gradient descent step toward the minimization of $\mathcal{L}^n$:

$$\hat{\theta} = \theta_t - \alpha \nabla_{\theta} \Sigma_{i=1}^{|b_n|} \mathcal{L}^n(X^n(i), Y^n(i); \theta_t, W_0(i)) \tag{6.6}$$

where α is the gradient descent learning rate and W_0 in the initial spatial weight maps set to zero. Next, a mini-batch b_c of clean data is fed forwarded through the network with parameters $\hat{\theta}$ and the gradient of $\mathcal{L}^c$ with respect to the current batch weight maps $W^B = \{W(1), \ldots, W(|b_n|)\}$ is computed. We then take a single step toward the minimization of $\mathcal{L}^c$, as per (7.1), and pass the output to a rectifier function as follows:

$$U^B = W_0^B \Big|_{W_0^B = \mathbf{0}} - \eta \nabla_{W^B} \frac{1}{|b_c|} \Sigma_{i=1}^{|b_c|} \mathcal{L}^c(X^c(i), Y^c(i); \hat{\theta}(W)), \tag{6.7}$$

$$W^B = g(\max(\mathbf{0}, U^B)). \tag{6.8}$$

where η is a gradient descent learning rate, max is an element-wise max and g is the normalization function. Following the average loss over a mini-batch samples in training a deep network, g normalizes the learned weight maps such that $\Sigma_{i=1}^{|b_n|} \Sigma_{p \in \Omega_i} w_{ip} = 1$.

Equations (6.7) and (6.8) clarify how the learned weight maps prevents penalizing the pixels whose gradient direction is not similar to the direction of gradient on the clean data. A negative element u_{ip} in U (associated with pixel p of image i) implies a positive gradient $\nabla_{w_{ip}}\mathcal{L}^c$ in (6.7), meaning that increasing the assigned weight to pixel p, w_{ip}, increases the $\mathcal{L}^c$ loss value on clean data. So by rectifying the values of u_{ip} in (6.8), we assign zero weights w_{ip} to pixel p and prevent penalizing it in the loss function. In addition, the rectify function makes the $\mathcal{L}^n$ loss non-negative (cf. (6.3)) and results in more stable optimization.

Once the learning of spatially adaptive weight maps is performed, a final backward pass is needed to minimize the reweighted objective function and update the network parameters from θ_t to θ_{t+1}:

$$\theta_{t+1} = \theta_t - \alpha\nabla_{\theta_t}\Sigma_{i=1}^{|b_n|}\mathcal{L}^n(X^n(i);\theta_t, W^B). \tag{6.9}$$

6.3 Experiments and Discussion

6.3.1 Data Description

We validated our spatially adaptive reweighting approach on data provided by the International Skin Imaging Collaboration (ISIC) in 2017 [89]. The data consists of 2000 training, 150 validation and 600 test images with their corresponding segmentation masks. The same split of validation and test data are deployed for setting the hyper-parameters and reporting the final results. We re-sized all images to 96×96 pixels and normalized each RGB channel with the per channel mean and standard deviation of training data.

To create noisy ground truth annotations, we consider a lesion boundary as a closed polygon and simplify it by reducing its number of vertices: Less important vertices are discarded first, where the importance of each vertex is proportional to the acuteness of the angle formed by the two adjacent polygon line segments and their length. 7-vertex, 3-vertex and 4-axis-aligned-vertex polygons are generated to represent different levels of annotation noise for our experiments. To simulate an unsupervised setting, as an extreme level of noise, we automatically generated segmentation maps that cover the whole image (excluding a thin band around the image perimeter). Fig. 6.1 shows a sample lesion image and its associated ground truth as well as generated noisy annotations.

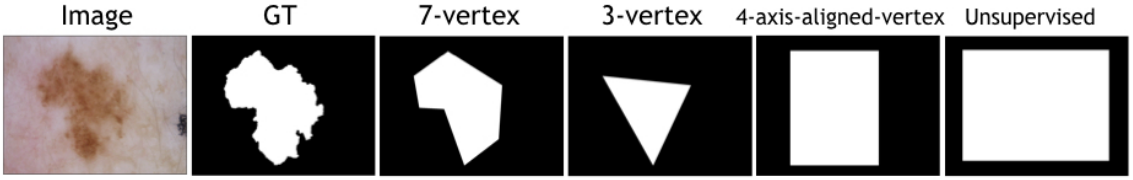


Figure 6.1: A skin image and its clean and various noisy segmentation maps.

6.3.2 Implementation

We utilize PyTorch framework to implement our segmentation reweighting network. We adopt the architecture of fully convolutional network U-Net [287] initialized by a random Gaussian distribution. We use SGD for learning the network parameters from scratch as well as the spatial weight maps over the mini-batch of sizes $|b_n| = 2$ and $|b_c| = 10$. We set the initial learning rate for both α and η to 10^{-4} and divide by 10 when the validation performance stops improving. We set the momentum and weight decay to 0.99 and 5×10^5, respectively. Training the deep reweighting network took three days on our 12 GB GPU memory.

6.3.3 Spatially Adaptive Reweighting vs. Image Reweighting and Fine-tuning

We compare our work with previous work on noisy labels which assign a weight per training images [281]. In addition, one popular way of training a deep network when a small set of clean data as well as a large set of noisy data are available is to pre-train the network on the noisy dataset and then fine-tune it using the clean dataset. By learning the spatially adaptive weight maps proposed in this work, we expect to leverage clean annotations more effectively for segmentation task and achieve an improved performance. We start with $|\mathcal{D}^n| = 2000$ images annotated by 3-vertex polygons and gradually replace some of the noisy annotation with expert-level clean annotations, i.e., increase $|\mathcal{D}^c|$. We report the Dice score on the test set in Fig. 6.2. The first (leftmost) point on the fine-tuning curve indicates the result of U-Net when all annotation are noisy and the last point corresponds to a fully-supervised U-Net. When all annotation are either clean or noisy, training the reweighting networks are not applicable. We observe a consistent improvement in the test Dice score when the proposed reweighting algorithm is deployed. In particular, a bigger boost in improvement when the size of the clean annotation is smaller signifies our method's ability to effectively utilize even a handful of clean samples.

6.3.4 Size of the Clean Dataset

Fig. 6.2 shows the effect of the clean data size, $|\mathcal{D}^c|$, on the spatial reweighting network performance. Our results show leveraging just 10 clean annotations in the proposed model improves the test Dice score by 21.79% in comparison to training U-Net on all noisy annotations. Also, utilizing 50 clean annotations in the spatial reweighting algorithm achieves a test Dice score ($\sim$80%) almost equal to that of the fully supervised approach. With only $\sim$100 clean image annotations, the spatial reweighting method outperforms the fully-supervised with 2000 clean annotations. Incrementing $|\mathcal{D}^c|$ from 50 to 1990, the reweighting approach

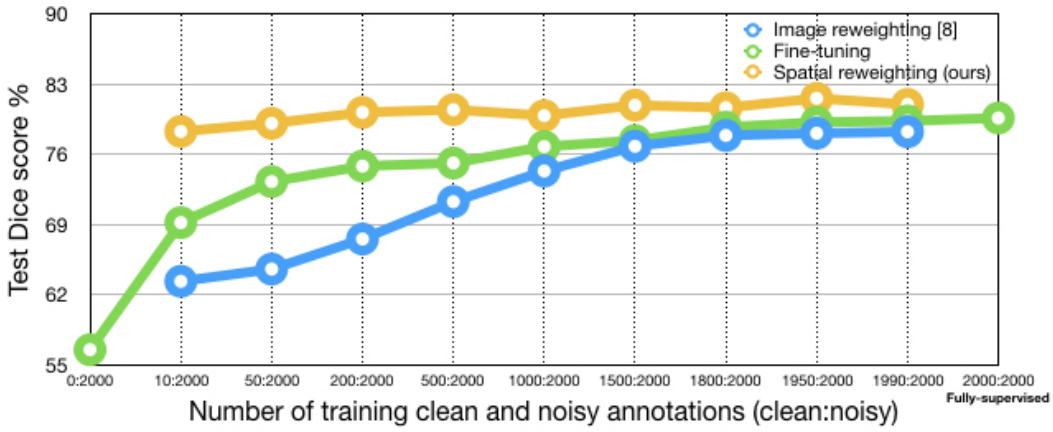


Figure 6.2: Test Dice score comparison for fine-tuning, per image reweighting [281] and, spatially adaptive reweighting (ours) models.

Table 6.1: Dice score using fine-tuning and reweighting methods for various noise levels.

	noise type	fine-tuning	proposed reweighting
A	no noise (fully-supervised)	78.63%	not applicable
B	7-vertex	76.12%	80.72%
C	4-axis-aligned-vertex	75.04%	80.29%
D	3-vertex	73.02%	79.45%
E	maximal (unsupervised)	70.45%	73.55%

improves the test Dice score by about 2%, questioning whether a 2% increase in accuracy is worth the ∼40-fold increase in annotation effort. Outperforming the supervised setting using spatial reweighting algorithm suggests that the adaptive loss reweighting strategy works like a regularizer and improves the generalization ability of the deep network.

6.3.5 Robustness to Noise

In our next experiment, we examine how the level of noise in the training data affect the performance of the spatial reweighting network in comparison to fine-tuning. We utilized four sets of (i) 7-vertex; (ii) 3-vertex; (iii) 4-axis-aligned-vertex simplified polygons as segmentation maps; and (iv) unsupervised coarse segmentation masks where each set corresponds to a level of annotation noise (Fig. 6.1). Setting $|\mathcal{D}^c| = 100$ and $|\mathcal{D}^n| = 1600$, the segmentation Dice score of test images for reweighting and fine-tuning approaches are reported in Table 6.1. We observe that deploying the proposed reweighting algorithm for 3-vertex annotations outperforms learning from accurate delineation without reweighting. Also, increasing the level of noise, from 7-vertex to 3-vertex polygon masks in noisy data, results in

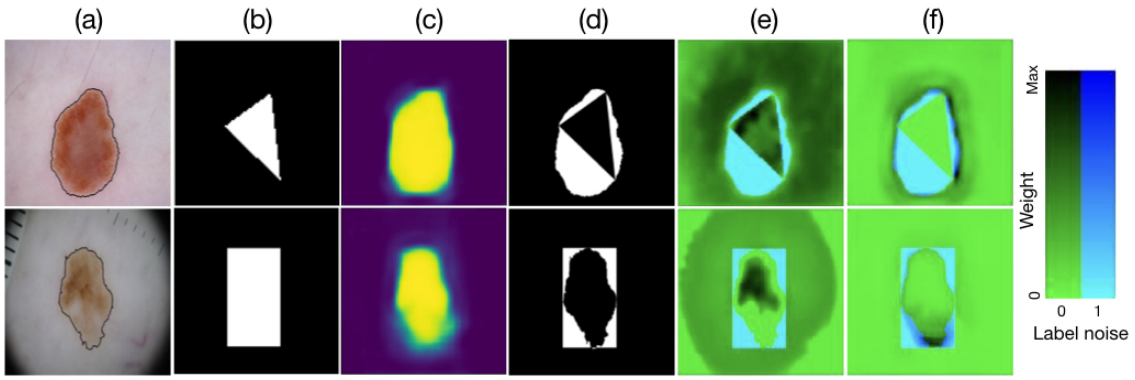


Figure 6.3: (a) Sample skin images and expert lesion delineations (thin black contour), (b) noisy ground truth, (c) network output, (d) the erroneously labelled pixels (i.e. noisy pixels) and learned weight maps in iterations (e) 1K and (f) 100K overlaid over the noisy pixel masks using the following coloring scheme: Noisy pixels are rendered via the blue channel: mislabelled pixels are blue, and weights via the green channel: the *lower* the weight the greener the rendering. The cyan color is produced when mixing green and blue, i.e. when low weights (green) are assigned to mislabelled pixels (blue). Note how the cyan very closely matches (d), i.e. mislabelled pixels are ca. null-weighted.

just ∼1% Dice score drop when deploying reweighting compared to ∼3% drop in fine-tuning.

6.3.6 Qualitative Results

To examine the spatially adaptive weights more closely, for some sample images, we overlay the learned weight maps, in training iterations 1K and 100K, over the incorrectly annotated pixels mask (Fig. 6.3). To avoid overfitting to annotation noise, we expect the meta-learning step to assign zero weights to noisy pixels (the white pixels in Fig. 6.3-(d)). Looking into Fig. 6.3-(e,f) confirms that the model consistently learns to assign zero (or very close to zero) weights to noisy annotated pixels (cyan pixels), which ultimatly results in the prediction of the segmentation maps in Fig. 6.3-(c) that, qualitatively, closely resemble the unseen expert delineated contours shown in Fig. 6.3-(a).

6.4 Conclusion

By learning a spatially-adaptive map to perform pixel-wise weighting of a segmentation loss, we were able to effectively leverage a limited amount of cleanly annotated data in training a deep segmentation network that is robust to annotation noise. We demonstrated, on a skin lesion image dataset, that our method can greatly reduce the requirement for careful labelling of images without sacrificing segmentation accuracy. Our reweighting segmentation network is trained end-to-end, can be combined with any segmentation network architecture, and does not require any additional hyper-parameter tuning.

Chapter 7

Deep Learning Ensembles from Potentially Contradictory Multiple Annotations

7.1 Introduction

Medical image segmentation annotations suffer from inter- and intra-observer variations even among experts due to intrinsic differences in human annotators and ambiguous boundaries. Leveraging a collection of annotators' opinions for an image is an interesting way of estimating a gold standard. Although training deep models in a supervised setting with a single annotation per image has been extensively studied, generalizing their training to work with datasets containing multiple annotations per image remains a fairly unexplored problem. In this Chapter, we propose an approach to handle annotators' disagreements when training a deep model. We utilized the proposed robust to noise segmentation model from Chapter 6 to model experts' knowledge independently while utilizing all available annotations. Two binary segmentation masks are called contradictory if they differ in at least one pixel-level annotation. An ensemble of Bayesian fully convolutional networks (FCNs) for the segmentation task is proposed by considering two major factors in the aggregation of multiple ground truth annotations: (1) handling contradictory annotations in the training data originating from inter-annotator disagreements and (2) improving confidence calibration through the fusion of base models' predictions.

7.1.1 Supervised Semantic Segmentation and Annotation Limitations

The majority of deep learning-based semantic segmentation models, however, rely on supervised learning of dense pixel annotations for the labels in images. State of the art supervised learning algorithms rely upon training using large volumes of data to yield acceptable results, and previous work has shown the importance of sufficient annotated data for visual tasks [256, 158, 317]. Particularly, Sun et al. [317] showed that the performance of segmen-

tation models in terms of overlap based measures exhibits a logarithmic relationship with the amount of training data used for representation learning for semantic segmentation.

Collecting ground truth annotations for semantic segmentation is considerably more expensive than doing so for other visual tasks such as classification and object detection because of the dense annotations involved. While this can partly be ameliorated by crowdsourcing the annotation process to non-experts, the presence of multiple object classes in a scene, coupled with factors such as illumination, shading, and occlusion, makes delineating the exact object boundaries an ambiguous and tedious task, leading to inter-annotator disagreements. The presence of multiple annotations (Fig. 7.1) further leads to the challenge of deciding upon an ideal ground truth against which the model's performance is assessed. Moreover, there exists a tradeoff between the precision and the generalizability of an 'ideal' segmentation ground truth, since aoverly precise delineation may not be reflective of the typical uncertainty encountered in practice when localizing the boundary [359]. A similar trade-off exists between the quality and the efficiency of these annotations: High quality dense annotations, although useful, take up more time to collect than relatively less informative approximate annotations (e.g., bounding boxes or simplified polygons). These problems are exacerbated further for medical images since medical imaging datasets with accurate pixel-level annotations are much smaller than their natural image counterparts [318], which can be attributed to the high cost associated with expert annotations, the difficulty in quantifying a true reference standard, the laborious nature of making dense annotations, which is even more difficult for 3D medical image volumes, and patient data privacy concerns. To add to this, the manual annotation of anatomical regions of interest can be very subjective and presents considerable inter- and intra-annotator disagreements even amongst experts across multiple medical imaging modalities [347, 119, 320, 283, 135], making it difficult to converge on a single gold standard annotation for model training and evaluation.

7.1.2 Related Works

One of the seminal works on comparing a segmentation model's performance by comparing against a collection of (human-annotated) segmentations is that proposed by Warfield et al. [359], where they proposed an expectation maximization algorithm for the simultaneous truth and performance level estimation (STAPLE). Given a collection of segmentation masks, STAPLE generates a probabilistic estimate of the true segmentation mask as well as the segmentation performance of each of the segmentations in the collection. This was followed by several other extensions of STAPLE which addressed its limitations such as susceptibilities to large variations in inter-annotator uncertainty and annotator performance [52, 181, 208, 217].

More recently, Mirikharaji et al. [243] showed that leveraging different levels of annotation reliability, using spatially-adaptive reweighting while learning deep learning based segmentation model parameters, helps improve performance, and demonstrated superior

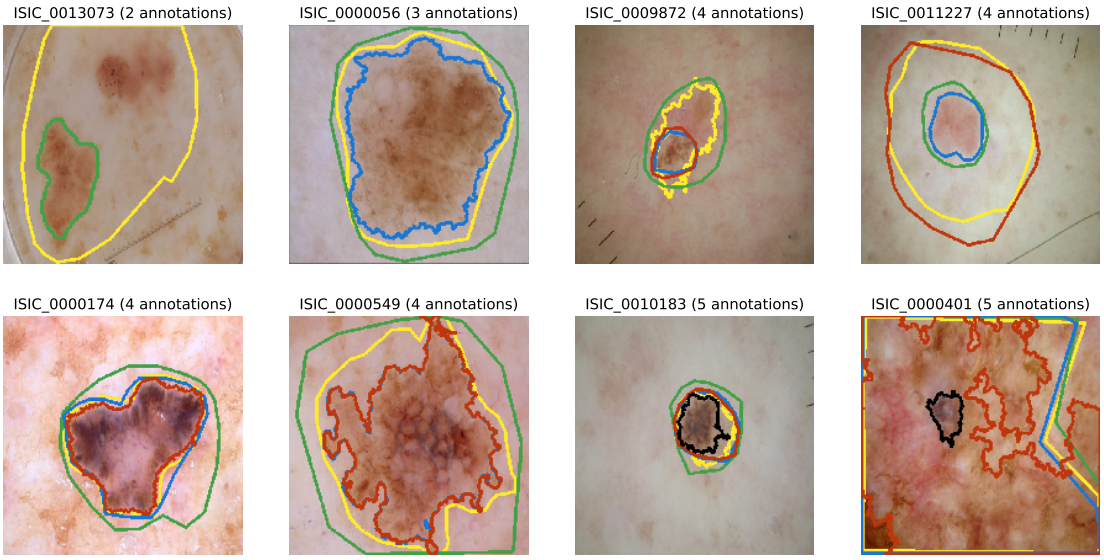


Figure 7.1: Sample skin lesion images from the ISIC Archive which contain multiple lesion boundary annotations (denoted by different colors).

segmentation accuracy using a large number of low quality, 'noisy' annotations along with only a small fraction of precise annotations. Hu et al. [156] used a modified probabilistic U-Net [198] model to generate quantifiable aleatoric and epistemic uncertainty estimates for segmentation using a supervised learning framework which modeled inter-annotator variability as aleatoric uncertainty ground truth. Ribeiro et al. [283] proposed an approach to improve inter-annotator agreement by conditioning the segmentation masks using morphological image processing operations (opening and closing), convex hulls and bounding boxes to remove details specific to any single particular annotator. They argue that the conditioning could be deemed as denoising operations, removing the annotator specific details from the segmentation masks. The same authors then proposed to train their segmentation model on a subset of the images, derived by filtering out all samples whose mean pairwise Cohen's kappa score was less than 0.5, thus using only those segmentations which largely agree between annotators [284].

7.1.3 Predictive Uncertainty

Despite the obvious benefits of improving segmentation performance, it is also crucial to analyze the predictive uncertainty of deep networks in medical image segmentation. In machine learning, the uncertainty has been classified into aleatoric and epistemic types. The aleatoric, which reflects the inherent noise in the data, has been estimated using a second auxiliary output in the network [193]. Bayesian neural networks (BNNs) have adopted Monte Carlo (MC) dropout [121] to reflect the epistemic uncertainty associated with the net-

work parameters. Thanks to their simplicity, MC dropout uncertainty estimation has been studied in the context of general semantic segmentation [192] as well as medical image segmentation [205, 304]. However, the uncertainty estimates obtained using MC dropout tend to be miscalibrated, i.e., they do not correspond well with the model error [209]. Recently, there have been efforts to improve the uncertainty calibration using ensemble learning. Particularly, Lakshminarayanan et al. [206] demonstrated the advantage of ensemble learning, i.e., averaging a collection of models trained from different initializations, in yielding more accurate predictive uncertainty estimates for classification and regression tasks. Mehrtash et al. [236] studied the performance of ensemble learning for predictive uncertainty in medical image segmentation. Particular to skin lesion segmentation, Jungo et al. [179] thoroughly studied the reliability of existing uncertainty estimation methods and showed their benefits and limitations [179].

7.1.4 Contribution Claims

Deep neural networks have been shown to potentially overfit to noisy labels [381] and our motivation for this work is to avoid single annotator bias [207]. Therefore, we seek training deep segmentation models to learn from multiple annotations as available instead of discarding some annotations. Rather than selecting a subset of images to learn from Ribeiro et al. [284], we instead propose a generalized approach of annotation weighting by leveraging different groups of consistent annotations in an ensemble method towards efficiently learning from all available annotations. We also utilize uncertainty estimates [193, 206] in an ensemble learning framework to improve predictive uncertainty and calibration confidence in the final prediction.

We consider two major factors in the aggregation of multiple ground truth annotations: (1) handling contradictory annotations in the training data originating from inter-annotator disagreements, and (2) improving the model's confidence calibration through deep ensembling. Our hypothesis is that given a new image, leveraging different experts' skills independently and fusing them in an ensemble model, while considering their estimated uncertainty, makes for a more reliable final prediction.

7.2 Method

7.2.1 Problem Statement and Method Overview

Let $\mathcal{X} = \{X_n\}_1^N$ and $\mathcal{Y} = \{Y_n\}_1^N$ be a set of N images and segmentation ground truth masks, respectively. In a supervised learning scheme, a network is trained to learn a function $f_\theta : X_n \mapsto \hat{Y}_n$ parameterized by θ, which maps an image X_n to the corresponding estimated segmentation mask $\hat{Y}_n$. Approximating the mapping function f_θ using a single annotation per image has been well studied in the literature. However, training supervised models in the presence of multiple annotations remains largely unexplored.

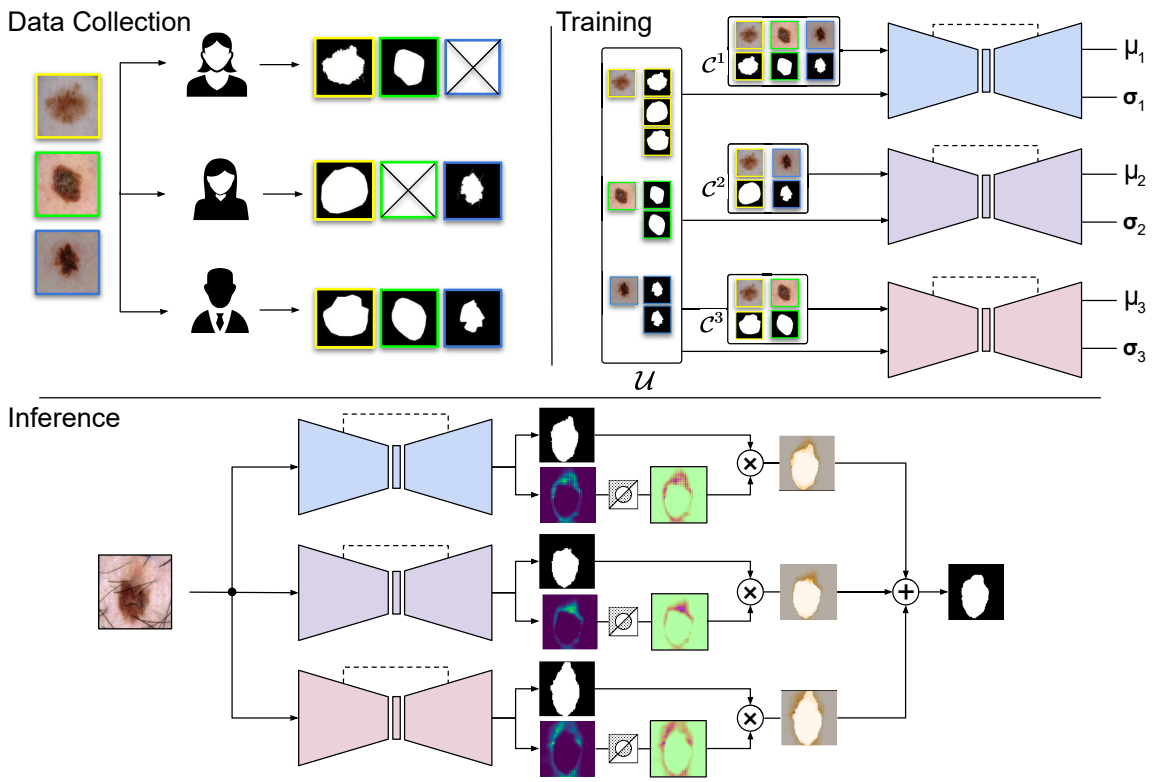


Figure 7.2: An overview of our proposed framework for skin lesion segmentation with multiple annotations. (top left) Multiple users annotating different, potentially overlapping, subsets of the original data. (top right) Each set of non-contradictory labels is considered as ground truth and, along with the remaining annotations that are deemed potentially noisy, are used to train a different base model. (bottom) At inference, each base model's prediction, along with its estimated aleatoric uncertainty maps are fused to obtain the final prediction.

Let us assume that K annotators have independently annotated different subsets of the images resulting in a set of segmentation ground truths $\mathcal{Y} = \{\{Y_{mn}\}_{m=1}^{M_n}\}_{n=1}^{N}$, where M_n denotes the number of available annotations for X_n. Inconsistent annotations for a given image could mislead the network and substantially deteriorate the performance of the model. Let M indicate the maximum number of annotations per image over the entire dataset. Instead of aggregating multiple annotations to estimate a single ground truth before the training phase, we propose to (1) learn a set of M mapping functions $\mathcal{F} = \{f_{\theta^i}\}$ through ensembling M base deep models trained over the union of available annotations and (2) minimize the confusion induced from observing multiple annotations through a spatial re-weighting scheme during training. (3) Lastly, we demonstrate that our proposed ensemble learning framework not only improves the segmentation performance but also provides a well-calibrated predictive uncertainty. Fig. 7.2 illustrates the overview of our ensemble learning framework for skin lesion segmentation with multiple annotations.

7.2.2 Detailed Method

Non-contradictory Subsets Selection: To handle contradictory annotations arising from having multiple annotations per image during the training, we partition the entire dataset into M disjoint subsets, denoted by $\{\mathcal{C}^i\}_{i=1}^{M}$, such that each $\mathcal{C}^i$ includes at most one unique annotation for every image. In particular, for each image, with $M_n \leq M$ annotations, we randomly assign the M_n annotations to $\{\mathcal{C}^i\}_{i=1}^{M_n}$ subsets.

A naïve approach is to utilize these disjoint subsets to train individual base models independently. Even though this solution prevents exposing each ensemble base model to multiple annotations per image and encourages a diverse set of model performance, however, each disjoint set includes a small number of training samples which can adversely affect the generalization capability of individual base models. To address this issue, we combine all images along with all available annotations into a *union* dataset, denoted as $\mathcal{U}$, and use it to train M base networks. Following Mirikharaji et al. [243], we utilize these non-contradictory subsets to assess the quality of annotations in $\mathcal{U}$. Specifically, spatially-adaptive weight maps associated with varying annotations in $\mathcal{U}$ are learned to adjust the contribution of each annotated pixel in the optimization of deep network based on its consistency with clean annotations in $\{\mathcal{C}^i\}$.

Learning Models: In more details, for each base model i, $i \in 1, ..., M$, we define a cross entropy loss, denoted as $\mathcal{L} = \{L_{ce}^{\mathcal{C}^i}\}$ over each non-contradictory set $\mathcal{C}^i$. We then, in a meta-learning paradigm, learn a set of spatial weight maps $\mathcal{W}^i = \{\{W_{mn}^i\}_{m=1}^{M_n}\}_{n=1}^{N}$ for all annotations $\mathcal{U}$ based on the gradients of the cross entropy losses with respect to the weights maps, i.e. $\nabla_{W^i}\mathcal{L}_{ce}^{\mathcal{C}^i}$. This way, $\mathcal{W}^i$ is optimized to cancel out the contributions of annotations inconsistent with $\mathcal{C}^i$ while optimizing the parameters for i^{th} base network, i.e. θ^i. Mathematically:

$$\mathcal{W}^{i*} = \underset{\mathcal{W}^i,\ \mathcal{W}^i \geqslant 0}{\arg\min} \sum_{n \in \mathcal{C}^i} L_{ce}^n(\hat{Y}_n^i, Y_n; \theta^i(\mathcal{W}^i)). \tag{7.1}$$

Note that every image in $\mathcal{C}^i$ has only one ground truth. $\mathcal{W}^i$ are encoded in $\mathcal{L}$ and they are optimized along with the network parameters θ^i for each individual base model. By integrating the information in the optimized $\mathcal{W}^i$, we can determine the degree by which a pixel-level annotation from any of annotators is considered noisy for model i, depending on how similar this annotation is to the annotations in $\mathcal{C}^i$. Therefore:

$$\mathcal{L}(\hat{Y}_n^i, Y_{mn}; \theta^i, W_{mn}^i) = -\sum_{q \in X_n} W_{mnq}^i Y_{mnq} \log \hat{Y}_{nq}^i, \tag{7.2}$$

$$\hat{Y}_{nq}^i = \text{softmax}(U_{nq}^i). \tag{7.3}$$

Fusion of Predictions: Once the individual base models are trained, the final prediction of the entire ensemble for the X_n is obtained by using a weighted fusion [293], that is:

$$\hat{Y}_n = \sum_{i=1}^{M} \alpha_n^i \hat{Y}_n^i, \tag{7.4}$$

where α_n^i is the combination coefficient for prediction by model i. The simplest way to determine α_n^i is to consider equally weighted averaging and set them to $1/M$. Another popular technique is to set α_n^i coefficients according to the confidence of the model [303]. In this work, we explore both aggregation techniques in our experimental evaluations.

Uncertainty-driven Aggregation: For the uncertainty-driven aggregation of base models, we leverage aleatoric uncertainty, which models irreducible observation noise, to estimate how confident a base model is about its prediction, and utilize the confidence when combining the base models' prediction maps. Following Kendall et al. [193], we approximate the aleatoric uncertainty for each pixel $q \in X_n$ by placing a Gaussian distribution over the logit space before applying a sigmoid function in the last layer and reformulate the network output as:

$$U_{nq}^i \sim \mathcal{N}\left(f_{nq}^i, (\sigma_{nq}^i)^2\right), \tag{7.5}$$

where f_i and σ_i are the network i outputs.

We use the aleatoric uncertainty in two forms: (1) considering the pixel-wise uncertainty values as spatially-adaptive coefficients and (2) averaging the pixel-wise uncertainty into a single scalar image-level coefficient.

7.3 Experiments

7.3.1 Data

For training, we used the International Skin Imaging Collaboration (ISIC) Archive data [1, 89, 88], the largest dermoscopic public dataset with over 13,000 images, captured by diverse devices in international clinical centers. All images are 8-bit RGB color dermoscopy images. Similar to Ribeiro et al. [284], we utilized 2,223 images with more than one segmentation ground truth mask (2,094 with two, 100 with three and 36 with four and 3 with five) to train our models. We split all 2,223 images to 80% for training and 20% for validation. For model selection, we randomly selected which annotation to use in validation set. To create our non-contradictory annotation sets, all training data are randomly and uniformly partitioned into five groups of overlapping images but unique ground truth annotations. ISIC ground truth masks were generated using three different pipelines with different levels of border

irregularities all involving a dermatologist with expertise in dermoscopy: (1) an automatic algorithm followed by an expert review; (2) a semi-automatic algorithm controlled by an expert; and (3) manually drawing a polygon by an expert. A large variation of disagreement based on Cohen's kappa scores with the mean 0.67 is reported in Ribeiro et al. [283]. Fig. 7.1 shows some examples of skin lesion images with multiple lesion boundary annotations from this dataset.

To thoroughly assess the segmentation performance of our proposed ensemble framework, we leveraged three publicly available datasets in our evaluations. All the images in the used datasets are resized into 96×96 pixels and normalized using the per-channel mean and standard deviation across the entire dataset. A brief description of these test datasets are provided as follows:

- **ISIC**: Ribeiro et al. [284] randomly selected a test set of 2,000 images with just one segmentation ground truth from ISIC Archive. We used the exact set in our experimental evaluations for fair comparisons.
- **PH^2**: The PH^2 (Pedro Hispano Hospital) dataset contains 200 8-bit RGB color dermoscopic images [237]. All images are acquired under the same condition using Tuebinger Mole Analyzer system at $20\times$ magnification.
- **DermoFit**: This dataset has 1300 8-bit RGB color clinical images [33]. The images are captured with a Canon EOS 350D SLR camera at the same distance from the lesion under controlled lighting conditions.

7.3.2 Base Models and Implementation Details

Our architecture is an encoder-decoder architecture with residual and skip connections transferring the information in the encoder modules to the corresponding decoder modules [77]. Since the images in our training dataset are paired with at most five annotations ($M = 5$), our ensemble framework consists of five base deep neural networks. Each network outputs two spatial maps in the last layer: the dense segmentation prediction and the predicted aleatoric uncertainty map. In training the aleatoric loss, 10 Monte Carlo samples from logits are taken. SGD with an initial learning rate of 10^{-4} is used to optimize the network parameters. The batch size for optimizing the spatial weight maps and network parameters is 64 and 2. The momentum and weight decay are set to 0.99 and 5×10^5, respectively.

7.3.3 Results

Table 7.1 compares the segmentation performance of our baseline models as well as the individual base models, across different prediction fusion schemes, using the Jaccard index. To train the baseline model, for every image in the training batch, we randomly select which ground truth to use when optimizing the loss function (row A). While it is interesting to

Table 7.1: Comparing the segmentation performance based on Jaccard index reported in percent (% ± standard error) on three datasets.

	Method	ISIC Archive [1]	PH^2 [237]	DermoFit [33]
A	baseline	68.00 ± 0.56	81.30 ± 0.77	70.30 ± 0.54
B	model 0	69.22 ±0.53	82.82 ± 0.75	72.57 ± 0.50
C	model 1	69.75 ± 0.55	82.40 ± 0.75	71.05 ± 0.55
D	model 2	70.33 ± 0.52	83.46 ± 0.74	72.80 ± 0.51
E	model 3	70.37 ± 0.51	83.31 ± 0.70	73.04 ± 0.53
F	model 4	69.73 ± 0.52	82.29 ± 0.72	70.87 ± 0.48
G	equally weighted fusion (ours)	**72.11± 0.51**	84.96± 0.73	74.22± 0.51
H	pixel-level confidence (ours)	71.46± 0.49	84.52± 0.74	73.91± 0.53
I	image-level confidence (ours)	**72.08± 0.49**	**85.20 ± 0.70**	**74.33± 0.50**
J	less is more [284]	69.20	81.25	72.55

Table 7.2: Comparing predictive uncertainty based on negative log-likelihood (NLL) and Brier score (Br) on three datasets. Lower NLL and Br values correspond to a better predictive uncertainty estimate.

Dataset		ISIC Archive		PH^2		DermoFit	
Method		NLL	Br	NLL	Br	NLL	Br
A	MC dropout model 0	0.073	0.019	0.166	0.048	0.272	0.082
B	MC dropout model 1	0.075	0.020	0.151	0.044	0.310	0.099
C	MC dropout model 2	0.075	0.019	0.149	0.044	0.283	0.087
D	MC dropout model 3	0.078	0.020	0.152	0.042	0.291	0.091
E	MC dropout model 4	0.075	0.019	0.155	0.045	0.312	0.100
F	deep ensemble (ours)	**0.070**	**0.018**	**0.144**	**0.041**	**0.254**	**0.078**

consider each annotator separately and evaluate their performance, the assignments between annotators and ground truth are not stated in the ISIC Archive dataset. Instead, we evaluate the performance of each base model trained on non-contradictory annotations simulating an expert knowledge (rows B to F). In addition, we compare the performance of our proposed method against the work of Ribeiro et al. [284] where a subset of samples with small annotator disagreements is taken into account during the training.

For the fusion stage, we examine three approaches as listed below:

- **Uniformly weighted fusion**: The predictions from the base models are combined by averaging the output probabilities.

- **pixel-level confidence-based fusion**: The predictions from the models are fused using normalized confidence spatial maps computed by inverting the predicted aleatoric outputs.
- **Image-level confidence-based fusion**: The aleatoric uncertainty maps are aggregated into an image-level aleatoric scalars and the predictions of the base models are combined based on the image-level normalized confidence scalars computed by inverting the uncertainty scalars.

Our results demonstrate that leveraging all available annotations effectively in an ensemble framework consistently improves the performance of the segmentation performance both in a held-out test set and over two other distinct datasets. Looking into different variants of our deep ensemble method, it is evident that aggregating the aleatoric uncertainty into the image-level scalar and leveraging them in the fusion stage (row H) either outperforms or exhibits competitive performance against the uniform averaging scheme (row G).

While modeling predictive uncertainty in clinical applications without a 'real' gold standard is helpful in decision making, miscalibrated uncertainty with overconfident predictions leads to an unreliable outcome. To evaluate the calibration quality of our ensemble annotation aggregation against Bayesian FCNs, we implemented Bayesian epistemic uncertainty using dropout for each base model. Similar to Bayesian SegNet [192], we added five dropout layers in the central part of the encoder and the decoder after each convolutional layer. Dropout probability is set to 0.3 and they are kept active at the inference time. Fifteen feed-forwards are executed to perform MC sampling and the output mean is considered as the final segmentation prediction.

To evaluate the quality of the predictive uncertainty, we use two widely used metric in the literature [206, 121]; negative log-likelihood (NLL) and Brier score (Br). Given a segmentation network with sigmoid non-linearity in the output layer, NLL and Br for X_n are calculated as follows:

$$NLL = \frac{-1}{|X_n|} \sum_{q \in X_n} Y_{nq} \log \hat{Y}_{nq} + (1 - Y_{nq}) \log(1 - \hat{Y}_{nq}) \tag{7.6}$$

$$Br = \frac{1}{|X_n|} \sum_{q \in X_n} [Y_{nq} - \hat{Y}_{nq}]^2 \tag{7.7}$$

Consistent with prior studies on deep ensembling [206, 236], Table 7.2 indicates that our annotation aggregation ensemble with five base models consistently improves the confidence calibration and predictive uncertainty for three datasets in comparison to modeling epistemic uncertainty by MC dropout.

The spatially adaptive weight maps for model i, $\mathcal{W}^i$, are learned to prevent penalizing the pixels whose feature maps are similar to the feature maps of data in $\mathcal{C}^i$ while their gradient

direction is not similar to the direction of loss gradient on annotations in $\mathcal{C}^i$. To qualitatively evaluate matrices $\mathcal{W}^i$, in Figures 7.3 and 7.4, we overlay the learned weight maps, in training iteration 100K, over the inconsistency maps (absolute differences of ground truth masks). Looking into the color-coded boxes shows how the location of the cyan pixels matches the inconsistency maps (zero or very close to zero weights are assigned to inconsistent annotated pixels), which results in exclusively leveraging the experts knowledge in $\mathcal{C}^i$ when learning θ^i.

7.4 Conclusion

Approaches to train deep segmentation models do not trivially generalize to datasets with multiple image annotations. We propose an ensemble paradigm to deal with discrepancies in segmentation annotations. A robust-to-annotation-noise learning scheme is utilized to efficiently leverage the multiple experts' opinions toward learning from all available annotations and improve the generalization performance of deep segmentation models. The quality of predictive uncertainty in clinical applications without true gold standards is critical. Our model captures two types of uncertainty, aleatoric uncertainty modeled in the training loss function and epistemic uncertainty modeled in the ensemble framework to improve confidence calibration.

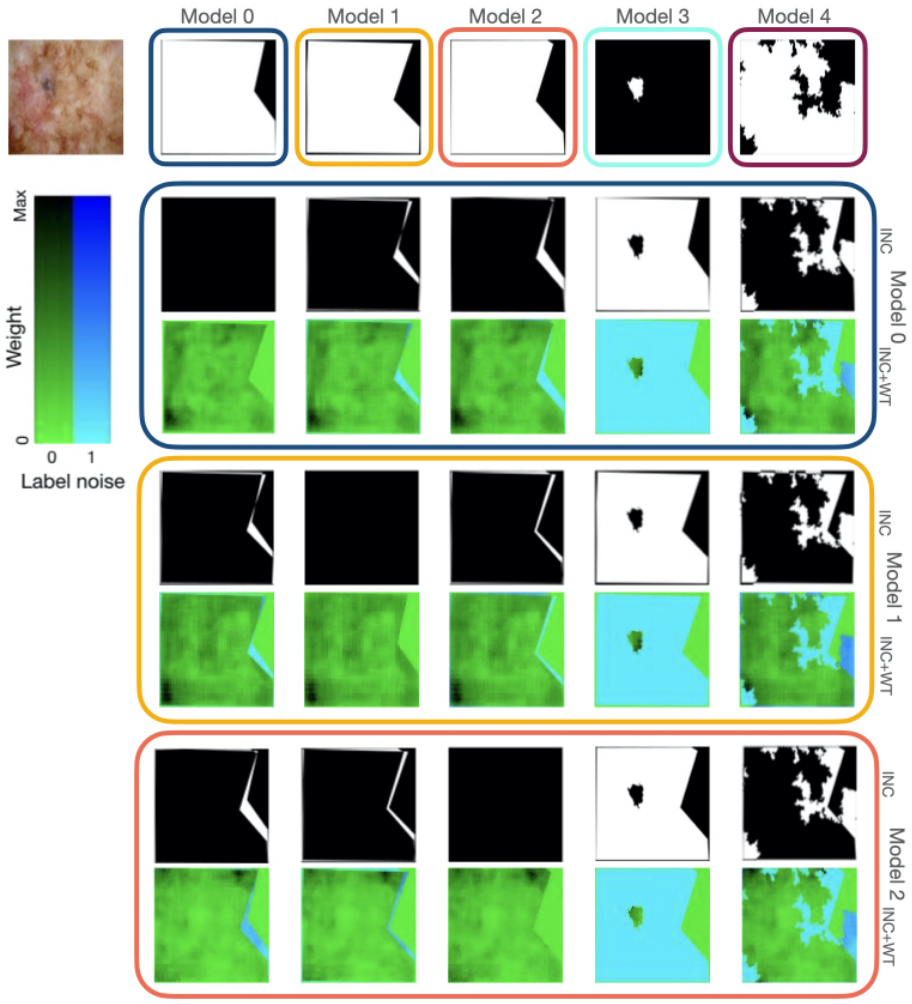


Figure 7.3: Qualitative evaluation of weighting matrices: (first row) a sample training image and trusted annotations in base models 0 to 4. (second row) inconsistency maps (INC) between the trusted ground truth in Model 0 and other ground truth annotations. (third row) learned weight maps in iteration 100K overlaid over the inconsistency maps (INC+WT). Color-coded boxes indicates the change when the trusted annotations in base models 0, 1 and 2 are different.

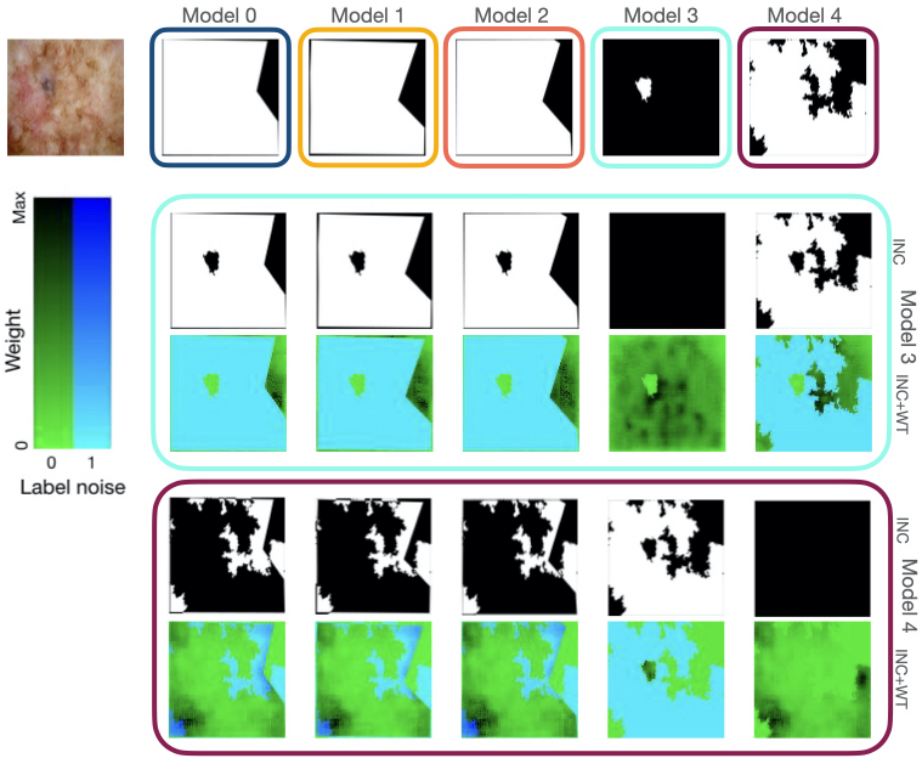


Figure 7.4: Qualitative evaluation of weighting matrices: (first row) a sample training image and trusted annotations in base models 0 to 4. (second row) inconsistency maps (INC) between the trusted ground truth in Model 3 and other ground truth annotations. (third row) learned weight maps in iteration 100K overlaid over the inconsistency maps (INC+WT). Color-coded boxes indicate the changes when the trusted annotations in base models 3 and 4 are different.

Chapter 8

Conclusions

8.1 Thesis Summary

Through the proposed approaches presented in the different Chapters of this thesis, we explored different ways to automate skin lesion segmentation task by advancing state-of-the-art deep models while considering the limitation of available data. Chapter 2 reviewed the literate utilizing the capability of deep learning models to segment skin lesions, discussed input data (datasets, preprocessing and synthetic data generation) (Section 2.2) and analyzed utilization of different architectural modules and losses (Section 2.3) as well as evaluation aspects (Section 2.4).

In the first part of the thesis, we leveraged the auxiliary information in the form of domain knowledge, contextual information, and labels consistency in deep segmentation models to regularize model parameters toward a more generalizable solution. In Chapter 3, we explicitly encoded the contextual information in the learning of the deep models' parameters by training a sequence of models. We also proposed to use degraded probability maps to avoid overfitting in subsequent models. Chapter 4 presented how to impose high-order consistency in predicted segmentation maps by utilizing a discriminator on the top of generative model in GANs. Chapter 5 proposed to encode high order shape prior knowledge in the form of a differentiable regularization term in the loss function, preserve global structures in the output space and, generate plausible skin lesion segmentation.

In addition to advancing the performance of segmentation models, in the second part of this thesis, we studied the limitations of ground truth pixels level annotations to effectively leverage limited reliable annotations. Chapter 6 introduced the first robust to noise deep network for segmentation task to reduce the requirement of careful labeling. Spatially adaptive reliability maps were learned in a meta-learning paradigm to treat noisy pixels based on a small set of reliable expert-level segmentation annotations. Finally, Chapter 7 discussed inter-annotator disagreements toward avoiding single annotator bias and proposed an ensemble paradigm modeling multiple experts' opinions toward learning effectively from all available annotations.

8.2 Future Directions

Following the contributions in this thesis applied to the challenging task of skin lesion segmentation, aiming at image per-pixel label prediction and performing relatively well on benchmark skin lesion datasets with minimal pre-processing requirements, there are still some substantial challenges to be addressed. In this section, we give a list of different directions and related questions to be further explored in future research.

- **Mobile dermoscopic image analysis:** There are various inexpensive dermoscopes designed for smartphones. Thus, mobile dermoscopic image analysis is of great interest worldwide, especially in regions where access to dermatologists is limited. Typical CNN-based image segmentation algorithms have millions of weights. In addition, classical CNN architectures are known to have difficulty dealing with certain image distortions such as noise and blur [109]. Therefore, the current dermoscopic image segmentation algorithms may not be ideal for execution on resource-constrained mobile devices. Leaner CNN architectures (e.g., MobileNet [154], ShuffleNet [386], EfficientNet [324], and MnasNet [323]) should be investigated in addition to the robustness of such architectures with respect to image noise and blur.

- **Image Sets:** To train more accurate and robust deep neural segmentation architectures, we need larger, more diverse, and more representative dermatological image sets with multiple manual segmentations per image.

- **Collecting Manual Segmentations:** At the time of this writing, the ISIC Archive contains over $69,000$ publicly available images. Considering that the largest public dermoscopic image set contained a little over $1,000$ images less than five years ago, we have come a long way. The more pressing problem now is the lack of manual segmentations for most of these images. Since manual segmentation by medical experts is laborious and costly, crowdsourcing techniques [201] could be explored to collect annotations from non-experts. Experts could then revise these initial annotations. Note that the utility of crowdsourcing in medical image annotation has been demonstrated in multiple studies [115, 143, 307, 128].

- **Ground truth Annotations:** The quality of dataset ground truths are affected by laborious and costly nature of pixel-wise annotations, ambiguous lesion boundaries, and inter- and intra-annotator disagreements amongst experts. Manual segmentations outlined by multiple experts must be approached as samples of an estimator about the true label, which can never be directly observed. Although in Chapter 7, contradictory annotations are defined in a binary way, having a fuzzy definition of contradictory help to encode different level of trust to available annotations.

- **Segmentation Fusion:** If the dermatological image set at hand contains multiple manual segmentations per image, one should consider fusing the manual segmentations using an algorithm such as STAPLE [359] (see Section 2.4). Such a fusion algorithm can also be used to build an ensemble of multiple automated segmentations.

- **Supervised Segmentation Evaluation Measures:** Supervised segmentation evaluation measures popular in the dermatological image analysis literature (see subsection 2.4.3) are often region-based, pair-counting measures. Other region-based measures, such as information-theoretic measures, as well as boundary-based measures [321] should be explored as well.

- **Unsupervised Segmentation and Unsupervised Segmentation Evaluation:** Current CNN-based dermatological image segmentation algorithms are all based on supervised deep learning, meaning that these algorithms require manual segmentations for training a CNN classifier. Nearly all of these segmentation studies employ supervised segmentation evaluation, meaning that they also require manual segmentations for testing. Due to the scarcity of annotated dematological images, it may be beneficial to investigate unsupervised deep learning [173] as well as unsupervised segmentation evaluation [75, 383].

- **Weakly-Supervised Annotation:** Supervised annotation (manual segmentation) is time-consuming. An alternative is to use weakly-supervised annotation to decrease the burden of pixel-level annotation and provide larger data sets. While more than 95% of deep skin lesion papers proposed fully-supervised models (Fig. 2.7), image-level annotations [361], point supervision [272, 39], scribbles [229] and bounding-box annotations [100, 259] are recently leveraged for segmentation tasks. These weakly-supervised annotations are more amenable to crowdsourcing as well, especially for non-experts, and can be effectively utilized to alleviate the need for object boundary delineation.

- **Statistical Significance Analysis:** Only a few of the prior studies in dermatological image segmentation (e.g., [116]) conducted a statistical significance analysis of their results.

- **Systematic evaluations:** Systematic evaluations that have been performed for skin-lesion diagnosis [340, 57, 265] are, so far, inexistent in the literature of skin-lesion segmentation. (Except for [15] — does it count? Any others?)

- **Fusion of Hand-Crafted and Deep Features:** Can we integrate the deep features extracted by CNNs and hand-crafted features synergistically?

- **Loss of Spatial Resolution:** The use of repeated subsampling in CNNs leads to coarse segmentations. Various approaches have been proposed to minimize the loss of

spatial resolution, including fractionally-strided convolution (or, deconvolution) [228], atrous (or dilated) convolution [79], and conditional random fields [202]. More research needs to be conducted to determine the most appropriate strategy for dermatological image segmentation.

- **Hyperparameter Tuning:** Compared to traditional machine learning classifiers (e.g., nearest neighbors, decision trees, and support vector machines), deep neural networks have a large number of hyperparameters related to their architecture, optimization, and regularization. An average CNN classifier has about a dozen or more hyperparameters [40] and tuning these hyperparameters systematically is a laborious undertaking. *Neural architecture search* is an active area of research [111] and some of these model selection approaches have already been applied to segmentation [225].
- **Limited research on clinical data**: Another limitation is the insufficient benchmark clinical skin lesion dataset with expert pixel-level annotations. Fig. 2.8 shows that while the number of dermoscopic images with ground truth segmentation masks has been increasing over the last few years, a few clinical data are available. In contrast to dermoscopic images requiring a special tool that is not always utilized even by dermatologists [112], clinical images captured by a digital camera and smartphones have the advantage of easy accessibility, which can be utilized to evaluate the priority of patients by their lesion severity level. Most of the deep skin lesion segmentation models are performed on dermoscopic images, leaving the need to develop automatic tools for non-specialists unattended.
- **Multi-class segmentation toward dermoscopy feature extraction:** Instead of extracting general appearance features on skin lesion images, salient properties of skin conditions (e.g. an atypical pigment network or irregular streaks) can be leveraged to improve the diagnosis of skin lesions. Identifying visual criteria associated with different skin conditions can be performed by not only classifying dermoscopic features but also localizing the regions of image containing those features. Dermoscopic feature extraction can be formulated as a multi-class segmentation [188] problem addressed by deep learning models.
- **Transferability of models across populations:** As the majority of skin lesion datasets are from fair-skinned people, the generalizability of deep models over racially diverse skin tons is questionable.

- **3D total-body skin images**: Instead of 2D skin images, 3D skin images from total body skin surface can be used for skin image analysis. Acquisition and analysis of a wider imaging field of view using total body 3D imaging [59], especially using state-of-the-art deep learning-based approaches is an avenue worth further explorations [390].

Bibliography

[1] International Skin Imaging Collaboration: Melanoma Project. `https://www.isic-archive.com/`. [Online. Accessed December 11, 2020].

[2] Dermquest. `http://www.dermquest.com`, 2012. cited: 2020-04-28.

[3] Cancer facts and figures 2020. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2020/cancer-facts-and-figures-2020.pdf, January 8, 2020.

[4] Q. Abbas, M. E. Celebi, and I. F. Garcia. Hair Removal Methods: A Comparative Study for Dermoscopy Images. *Biomedical Signal Processing and Control*, 6(4):395–404, 2011.

[5] Naheed R Abbasi, Helen M Shaw, Darrell S Rigel, Robert J Friedman, William H McCarthy, Iman Osman, Alfred W Kopf, and David Polsky. Early diagnosis of cutaneous melanoma: revisiting the abcd criteria. *Jama*, 292(22):2771–2776, 2004.

[6] I. S. A. Abdelhalim, M. F. Mohamed, and Y. B. Mahdy. Data Augmentation For Skin Lesion Using Self-Attention Based Progressive Generative Adversarial Network. *Expert Systems with Applications*, 165:113922, 2021.

[7] Kumar Abhishek and Ghassan Hamarneh. Mask2Lesion: Mask-constrained adversarial skin lesion image synthesis. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 71–80. Springer, 2019.

[8] Kumar Abhishek and Ghassan Hamarneh. Matthews correlation coefficient loss for deep convolutional networks: Application to skin lesion segmentation. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 225–229. IEEE, 2021.

[9] Kumar Abhishek, Ghassan Hamarneh, and Mark S Drew. Illumination-based transformations improve skin lesion segmentation in dermoscopic images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 728–729, 2020.

[10] Kumar Abhishek, Jeremy Kawahara, and Ghassan Hamarneh. Predicting the clinical management of skin lesions using deep learning. *Scientific reports*, 11(1):1–14, 2021.

[11] Nabila Abraham and Naimul Mefraz Khan. A novel focal tversky loss function with improved attention U-Net for lesion segmentation. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 683–687. IEEE, 2019.

[12] Adekanmi Adegun and Serestina Viriri. An enhanced deep learning framework for skin lesions segmentation. In *International conference on computational collective intelligence*, pages 414–425. Springer, 2019.

[13] Adekanmi Adegun and Serestina Viriri. Deep learning techniques for skin lesion analysis and melanoma cancer detection: a survey of state-of-the-art. *Artificial Intelligence Review*, pages 1–31, June 2020.

[14] Adekanmi A Adegun and Serestina Viriri. Fcn-based densenet framework for automated detection and classification of skin lesions in dermoscopy images. *IEEE Access*, 8:150377–150396, 2020.

[15] Adegun Adekanmi Adeyinka and Serestina Viriri. Skin lesion images segmentation: a survey of the state-of-the-art. In *International Conference on Mining Intelligence and Knowledge Exploration*, pages 321–330. Springer, 2018.

[16] Mohammed A Al-Masni, Mugahed A Al-antari, Mun-Taek Choi, Seung-Moo Han, and Tae-Seong Kim. Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks. *Computer methods and programs in biomedicine*, 162:221–231, 2018.

[17] Mohammed A Al-masni, Mugahed A Al-antari, Hye Min Park, Na Hyeon Park, and Tae-Seong Kim. A deep learning model integrating FrCN and residual convolutional networks for skin lesion segmentation and classification. In *2019 IEEE Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS)*, pages 95–98. IEEE, 2019.

[18] Mohammed A Al-Masni, Dong-Hyun Kim, and Tae-Seong Kim. Multiple skin lesions diagnostics via integrated deep convolutional networks for segmentation and classification. *Computer methods and programs in biomedicine*, 190:105351, 2020.

[19] Zabir Al Nazi and Tasnim Azad Abir. Automatic skin lesion segmentation and melanoma detection: Transfer learning approach with U-Net and DCNN-SVM. In *Proceedings of International Joint Conference on Computational Intelligence*, pages 371–381. Springer, 2020.

[20] Md Zahangir Alom, Theus Aspiras, Tarek M Taha, and Vijayan K Asari. Skin cancer segmentation and classification with NABLA-N and inception recurrent residual convolutional networks. *arXiv preprint arXiv:1904.11126*, 2019.

[21] Md Zahangir Alom, Chris Yakopcic, Mahmudul Hasan, Tarek M Taha, and Vijayan K Asari. Recurrent residual u-net for medical image segmentation. *Journal of Medical Imaging*, 6(1):014006, 2019.

[22] Giuseppe Argenziano, Gabriella Fabbrocini, Paolo Carli, Vincenzo De Giorgi, Elena Sammarco, and Mario Delfino. Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: comparison of the abcd rule of dermatoscopy and a new 7-point checklist based on pattern analysis. *Archives of dermatology*, 134(12):1563–1570, 1998.

[23] Giuseppe Argenziano, H Peter Soyer, Vincenzo De Giorgio, Domenico Piccolo, Paolo Carli, Mario Delfino, Angela Ferrari, Rainer Hofmann-Wellenhof, Daniela Massi, Giampero Mazzocchetti, et al. *Interactive Atlas of Dermoscopy*. Edra Medical Publishing and New Media, 2000.

[24] Ridhi Arora, Balasubramanian Raman, Kritagya Nayyar, and Ruchi Awasthi. Automated skin lesion segmentation using attention-based deep convolutional neural network. *Biomedical Signal Processing and Control*, 65:102358, 2021.

[25] Mohamed Attia, Mohamed Hossny, Saeid Nahavandi, and Anousha Yazdabadi. Skin melanoma segmentation using recurrent and convolutional neural networks. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 292–296. IEEE, 2017.

[26] Reza Azad, Maryam Asadi-Aghbolaghi, Mahmood Fathy, and Sergio Escalera. Bidirectional ConvLSTM U-Net with densley connected convolutions. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[27] Reza Azad, Maryam Asadi-Aghbolaghi, Mahmood Fathy, and Sergio Escalera. Attention deeplabv3+: Multi-level context attention mechanism for skin lesion segmentation. In *European Conference on Computer Vision Workshops*, pages 251–266. Springer, 2020.

[28] Fatemeh Bagheri, Mohammad Jafar Tarokh, and Majid Ziaratban. Skin lesion segmentation based on mask rcnn, multi atrous full-cnn, and a geodesic method. *International Journal of Imaging Systems and Technology*, 2021.

[29] Fatemeh Bagheri, Mohammad Jafar Tarokh, and Majid Ziaratban. Skin lesion segmentation from dermoscopic images by using mask r-cnn, retina-deeplab, and graph-based methods. *Biomedical Signal Processing and Control*, 67:102533, 2021.

[30] Saleh Baghersalimi, Behzad Bozorgtabar, Philippe Schmid-Saugeon, Hazım Kemal Ekenel, and Jean-Philippe Thiran. DermoNet: densely linked convolutional neural network for efficient skin lesion segmentation. *EURASIP Journal on Image and Video Processing*, 2019(1):71, 2019.

[31] Ramsha Baig, Maryam Bibi, Anmol Hamid, Sumaira Kausar, and Shahzad Khalid. Deep learning approaches towards skin lesion segmentation and classification from dermoscopic images - a review. *Current Medical Imaging Reviews*, 15:1–20, 2019.

[32] Pierre Baldi, Søren Brunak, Yves Chauvin, Claus AF Andersen, and Henrik Nielsen. Assessing the Accuracy of Prediction Algorithms for Classification: An Overview. *Bioinformatics*, 16(5):412–424, 2000.

[33] Lucia Ballerini, Robert B Fisher, Ben Aldridge, and Jonathan Rees. A color and texture based hierarchical k-nn approach to the classification of non-melanoma skin lesions. In *Color Medical Image Analysis*, pages 63–86. Springer, 2013.

[34] C. Barata, M. E. Celebi, and J. S. Marques. Improving Dermoscopy Image Classification Using Color Constancy. *IEEE Journal of Biomedical and Health Informatics*, 19(3):1146–1152, 2015.

[35] C. Barata, M. E. Celebi, and J. S. Marques. Toward a Robust Analysis of Dermoscopy Images Acquired Under Different Conditions. In M. E. Celebi, T. Mendonca, and J. S. Marques, editors, *Dermoscopy Image Analysis*, pages 1–22. CRC Press, 2015.

[36] C. Barata, M. E. Celebi, and J. S. Marques. A Survey of Feature Extraction in Dermoscopy Image Analysis of Skin Cancer. *IEEE Journal of Biomedical and Health Informatics*, 23(3):1096–1109, 2019.

[37] C. Barata, M. Ruela, M. Francisco, T. Mendonca, and J. S. Marques. Two Systems for the Detection of Melanomas In Dermoscopy Images Using Texture and Color Features. *IEEE Systems Journal*, 8(3):965–979, 2014.

[38] C. Baur, S. Albarqouni, and N. Navab. Generating Highly Realistic Images of Skin Lesions with GANs. In *Proceedings of the Third ISIC Workshop on Skin Image Analysis*, pages 260–267, 2018.

[39] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *European conference on computer vision*, pages 549–565. Springer, 2016.

[40] Y. Bengio. Practical Recommendations for Gradient-Based Training of Deep Architectures. In G. Montavon, G. Orr, and K. R. Muller, editors, *Neural networks: Tricks of the Trade*, pages 437–478. Springer, Second edition, 2012.

[41] Y. Bengio, A. Courville, and P. Vincent. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[42] Aïcha BenTaieb and Ghassan Hamarneh. Topology aware fully convolutional networks for histology gland segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 460–468, 2016.

[43] Matt Berseth. ISIC 2017-skin lesion analysis towards melanoma detection. *arXiv:1703.00523*, 2017.

[44] Lei Bi, Dagan Feng, Michael Fulham, and Jinman Kim. Improving skin lesion segmentation via stacked adversarial learning. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1100–1103. IEEE, 2019.

[45] Lei Bi, Dagan Feng, and Jinman Kim. Improving automatic skin lesion segmentation using adversarial learning based data augmentation. *arXiv preprint arXiv:1807.08392*, 2018.

[46] Lei Bi, Michael Fulham, and Jinman Kim. Hyper-fusion network for semi-automatic segmentation of skin lesions. *Medical Image Analysis*, 76:102334, 2022.

[47] Lei Bi, Jinman Kim, Euijoon Ahn, and Dagan Feng. Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks. *arXiv:1703.04197*, 2017.

[48] Lei Bi, Jinman Kim, Euijoon Ahn, Dagan Feng, and Michael Fulham. Semi-automatic skin lesion segmentation via fully convolutional networks. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 561–564. IEEE, 2017.

[49] Lei Bi, Jinman Kim, Euijoon Ahn, Ashnil Kumar, Dagan Feng, and Michael Fulham. Step-wise integration of deep class-specific learning for dermoscopic image segmentation. *Pattern recognition*, 85:78–89, 2019.

[50] Lei Bi, Jinman Kim, Euijoon Ahn, Ashnil Kumar, Michael Fulham, and Dagan Feng. Dermoscopic image segmentation via multistage fully convolutional networks. *IEEE Transactions on Biomedical Engineering*, 64(9):2065–2074, 2017.

[51] Alberto M Biancardi, Artit C Jirapatnakul, and Anthony P Reeves. A Comparison of Ground Truth Estimation Methods. *International Journal of Computer Assisted Radiology and Surgery*, 5(3):295–305, 2010.

[52] Alberto M Biancardi and Anthony P Reeves. TESD: a novel ground truth estimation method. In *Medical Imaging 2009: Computer-Aided Diagnosis*, volume 7260, page 72603V. International Society for Optics and Photonics, February 2009.

[53] M. Binder, A. Steiner, M. Schwarz, S. Knollmayer, K. Wolff, and H. Pehamberger. Application of an Artificial Neural Network in Epiluminescence Microscopy Pattern Analysis of Pigmented Skin Lesions: A Pilot Study. *British Journal of Dermatology*, 130(4):460–465, 1994.

[54] Michael Binder, Margot Schwarz, Alexander Winkler, Andreas Steiner, Alexandra Kaider, Klaus Wolff, and Hubert Pehamberger. Epiluminescence microscopy. a useful tool for the diagnosis of pigmented skin lesions for formally trained dermatologists. *Archives of Dermatology*, 131(3):286–291, 1995.

[55] Alceu Bissoto, Michel Fornaciali, Eduardo Valle, and Sandra Avila. (de)constructing bias on skin lesion datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[56] Alceu Bissoto, Fábio Perez, Eduardo Valle, and Sandra Avila. Skin lesion synthesis with generative adversarial networks. In *OR 2.0 context-aware operating theaters, computer assisted robotic endoscopy, clinical image-based procedures, and skin image analysis*, pages 294–302. 2018.

[57] Alceu Bissoto, Eduardo Valle, and Sandra Avila. Gan-based data augmentation and anonymization for skin-lesion analysis: A critical review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1847–1856, June 2021.

[58] F Bogo, F Peruch, A B Fortina, and E Peserico. Where's the Lesion? Variability in Human and Automated Segmentation of Dermoscopy Images of Melanocytic Skin Lesions. In M. E. Celebi, T. Mendonca, and J. S. Marques, editors, *Dermoscopy Image Analysis*, pages 67–95. CRC Press, 2015.

[59] Federica Bogo, Javier Romero, Enoch Peserico, and Michael J Black. Automated detection of new or evolving melanocytic lesions using a 3d body model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 593–600. Springer, 2014.

[60] Sabri Boughorbel, Fethi Jarray, and Mohammed El-Anbari. Optimal Classifier for Imbalanced Data Using Matthews Correlation Coefficient Metric. *PLOS One*, 12(6):e0177678, 2017.

[61] Behzad Bozorgtabar, Zongyuan Ge, Rajib Chakravorty, Mani Abedini, Sergey Demyanov, and Rahil Garnavi. Investigating deep side layers for skin lesion segmentation. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 256–260. IEEE, 2017.

[62] Behzad Bozorgtabar, Suman Sedai, Pallab Kanti Roy, and Rahil Garnavi. Skin lesion segmentation using deep convolution networks guided by local unsupervised learning. *IBM Journal of Research and Development*, 61(4/5):6–1, 2017.

[63] L. Busin, N. Vandenbroucke, and L. Macaire. Color Spaces and Image Segmentation. In P. W. Hawkes, editor, *Advances in Imaging and Electron Physics*, volume 151, pages 65–168. Academic Press, 2008.

[64] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin. Albumentations: Fast and Flexible Image Augmentations. *Information*, 11(2):125, 2020.

[65] L. J. Caffery, D. Clunie, C. Curiel-Lewandrowski, J. Malvehy, H. P. Soyer, and A. C. Halpern. Transforming Dermatologic Imaging for the Digital Era: Metadata and Standards. *Journal of Digital Imaging*, 31:pages568–577, 2018.

[66] Laura Canalini, Federico Pollastri, Federico Bolelli, Michele Cancilla, Stefano Allegretti, and Costantino Grana. Skin lesion segmentation ensemble with diverse training strategies. In *International Conference on Computer Analysis of Images and Patterns*, pages 89–101. Springer, 2019.

[67] M. E. Celebi, A. Aslandogan, and W. V. Stoecker. Unsupervised Border Detection in Dermoscopy Images. *Skin Research and Technology*, 13(4):454–462, 2007.

[68] M. E. Celebi, T. Mendonca, and J. S. Marques, editors. *Dermoscopy Image Analysis*. CRC Press, 2015.

[69] M Emre Celebi, Noel Codella, and Allan Halpern.

[70] M Emre Celebi, Hitoshi Iyatomi, Gerald Schaefer, and William V Stoecker. Lesion border detection in dermoscopy images. *Computerized medical imaging and graphics*, 33(2):148–153, 2009.

[71] M Emre Celebi, Hitoshi Iyatomi, William V Stoecker, Randy H Moss, Harold S Rabinovitz, Giuseppe Argenziano, and H Peter Soyer. Automatic Detection of Blue-White Veil and Related Structures in Dermoscopy Images. *Computerized Medical Imaging and Graphics*, 32(8):670–677, 2008.

[72] M Emre Celebi, Hassan A Kingravi, Bakhtiyar Uddin, Hitoshi Iyatomi, Y Alp Aslandogan, William V Stoecker, and Randy H Moss. A Methodological Approach to the Classification of Dermoscopy Images. 31(6):362–373, 2007.

[73] M Emre Celebi, Gerald Schaefer, Hitoshi Iyatomi, William V Stoecker, Joseph M Malters, and James M Grichnik. An Improved Objective Evaluation Measure for Border Detection in Dermoscopy Images. *Skin Research and Technology*, 15(4):444–450, 2009.

[74] M Emre Celebi, QUAN Wen, HITOSHI Iyatomi, KOUHEI Shimizu, Huiyu Zhou, and Gerald Schaefer. A State-of-the-Art Survey on Lesion Border Detection in Dermoscopy Images. In M. E. Celebi, T. Mendonca, and J. S. Marques, editors, *Dermoscopy Image Analysis*, pages 97–129. CRC Press, 2015.

[75] Sebastien Chabrier, Bruno Emile, Christophe Rosenberger, and Helene Laurent. Unsupervised Performance Evaluation of Image Segmentation. *EURASIP Journal on Advances in Signal Processing*, 2006:1–12, 2006.

[76] Vikram Chalana and Yongmin Kim. A Methodology for Evaluation of Boundary Detection Algorithms on Medical Images. *IEEE Transactions on Medical Imaging*, 16(5):642–652, 1997.

[77] Abhishek Chaurasia and Eugenio Culurciello. LinkNet: Exploiting encoder representations for efficient semantic segmentation. In *2017 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE, 2017.

[78] Fei Chen, Huimin Yu, Roland Hu, and Xunxun Zeng. Deep learning shape priors for object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1870–1877, 2013.

[79] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017.

[80] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[81] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[82] Shenchang Eric Chen and Richard E Parent. Shape Averaging and its Applications to Industrial Design. *IEEE Computer Graphics and Applications*, 9(1):47–54, 1989.

[83] Sheng Chen, Zhe Wang, Jianping Shi, Bin Liu, and Nenghai Yu. A multi-task framework with feature passing module for skin lesion classification and segmentation. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 1126–1129. IEEE, 2018.

[84] Davide Chicco and Giuseppe Jurman. The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation. *BMC Genomics*, 21(1), 2020.

[85] Deepak Roy Chittajallu, Shishir K Shah, and Ioannis A Kakadiaris. A shape-driven MRF model for the segmentation of organs in medical images. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3233–3240, 2010.

[86] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[87] P. Y Chou and G D Fasman. Prediction of the Secondary Structure of Proteins from Their Amino Acid Sequence. In A. Meister, editor, *Advances in Enzymology and Related Areas of Molecular Biology*, volume 47, pages 45–148. John Wiley & Sons, 1978.

[88] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). *arXiv preprint arXiv:1902.03368*, 2019.

[89] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald 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). pages 168–172, 2018.

[90] Noel CF Codella, Q-B Nguyen, Sharath Pankanti, David A Gutman, Brian Helba, Allan C Halpern, and John R Smith. Deep Learning Ensembles for Melanoma Recognition in Dermoscopy Images. *IBM Journal of Research and Development*, 61(4/5):5:1–5:15, 2017.

[91] Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational andPsychological Measurement*, 20(1):37–46, 1960.

[92] Marc Combalia, Noel CF Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C Halpern, Susana Puig, et al. BCN20000: Dermoscopic Lesions in the Wild. https://arxiv.org/abs/1908.02288.

[93] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685, 2001.

[94] Daniel Cremers, Mikael Rousson, and Rachid Deriche. A review of statistical approaches to level set segmentation: integrating color, texture, motion and shape. *International journal of computer vision*, 72(2):195–215, 2007.

[95] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath. Generative Adversarial Networks: An Overview. *IEEE Signal Processing Magazine*, 35(1):53–65, 2018.

[96] William R Crum, Oscar Camara, and Derek LG Hill. Generalized Overlap Measures for Evaluation and Validation in Medical Image Analysis. *IEEE Transactions on Medical Imaging*, 25(11):1451–1461, 2006.

[97] Zhiying Cui, Longshi Wu, Ruixuan Wang, and Wei-Shi Zheng. Ensemble transductive learning for skin lesion segmentation. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 572–581. Springer, 2019.

[98] C. Curiel-Lewandrowski, R. A. Novoa, E. Berry, M. E. Celebi, N. Codella, F. Giuste, D. Gutman, A. Halpern, S. Leachman, Y. Liu, Y. Liu, O. Reiter, and P. Tschandl. Artificial Intelligence Approach in Melanoma. In D. E. Fisher and B. C. Bastian, editors, *Melanoma*, pages 599–628. Spriner, 2019.

[99] Duwei Dai, Caixia Dong, Songhua Xu, Qingsen Yan, Zongfang Li, Chunyan Zhang, and Nana Luo. Ms RED: A novel multi-scale residual encoding and decoding network for skin lesion segmentation. *Medical Image Analysis*, 75:102293, 2022.

[100] Jifeng Dai, Kaiming He, and Jian Sun. BoxSup: Exploiting Bounding Boxes to Supervise Convolutional Networks for Semantic Segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1635–1643, 2015.

[101] Gabriel G De Angelo, Andre GC Pacheco, and Renato A Krohling. Skin lesion segmentation using deep learning for images acquired from smartphones. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

[102] Zhuofu Deng, Yi Xin, Xiaolin Qiu, and Yeda Chen. Weakly and semi-supervised deep level set network for automated skin lesion segmentation. In *Innovation in Medicine and Healthcare*, pages 145–155. Springer, 2020.

[103] Zilin Deng, Haidi Fan, Fengying Xie, Yong Cui, and Jie Liu. Segmentation of dermoscopy images based on fully convolutional neural network. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1732–1736. IEEE, 2017.

[104] Emily Denton, Soumith Chintala, Arthur Szlam, and Rob Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, pages 1486–1494, 2015.

[105] Terrance DeVries and Graham W Taylor. Leveraging uncertainty estimates for predicting segmentation quality. *arXiv preprint arXiv:1807.00502*, 2018.

[106] Lee R Dice. Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26(3):297–302, 1945.

[107] Sander Dieleman et al. Lasagne: First release., August 2015.

[108] S. Ding, J. Zheng, Z. Liu, Y. Zheng, Y. Chen, X. Xu, J. Lu, and J. Xie. High-Resolution Dermoscopy Image Synthesis with Conditional Generative Adversarial Networks. *Biomedical Signal Processing and Control*, 64:102224, 2021.

[109] S. Dodge and L. Karam. Understanding How Image Quality Affects Deep Neural Networks. In *Proceedings of the 2016 International Conference on Quality of Multimedia Experience*, 2016.

[110] Joshua Peter Ebenezer and Jagath C Rajapakse. Automatic segmentation of skin lesions using deep learning. *arXiv preprint arXiv:1807.04893*, 2018.

[111] T. Elsken, J. H. Metzen, and F. Hutter. Neural Architecture Search: A Survey. *Journal of Machine Learning Research*, 20:1–21, 2019.

[112] Holly C Engasser and Erin M Warshaw. Dermatoscopy use by us dermatologists: a cross-sectional survey. *Journal of the American Academy of Dermatology*, 63(3):412–419, 2010.

[113] Bulent Erkol, Randy H Moss, R Joe Stanley, William V Stoecker, and Erik Hvatum. Automatic Lesion Boundary Detection in Dermoscopy Images Using Gradient Vector Flow Snakes. *Skin Research and Technology*, 11(1):17–26, 2005.

[114] Pedro M Ferreira, Teresa Mendonça, Jorge Rozeira, and Paula Rocha. An Annotation Tool for Dermoscopic Image Segmentation. In *Proceedings of the 1st International Workshop on Visual Interfaces for Ground Truth Collection in Computer Vision Applications*, 2012.

[115] A. Foncubierta-Rodriguez and H. Muller. Ground Truth Generation in Medical Imaging: A Crowdsourcing-Based Iterative Approach. In *Proceedings of the ACM Multimedia 2012 Workshop on Crowdsourcing for Multimedia*, pages 9–14, 2012.

[116] Anna Belloni Fortina, Enoch Peserico, Alberto Silletti, and Edoardo Zattra. Where's the Naevus? Inter-Operator Variability in the Localization of Melanocytic Lesion Border. *Skin Research and Technology*, 18(3):311–315, 2012.

[117] Daniel Freedman and Tao Zhang. Interactive graph cut based segmentation with shape priors. In *Proceedings of the 2005 IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 755–762, 2005.

[118] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019.

[119] Michael C. Fu, Rafael A. Buerba, William D. Long, Daniel J. Blizzard, Andrew W. Lischuk, Andrew H. Haims, and Jonathan N. Grauer. Interrater and intrarater agreements of magnetic resonance imaging findings in the lumbar spine: significant variability across degenerative conditions. *The Spine Journal*, 14(10):2442–2448, October 2014.

[120] Julie Gachon, Philippe Beaulieu, Jean Francois Sei, Johanny Gouvernet, Jean Paul Claudel, Michel Lemaitre, Marie Aleth Richard, and Jean Jacques Grob. First prospective study of the recognition process of melanoma in dermatological practice. *Archives of dermatology*, 141(4):434–438, 2005.

[121] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059, 2016.

[122] Harald Ganster, Margit Gelautz, Axel Pinz, Michael Binder, Hubert Pehamberger, Manfred Bammer, and Johann Krocza. Initial Results of Automated Melanoma Recognition. In G. Borgefors, editor, *Theory and Applications of Image Analysis II: Selected Papers from the 9th Scandinavian Conference on Image Analysis*, pages 343–354. World Scientific Publishing Co. Pte. Ltd., 1995.

[123] Rahil Garnavi and Mohammad Aldeen. Optimized Weighted Performance Index for Objective Evaluation of Border-Detection Methods in Dermoscopy Images. *IEEE Transactions on Information Technology in Biomedicine*, 15(6):908–917, 2011.

[124] Rahil Garnavi, Mohammad Aldeen, M Emre Celebi, George Varigos, and Sue Finch. Border Detection in Dermoscopy Images Using Hybrid Thresholding on Optimized Color Channels. *Computerized Medical Imaging and Graphics*, 35(2):105–115, 2011.

[125] Rahil Garnavi, Mohammad Aldeen, and ME Celebi. Weighted Performance Index for Objective Evaluation of BorderDetection Methods in Dermoscopy Images. *Skin Research and Technology*, 17(1):35–44, 2011.

[126] Caroline Gaudy-Marqueste, Yanal Wazaefi, Yvane Bruneu, Raoul Triller, Luc Thomas, Giovanni Pellacani, Josep Malvehy, Marie-Françoise Avril, Sandrine Monestier, Marie-Aleth Richard, et al. Ugly duckling sign as a major factor of efficiency in melanoma detection. *JAMA dermatology*, 153(4):279–284, 2017.

[127] S. L. Gish and W. E. Blanz. Comparing the Performance of Connectionist and Statistical Classifiers on an Image Segmentation Problem. In *Proceedings of the Second International Conference on Neural Information Processing Systems*, pages 614–621, 1989.

[128] S. Goel, Y. Sharma, M. L. Jauer, and T. M. Deserno. WeLineation: Crowdsourcing Delineations for Reliable Ground Truth Estimation. In *Proceedings of the Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications*, pages 113180C–1–113180C–8, 2020.

[129] Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. In *International Conference on Learning Representations*, 2017.

[130] David Delgado Gómez, Constantine Butakoff, Bjarne Kjaer Ersboll, and William Stoecker. Independent histogram pursuit for segmentation of skin lesions. *IEEE transactions on biomedical engineering*, 55(1):157–161, 2007.

[131] Ivan Gonzalez-Diaz. Dermaknet: Incorporating the knowledge of dermatologists to convolutional neural networks for skin lesion diagnosis. *IEEE journal of biomedical and health informatics*, 23(2):547–559, 2018.

[132] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Networks. *Communications of the ACM*, 63(11):139–144, 2020.

[133] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[134] Manu Goyal, Jiahua Ng, Amanda Oakley, and Moi Hoon Yap. Skin lesion boundary segmentation with fully automated deep extreme cut methods. In *Medical Imaging 2019: Biomedical Applications in Molecular, Structural, and Functional Imaging*, volume 10953, page 109530Q. International Society for Optics and Photonics, 2019.

[135] Manu Goyal, Amanda Oakley, Priyanka Bansal, Darren Dancey, and Moi Hoon Yap. Skin lesion segmentation in dermoscopic images with ensemble deep learning methods. *IEEE Access*, 8:4171–4181, 2020.

[136] Manu Goyal, Moi Hoon Yap, and Saeed Hassanpour. Multi-class semantic segmentation of skin lesions via fully convolutional networks. *arXiv preprint arXiv:1711.10449*, 2017.

[137] Vicente Grau, AUJ Mewes, M Alcaniz, Ron Kikinis, and Simon K Warfield. Improved Watershed Transform for Medical Image Segmentation Using Prior Information. *IEEE Transactions on Medical Imaging*, 23(4):447–458, 2004.

[138] Adele Green, Nicholas Martin, John Pfitzner, Michael O'Rourke, and Ngaire Knight. Computer image analysis in the diagnosis of melanoma. *Journal of the American Academy of Dermatology*, 31(6):958–964, 1994.

[139] Pengfei Gu, Hao Zheng, Yizhe Zhang, Chaoli Wang, and Danny Z Chen. kCBAC-Net: Deeply supervised complete bipartite networks with asymmetric convolutions for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 337–347. Springer, 2021.

[140] Ran Gu, Guotai Wang, Tao Song, Rui Huang, Michael Aertsen, Jan Deprest, Sébastien Ourselin, Tom Vercauteren, and Shaoting Zhang. CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. *IEEE transactions on medical imaging*, 40(2):699–711, 2020.

[141] Naga Raju Gudhe, Hamid Behravan, Mazen Sudah, Hidemi Okuma, Ritva Vanninen, Veli-Matti Kosma, and Arto Mannermaa. Multi-level dilated residual network for biomedical image segmentation. *Scientific Reports*, 11(1):1–18, 2021.

[142] Xiaoqing Guo, Zhen Chen, and Yixuan Yuan. Complementary network with adaptive receptive fields for melanoma segmentation. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 2010–2013. IEEE, 2020.

[143] D. Gurari, D. Theriault, M. Sameki, B. Isenberg, T. A. Pham, A. Purwada, P. Solski, M. Walker, C. Zhang, J. Y. Wong, and M. Betke. How to Collect Segmentations for Biomedical Images? A Benchmark Evaluating the Performance of Experts, Crowdsourced Non-Experts, and Algorithms. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 1169–1176, 2015.

[144] David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan 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.

[145] Gery P Guy Jr, Steven R Machlin, Donatus U Ekwueme, and K Robin Yabroff. Prevalence and costs of skin cancer treatment in the us, 2002- 2006 and 2007- 2011. *American journal of preventive medicine*, 48(2):183–187, 2015.

[146] Ghassan Hamarneh et al. Simulation of ground-truth validation data via physically- and statistically-based warps. In *International Conference on Medical image computing and computer-assisted intervention*, pages 459–467, 2008.

[147] Gregory A Hance, Scott E Umbaugh, Randy H Moss, and William V Stoecker. Unsupervised Color Image Segmentation with Application to Skin Tumor Borders. *IEEE Engineering in Medicine and Biology Magazine*, 15(1):104–111, 1996.

[148] Md Hasan, Shidhartho Roy, Chayan Mondal, Md Alam, Md Elahi, E Toufick, Aishwariya Dutta, SM Raju, Mohiuddin Ahmad, et al. Dermo-doctor: A framework for concurrent skin lesion detection and recognition using a deep convolutional neural network with end-to-end dual encoders, 2021.

[149] Md Kamrul Hasan, Lavsen Dahal, Prasad N Samarakoon, Fakrul Islam Tushar, and Robert Martí. DSNet: Automatic dermoscopic skin lesion segmentation. *Computers in Biology and Medicine*, page 103738, 2020.

[150] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[151] Xinzi He, Zhen Yu, Tianfu Wang, and Baiying Lei. Skin lesion segmentation via deep RefineNet. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 303–311. Springer, 2017.

[152] Xinzi He, Zhen Yu, Tianfu Wang, Baiying Lei, and Yiyan Shi. Dense deconvolution net: Multi path fusion and dense deconvolution for high resolution skin lesion segmentation. *Technology and Health Care*, 26(S1):307–316, 2018.

[153] H Yu Henry, Xue Feng, Ziwen Wang, and Hao Sun. Mixmodule: Mixed cnn kernel module for medical image segmentation. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1508–1512. IEEE, 2020.

[154] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for MobileNetV3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1314–1324, 2019.

[155] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[156] Shi Hu, Daniel Worrall, Stefan Knegt, Bas Veeling, Henkjan Huisman, and Max Welling. Supervised uncertainty quantification for segmentation with multiple annotations. In *International Conference on Medical image computing and computer-assisted intervention*, pages 137–145. Springer, 2019.

[157] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[158] Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. What makes ImageNet good for transfer learning? *arXiv:1608.08614*, 2016.

[159] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge. Comparing Images Using the Hausdorff Distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):850–863, 1993.

[160] Sergey Ioffe et al. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *International Conference on Machine Learning*, volume 37, pages 448–456, 2015.

[161] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[162] H. Iyatomi, H. Oka, M. E. Celebi, M. Hashimoto, M. Hagiwara, M. Tanaka, and K. Ogawa. An Improved Internet-Based Melanoma Screening System with Dermatologist-Like Tumor Area Extraction Algorithm. *Computerized Medical Imaging and Graphics*, 32(7):566–579, 2008.

[163] H. Iyatomi, H. Oka, M. Saito, A. Miyake, M. Kimoto, J. Yamagami, S. Kobayashi, A. Tanikawa, M. Hagiwara, K. Ogawa, G. Argenziano, H. P. Soyer, and M. Tanaka. Quantitative Assessment of Tumor Extraction from Dermoscopy Images and Evaluation of Computer-Based Extraction Methods for Automatic Melanoma Diagnostic System. *Melanoma Research*, 16(2):183–190, 2006.

[164] Saeed Izadi, Zahra Mirikharaji, Jeremy Kawahara, and Ghassan Hamarneh. Generative adversarial networks to segment skin lesions. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 881–884. IEEE, 2018.

[165] Paul Jaccard. Distribution de la Flore Alpine dans le Bassin des Dranses et dans Quelques Regions Voisines. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, 37(140):241–272, 1901.

[166] M Hossein Jafari, Ebrahim Nasr-Esfahani, Nader Karimi, SM Reza Soroushmehr, Shadrokh Samavi, and Kayvan Najarian. Extraction of skin lesions from non-dermoscopic images for surgical excision of melanoma. *International journal of computer assisted radiology and surgery*, 12(6):1021–1030, 2017.

[167] Mina Jafari, Dorothee Auer, Susan Francis, Jonathan Garibaldi, and Xin Chen. Drunet: An efficient deep convolutional neural network for medical image segmentation. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1144–1148. IEEE, 2020.

[168] Mohammad H Jafari, Nader Karimi, Ebrahim Nasr-Esfahani, Shadrokh Samavi, S Mohamad R Soroushmehr, K Ward, and Kayvan Najarian. Skin lesion segmentation in clinical images using deep learning. In *2016 23rd International conference on pattern recognition (ICPR)*, pages 337–342. IEEE, 2016.

[169] Mostafa Jahanifar, Neda Zamani Tajeddin, Navid Alemi Koohbanani, Ali Gooya, and Nasir Rajpoot. Segmentation of skin lesions and their attributes using multi-scale convolutional neural networks and domain specific augmentations. *arXiv preprint arXiv:1809.10243*, 2018.

[170] Nathalie Japkowicz and Mohak Shah. *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press, 2011.

[171] Kalyanakumar Jayapriya and Israel Jeena Jacob. Hybrid fully convolutional networks-based skin lesion segmentation and melanoma detection using deep feature. *International Journal of Imaging Systems and Technology*, 30(2):348–357, 2020.

[172] J Daniel Jensen and Boni E Elewski. The abcdef rule: combining the "abcde rule" and the "ugly duckling sign" in an effort to improve patient self-screening examinations. *The Journal of clinical and aesthetic dermatology*, 8(2):15, 2015.

[173] Xu Ji, Joao F Henriques, and Andrea Vedaldi. Invariant Information Clustering for Unsupervised Image Classification and Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9865–9874, 2019.

[174] Feng Jiang, Feng Zhou, Jing Qin, Tianfu Wang, and Baiying Lei. Decision-augmented generative adversarial network for skin lesion segmentation. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 447–450. IEEE, 2019.

[175] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Regularizing very deep neural networks on corrupted labels. *arXiv preprint arXiv:1712.05055*, 4, 2017.

[176] Yun Jiang, Simin Cao, Shengxin Tao, and Hai Zhang. Skin lesion segmentation based on multi-scale attention convolutional neural network. *IEEE Access*, 8:122811–122825, 2020.

[177] Qiangguo Jin, Hui Cui, Changming Sun, Zhaopeng Meng, and Ran Su. Cascade knowledge diffusion network for skin lesion diagnosis and segmentation. *Applied Soft Computing*, 99:106881, 2021.

[178] Guillod Joel, Schmid-Saugeon Philippe, Guggisberg David, Cerottini Jean Philippe, Braun Ralph, Krischer Joakim, Saurat Jean-Hilaire, and Kunt Murat. Validation of Segmentation Techniques for Digital Dermoscopy. *Skin Research and Technology*, 8(4):240–249, 2002.

[179] Alain Jungo and Mauricio Reyes. Assessing reliability and challenges of uncertainty estimations for medical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 48–56. Springer, 2019.

[180] Anandhanarayanan Kamalakannan, Shiva Shankar Ganesan, and Govindaraj Rajamanickam. Self-learning ai framework for skin lesion image segmentation and classification. *International Journal of Computer Science and Information Technology*, 11(6):29–38, 2019.

[181] Joni-Kristian Kamarainen, Lasse Lensu, and Tomi Kauppi. Combining multiple image segmentations by maximizing expert agreement. In *International Workshop on Machine Learning in Medical Imaging*, pages 193–200. Springer, 2012.

[182] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

[183] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International journal of computer vision*, 1(4):321–331, 1988.

[184] W. T. Katz and M. B. Merickel. Translation-Invariant Aorta Segmentation from Magnetic Resonance Images. In *Proceedings of the 1989 International Joint Conference on Neural Networks*, pages 327–333, 1989.

[185] Chaitanya Kaul, Suresh Manandhar, and Nick Pears. FocusNet: an attention-based fully convolutional network for medical image segmentation. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 455–458. IEEE, 2019.

[186] Chaitanya Kaul, Nick Pears, Hang Dai, Roderick Murray-Smith, and Suresh Manandhar. Focusnet++: Attentive aggregated transformations for efficient and accurate medical image segmentation. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1042–1046. IEEE, 2021.

[187] Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. Seven-Point Checklist and Skin Lesion Classification Using Multitask Multimodal Neural Nets. *IEEE Journal of Biomedical and Health Informatics*, 23(2):538–546, 2019.

[188] Jeremy Kawahara and Ghassan Hamarneh. Fully convolutional neural networks to detect clinical dermoscopic features. *IEEE journal of biomedical and health informatics*, 23(2):578–585, 2018.

[189] Jeremy Kawahara, Chris McIntosh, Roger Tam, and Ghassan Hamarneh. Augmenting auto-context with global geometric features for spinal cord segmentation. In *International Workshop on Machine Learning in Medical Imaging*, pages 211–218. Springer, 2013.

[190] Ruya Kaymak, Cagri Kaymak, and Aysegul Ucar. Skin lesion segmentation using fully convolutional networks: A comparative experimental study. *Expert Systems with Applications*, 161:113742, 2020.

[191] S. Kazeminia, C. Baur, A. Kuijper, B. van Ginneken, N. Navab, S. Albarqouni, and A. Mukhopadhyay. GANs for Medical Image Analysis. *Artificial Intelligence in Medicine*, 109:101938, 2020.

[192] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv:1511.02680*, 2015.

[193] Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5574–5584, 2017.

[194] Abbas Khan, Hyongsuk Kim, and Leon Chua. Pmed-net: Pyramid based multi-scale encoder-decoder network for medical image segmentation. *IEEE Access*, 9:55988–55998, 2021.

[195] Sahib Khouloud, Melouah Ahlem, Touré Fadel, and Slim Amel. W-net and inception residual network for skin lesion segmentation and classification. *Applied Intelligence*, pages 1–19, 2021.

[196] Harold Kittler, H Pehamberger, K Wolff, and MJTIO Binder. Diagnostic accuracy of dermoscopy. *The lancet oncology*, 3(3):159–165, 2002.

[197] Simon Kohl et al. Adversarial networks for the detection of aggressive prostate cancer. *preprint arXiv:1702.08014*, 2017.

[198] Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R Ledsam, Klaus Maier-Hein, SM Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic U-Net for segmentation of ambiguous images. 31:6965–6975, 2018.

[199] Nikos Komodakis and Spyros Gidaris. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018.

[200] Gouse Mohiuddin Kosgiker, Anupama Deshpande, and Anjum Kauser. Segcaps: An efficient segcaps network-based skin lesion segmentation in dermoscopic images. *International Journal of Imaging Systems and Technology*, 2021.

[201] A. Kovashka, O. Russakovsky, L. Fei-Fei, and K. Grauman. Crowdsourcing in Computer Vision. *Foundations and Trends in Computer Graphics and Vision*, 10(3):177–243, 2016.

[202] Philipp Krähenbühl and Vladlen Koltun. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, pages 109–117, 2011.

[203] Miroslav Kubat, Robert C Holte, and Stan Matwin. Machine Learning for the Detection of Oil Spills in Satellite Radar Images. *Machine Learning*, 30(2–3):195–215, 1998.

[204] Sanjiv Kumar et al. Discriminative random fields: A discriminative framework for contextual interaction in classification. In *Proceedings ninth IEEE international conference on computer vision*, pages 1150–1157, 2003.

[205] Yongchan Kwon, Joong-Ho Won, Beom Joon Kim, and Myunghee Cho Paik. Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation. *Computational Statistics & Data Analysis*, 142:106816, 2020.

[206] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. 30:6402–6413, 2017.

[207] Thomas A Lampert, André Stumpf, and Pierre Gançarski. An Empirical Study into Annotator Agreement, Ground Truth Estimation, and Algorithm Evaluation. *IEEE Transactions on Image Processing*, 25(6):2557–2572, 2016.

[208] Thomas Robin Langerak, Uulke A van der Heide, Alexis NTJ Kotte, Max A Viergever, Marco Van Vulpen, and Josien PW Pluim. Label Fusion in Atlas-Based Segmentation Using a Selective and Iterative Method for Performance Level Estimation (SIMPLE). *IEEE Transactions on Medical Imaging*, 29(12):2000–2008, 2010.

[209] Max-Heinrich Laves, Sontje Ihler, Karl-Philipp Kortmann, and Tobias Ortmaier. Well-calibrated model uncertainty with temperature scaling for dropout variational inference. *arXiv preprint arXiv:1909.13550*, 2019.

[210] Y. LeCun, Y. Bengio, and G. Hinton. Deep Learning. *Nature*, 521(7553):436–444, 2015.

[211] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[212] T. K. Lee, D. I. McLean, and M. S. Atkins. Irregularity Index: A New Border Irregularity Measure for Cutaneous Melanocytic Lesions. *Medical Image Analysis*, 7(1):47–64, 2003.

[213] Baiying Lei, Zaimin Xia, Feng Jiang, Xudong Jiang, Zongyuan Ge, Yanwu Xu, Jing Qin, Siping Chen, Tianfu Wang, and Shuqiang Wang. Skin lesion segmentation via generative adversarial networks with dual discriminators. *Medical Image Analysis*, 64:101716, 2020.

[214] Hang Li, Xinzi He, Feng Zhou, Zhen Yu, Dong Ni, Siping Chen, Tianfu Wang, and Baiying Lei. Dense deconvolutional network for skin lesion segmentation. *IEEE journal of biomedical and health informatics*, 23(2):527–537, 2018.

[215] Ruizhe Li, Christian Wagner, Xin Chen, and Dorothee Auer. A generic ensemble based deep convolutional neural network for semi-supervised medical image segmentation. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1168–1172. IEEE, 2020.

[216] Wei Li, Alex Noel Joseph Raj, Tardi Tjahjadi, and Zhemin Zhuang. Digital hair removal by deep learning for skin lesion segmentation. *Pattern Recognition*, 117:107994, 2021.

[217] Xiang Li, Ben Aldridge, Robert Fisher, and Jonathan Rees. Estimating the ground truth from multiple individual segmentations incorporating prior pattern analysis with application to skin lesion segmentation. In *2011 IEEE ISBI: From Nano to Macro*, pages 1438–1441. IEEE, March 2011.

[218] Xiaomeng Li, Lequan Yu, Hao Chen, Chi-Wing Fu, Lei Xing, and Pheng-Ann Heng. Transformation consistent self-ensembling model for semi-supervised medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[219] Xiaomeng Li, Lequan Yu, Chi-Wing Fu, and Pheng-Ann Heng. Deeply supervised rotation equivariant network for lesion segmentation in dermoscopy images. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, pages 235–243. Springer, 2018.

[220] Yuexiang Li, Jiawei Chen, and Yefeng Zheng. A multi-task self-supervised learning framework for scopy images. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 2005–2009. IEEE, 2020.

[221] Yuexiang Li, Jiawei Chen, and Yefeng Zheng. A multi-task self-supervised learning framework for scopy images. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 2005–2009. IEEE, 2020.

[222] Yuexiang Li and Linlin Shen. Skin lesion analysis towards melanoma detection using deep learning network. *Sensors*, 18(2):556, 2018.

[223] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[224] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sanchez. A Survey on Deep Learning in Medical Image Analysis. *Medical Image Analysis*, 42:60–88, 2017.

[225] C. Liu, L. C. Chen, F. Schroff, H. Adam, W. Hua, A. L. Yuille, and L. Fei-Fei. Auto-DeepLab: Hierarchical Neural Architecture Search for Semantic Image Segmentation. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 82–92, 2019.

[226] Lina Liu, Lichao Mou, Xiao Xiang Zhu, and Mrinal Mandal. Skin lesion segmentation based on improved U-Net. In *2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE)*, pages 1–4. IEEE, 2019.

[227] Lina Liu, Ying Y Tsui, and Mrinal Mandal. Skin lesion segmentation using deep learning with auxiliary task. *Journal of Imaging*, 7(4):67, 2021.

[228] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[229] Wanxuan Lu, Dong Gong, Kun Fu, Xian Sun, Wenhui Diao, and Lingqiao Liu. Boundarymix: Generating pseudo-training images for improving segmentation with scribble annotations. *Pattern Recognition*, 117:107924, 2021.

[230] Pauline Luc et al. Semantic segmentation using adversarial networks. *preprint arXiv:1611.08408*, 2016.

[231] Amalia Luque, Alejandro Carrasco, Alejandro Martín, and Ana de las Heras. The Impact of Class Imbalance in Classification Performance Metrics Based on the Binary Confusion Matrix. *Pattern Recognition*, 91:216–231, 2020.

[232] Amirreza Mahbod, Philipp Tschandl, Georg Langs, Rupert Ecker, and Isabella Ellinger. The effects of skin lesion segmentation on the performance of dermatoscopic image classification. *Computer Methods and Programs in Biomedicine*, 197:105725, 2020.

[233] Roman C Maron, Achim Hekler, Eva Krieghoff-Henning, Max Schmitt, Justin G Schlager, Jochen S Utikal, and Titus J Brinker. Reducing the impact of confounding factors on skin cancer classification via image segmentation: Technical model study. *Journal of Medical Internet Research*, 23(3):e21695, 2021.

[234] Brian W Matthews. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochimica et Biophysica Acta*, 405(2):442–451, 1975.

[235] Tim McInerney and Demetri Terzopoulos. Deformable models in medical image analysis: a survey. *Medical image analysis*, 1(2):91–108, 1996.

[236] Alireza Mehrtash, William M Wells, Clare M Tempany, Purang Abolmaesumi, and Tina Kapur. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Transactions on Medical Imaging*, 39(12):3868–3878, 2020.

[237] T. Mendonca, P. M. Ferreira, J. S. Marques, A. R. S. Marcal, and J. Rozeira. PH^2—A Dermoscopic Image Database for Research and Benchmarking. In *Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5437–5440, 2013.

[238] Teresa Mendonça, Pedro M Ferreira, Jorge S Marques, André RS Marcal, and Jorge Rozeira. PH^2—A Dermoscopic Image Database for Research and Benchmarking. In M. E. Celebi, T. Mendonca, and J. S. Marques, editors, *Dermoscopy Image Analysis*, pages 419–439. CRC Press, 2015.

[239] Scott Menzies, Kerry Crotty, Christian Ingwar, and William H. McCarthy. *An Atlas of Surface Microscopy of Pigmented Skin Lesions: Dermoscopy.* McGraw-Hill, Second edition, 2003.

[240] Zahra Mirikharaji, Kumar Abhishek, Saeed Izadi, and Ghassan Hamarneh. D-LEMA: Deep learning ensembles from multiple annotations-application to skin lesion segmentation. pages 1837–1846, 2021.

[241] Zahra Mirikharaji and Ghassan Hamarneh. Star shape prior in fully convolutional networks for skin lesion segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 737–745. Springer, 2018.

[242] Zahra Mirikharaji, Saeed Izadi, Jeremy Kawahara, and Ghassan Hamarneh. Deep auto-context fully convolutional neural network for skin lesion segmentation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 877–880. IEEE, 2018.

[243] Zahra Mirikharaji, Yiqi Yan, and Ghassan 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*, pages 207–215. Springer, 2019.

[244] Rashika Mishra and Ovidiu Daescu. Deep learning for skin lesion segmentation. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1189–1194. IEEE, 2017.

[245] Pim Moeskops et al. Adversarial training and dilated convolutions for brain mri segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 56–64. 2017.

[246] Franz Nachbar, Wilhelm Stolz, Tanja Merkle, Armand B Cognetta, Thomas Vogt, Michael Landthaler, Peter Bilek, Otto Braun-Falco, and Gerd Plewig. The abcd rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions. *Journal of the American Academy of Dermatology*, 30(4):551–559, 1994.

[247] Ebrahim Nasr-Esfahani, Shima Rafiei, Mohammad H Jafari, Nader Karimi, James S Wrobel, Shadrokh Samavi, and SM Reza Soroushmehr. Dense pooling layers in fully convolutional network for skin lesion segmentation. *Computerized Medical Imaging and Graphics*, 78:101658, 2019.

[248] Sabari Nathan and Priya Kansal. Lesion net–skin lesion segmentation using coordinate convolution and deep residual units. *arXiv preprint arXiv:2012.14249*, 2020.

[249] Andrew Ng. Machine learning yearning: Technical strategy for ai engineers in the era of deep learning. *Retrieved online at https://www. mlyearning. org*, 2019.

[250] F. Ning, D. Delhomme, Y. LeCun, F. Piano, L. Bottou, and P. E. Barbano. Toward Automatic Phenotyping of Developing Embryos from Videos. *IEEE Transactions on Image Processing*, 14(9):1360–1371, 2005.

[251] K. A. Norton, H. Iyatomi, M. E. Celebi, S. Ishizaki, M. Sawada, R. Suzaki, K. Kobayashi, M. Tanaka, and K. Ogawa. Three-Phase General Border Detection Method for Dermoscopy Images Using Non-Uniform Illumination Correction. *Skin Research and Technology*, 18(3):290–300, 2012.

[252] Masoud S Nosrati and Ghassan Hamarneh. Segmentation of overlapping cervical cells: a variational method with star-shape prior. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pages 186–189, 2015.

[253] Masoud S Nosrati and Ghassan Hamarneh. Incorporating prior knowledge in medical image segmentation: a survey. *arXiv:1607.01092*, 2016.

[254] Ozan Oktay, Enzo Ferrante, Konstantinos Kamnitsas, Mattias Heinrich, Wenjia Bai, Jose Caballero, Stuart A Cook, Antonio De Marvao, Timothy Dawes, Declan P O'Regan, et al. Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation. *IEEE transactions on medical imaging*, 37(2):384–395, 2017.

[255] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention U-Net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[256] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1717–1724, 2014.

[257] Şaban Öztürk and Umut Özkaya. Skin lesion segmentation with improved convolutional neural network. *Journal of digital imaging*, 33:958–970, 2020.

[258] Junting Pan et al. Salgan: Visual saliency prediction with adversarial networks. In *CVPR Scene Understanding Workshop (SUNw)*, 2017.

[259] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly- and Semi-Supervised Learning of a Deep Convolutional Network for Semantic Image Segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1742–1750, 2015.

[260] Bo Peng and Tianrui Li. A Probabilistic Measure for Quantitative Evaluation of Image Segmentation. *IEEE Signal Processing Letters*, 20(7):689–692, 2013.

[261] Bo Peng, Xingzheng Wang, and Yan Yang. Region Based Exemplar References for Image Segmentation Evaluation. *IEEE Signal Processing Letters*, 23(4):459–462, 2016.

[262] Bo Peng, Lei Zhang, Xuanqin Mou, and Ming-Hsuan Yang. Evaluation of Segmentation Quality via Adaptive Composition of Reference Segmentations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(10):1929–1941, 2017.

[263] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters–improve semantic segmentation by global convolutional network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4353–4361, 2017.

[264] Yanjun Peng, Ning Wang, Yuanhong Wang, and Meiling Wang. Segmentation of dermoscopy image using adversarial networks. *Multimedia Tools and Applications*, 78(8):10965–10981, 2019.

[265] F. Perez, C. Vasconcelos, S. Avila, and E. Valle. Data Augmentation for Skin Lesion Analysis. In *Proceedings of the Third ISIC Workshop on Skin Image Analysis*, pages 303–311, 2018.

[266] E. Peserico and A. Silletti. Is (N)PRI Suitable for Evaluating Automated Segmentation of Cutaneous Lesions? *Pattern Recognition Letters*, 31(16):2464–2467, 2010.

[267] Pedro HO Pinheiro and Ronan Collobert. Recurrent convolutional neural networks for scene labeling. In *31st International Conference on Machine Learning (International Conference on Machine Learning)*, number CONF, 2014.

[268] Federico Pollastri, Federico Bolelli, Roberto Paredes Palacios, and Costantino Grana. Improving skin lesion segmentation with generative adversarial networks. In *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*, pages 442–443. IEEE, 2018.

[269] Federico Pollastri, Federico Bolelli, Roberto Paredes, and Costantino Grana. Augmenting Data with GANs to Segment Melanoma Skin Lesions. *Multimedia Tools and Applications*, 79(21):15575–15592, 2020.

[270] Sahadev Poudel and Sang-Woong Lee. Deep multi-scale attentional features for medical image segmentation. *Applied Soft Computing*, 109:107445, 2021.

[271] M. P. Pour and H. Seker. Transform Domain Representation-Driven Convolutional Neural Networks for Skin Lesion Segmentation. *Expert Systems with Applications*, 144:113129, 2020.

[272] Rui Qian, Yunchao Wei, Honghui Shi, Jiachen Li, Jiaying Liu, and Thomas Huang. Weakly supervised scene parsing with point-based distance metric learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8843–8850, 2019.

[273] Yuming Qiu, Jingyong Cai, Xiaolin Qin, and Ju Zhang. Inferring skin lesion deep convolutional neural networks. *IEEE Access*, 8:144246–144258, 2020.

[274] Dhanesh Ramachandram and Terrance DeVries. Lesionseg: semantic segmentation of skin lesions using deep convolutional neural network. *arXiv preprint arXiv:1703.03372*, 2017.

[275] Dhanesh Ramachandram and Graham W Taylor. Skin lesion segmentation using deep hypercolumn descriptors. *Journal of Computational Vision and Imaging Systems*, 3(1), 2017.

[276] D Roja Ramani and S. Siva Ranjani. U-Net based segmentation and multiple feature extraction of dermascopic images for efficient diagnosis of melanoma. In *Computer Aided Intervention and Diagnostics in Clinical and Medical Images*, pages 81–101. 2019.

[277] W. M. Rand. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[278] Hariharan Ravishankar, Rahul Venkataramani, Sheshadri Thiruvenkadam, Prasad Sudhakar, and Vivek Vaidya. Learning and incorporating shape models for semantic segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 203–211. Springer, 2017.

[279] Ekaterina Redekop and Alexey Chernyavskiy. Uncertainty-based method for improving poorly labeled segmentation datasets. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1831–1835. IEEE, 2021.

[280] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[281] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning*, pages 4334–4343. PMLR, 2018.

[282] Yuan Ren, Long Yu, Shengwei Tian, Junlong Cheng, Zhiqi Guo, and Yanhan Zhang. Serial attention network for skin lesion segmentation. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–12, 2021.

[283] Vinicius Ribeiro, Sandra Avila, and Eduardo Valle. Handling inter-annotator agreement for automated skin lesion segmentation. *arXiv:1906.02415*, 2019.

[284] Vinicius Ribeiro, Sandra Avila, and Eduardo Valle. Less is more: Sample selection and label conditioning improve skin lesion segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 738–739, 2020.

[285] Howard W Rogers, Martin A Weinstock, Steven R Feldman, and Brett M Coldiron. Incidence estimate of nonmelanoma skin cancer (keratinocyte carcinomas) in the us population, 2012. *JAMA dermatology*, 151(10):1081–1086, 2015.

[286] Torsten Rohlfing and Calvin R Maurer. Shape-Based Averaging. *IEEE Transactions on Image Processing*, 16(1):153–161, 2006.

[287] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241, 2015.

[288] Sara Ross-Howe and Hamid R Tizhoosh. The effects of image pre-and post-processing, wavelet decomposition, and local binary patterns on u-nets for skin lesion segmentation. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2018.

[289] Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Combalia, Stephen Dusza, Pascale Guitera, David Gutman, et al. A Patient-Centric Dataset of Images and Metadata for Identifying Melanomas Using Clinical Context. *Scientific Data*, 8:34, 2021.

[290] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning Representations by Back-Propagating Errors. *Nature*, 323(6088):533–536, 1986.

[291] T. Saba, M. A. Khan, A. Rehman, and S. L. Marie-Sainte. Region Extraction and Classification of Skin Cancer: A Heterogeneous Framework of Deep CNN Features Fusion and Reduction. *Journal of Medical Systems*, 43(9):289, 2019.

[292] TK Saj Sachin, V Sowmya, and KP Soman. Performance analysis of deep learning models for biomedical image segmentation. In *Deep Learning for Biomedical Applications*, pages 83–100. CRC Press, 2021.

[293] Omer Sagi and Lior Rokach. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249, 2018.

[294] Anindo Saha, Prem Prasad, and Abdullah Thabit. Leveraging adaptive color augmentation in convolutional neural networks for deep skin lesion segmentation. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 2014–2017. IEEE, 2020.

[295] Nurullah Şahin, Nuh Alpaslan, and Davut Hanbay. Robust optimization of SegNet hyperparameters for skin lesion segmentation. *Multimedia Tools and Applications*, pages 1–21, 2021.

[296] Shreshth Saini, Divij Gupta, and Anil Kumar Tiwari. Detector-segmentor network for skin lesion localization and segmentation. In *National Conference on Computer Vision, Pattern Recognition, Image Processing, and Graphics*, pages 589–599. Springer, 2019.

[297] Shreshth Saini, Young Seok Jeon, and Mengling Feng. B-segnet: branched-segmentor network for skin lesion segmentation. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 214–221, 2021.

[298] Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. Auto-context convolutional neural network (auto-net) for brain extraction in magnetic resonance imaging. *IEEE transactions on medical imaging*, 36(11):2319–2330, 2017.

[299] Md Sarker, Mostafa Kamal, Hatem A Rashwan, Mohamed Abdel-Nasser, Vivek Kumar Singh, Syeda Furruka Banu, Farhan Akram, Forhad UH Chowdhury, Kabir Ahmed Choudhury, Sylvie Chambon, et al. MobileGAN: Skin lesion segmentation using a lightweight generative adversarial network. *arXiv preprint arXiv:1907.00856*, 2019.

[300] Md Mostafa Kamal Sarker, Hatem A Rashwan, Farhan Akram, Syeda Furruka Banu, Adel Saleh, Vivek Kumar Singh, Forhad UH Chowdhury, Saddam Abdulwahab, Santiago Romani, Petia Radeva, et al. SLSDeep: Skin lesion segmentation based on dilated residual and pyramid pooling networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 21–29. Springer, 2018.

[301] Md Mostafa Kamal Sarker, Hatem A Rashwan, Farhan Akram, Vivek Kumar Singh, Syeda Furruka Banu, Forhad UH Chowdhury, Kabir Ahmed Choudhury, Sylvie Chambon, Petia Radeva, Domenec Puig, et al. SLSNet: Skin lesion segmentation using a lightweight generative adversarial network. *Expert Systems with Applications*, page 115433, 2021.

[302] G. Schaefer, M. I. Rajab, M. E. Celebi, and H. Iyatomi. Colour and Contrast Enhancement for Improved SkinLesion Segmentation. *Computerized Medical Imaging and Graphics*, 35(2):99–104, 2011.

[303] Robert E Schapire and Yoram Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297–336, 1999.

[304] Suman Sedai, Bhavna Antony, Dwarikanath Mahapatra, and Rahil Garnavi. Joint segmentation and uncertainty visualization of retinal layers in optical coherence tomography images using Bayesian deep learning. In *Computational Pathology and Ophthalmic Medical Image Analysis*, pages 219–227. Springer, 2018.

[305] Ahmed H Shahin, Karim Amer, and Mustafa A Elattar. Deep convolutional encoder-decoders with aggregated multi-resolution skip connections for skin lesion segmentation. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 451–454. IEEE, 2019.

[306] P. Shamsolmoali, M. Zareapoor, E. Granger, H. Zhou, R. Wang, M. E. Celebi, and J. Yang. Image Synthesis with Adversarial Networks: A Comprehensive Survey and Case Studies. *Information Fusion*, 72:126–146, 2021.

[307] M. Sharma, O. Saha, A. Sriraman, R. Hebbalaguppe, L. Vig, and S. Karande. Crowdsourcing for Chromosome Segmentation and Deep Classification. In *Proceedings of the IEEE Conference on Computer vision and Pattern Recognition Workshops*, pages 786–793, 2017.

[308] K. Shimizu, H. Iyatomi, M. E. Celebi, K. A. Norton, and M. Tanaka. Four-Class Classification of Skin Lesions with Task Decomposition Strategy. *IEEE Transactions on Biomedical Engineering*, 62(1):274–283, 2015.

[309] C. Shorten and T. M. Khoshgoftaar. A Survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1):60, 2019.

[310] M. Silveira, J. C. Nascimento, J. S. Marques, A. R. S. Marcal, T. Mendonca, S. Yamauchi, J. Maeda, and J. Rozeira. Comparison of Segmentation Methods for Melanoma Diagnosis in Dermoscopy Images. *IEEE Journal of Selected Topics in Signal Processing*, 3(1):35–45, 2009.

[311] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[312] Vivek Kumar Singh, Mohamed Abdel-Nasser, Hatem A Rashwan, Farhan Akram, Nidhi Pandey, Alain Lalande, Benoit Presles, Santiago Romani, and Domenec Puig. FCA-Net: Adversarial learning for skin lesion segmentation based on multi-scale features and factorized channel attention. *IEEE Access*, 7:130552–130565, 2019.

[313] Padhraic Smyth, Usama Fayyad, Michael Burl, Pietro Perona, and Pierre Baldi. Inferring Ground Truth from Subjective Labelling of Venus Images. In *Advances in Neural Information Processing Systems*, pages 1085–1092, 1995.

[314] Lei Song, Jianzhe Lin, Z Jane Wang, and Haoqian Wang. Dense-residual attention network for skin lesion segmentation. In *International Workshop on Machine Learning in Medical Imaging*, pages 319–327. Springer, 2019.

[315] Amira Soudani and Walid Barhoumi. An image-based segmentation recommender using crowdsourcing and transfer learning for skin lesion extraction. *Expert Systems with Applications*, 118:400–410, 2019.

[316] Sainbayar Sukhbaatar and Rob Fergus. Learning from noisy labels with deep neural networks. *arXiv preprint arXiv:1406.2080*, 2014.

[317] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *IEEE ICCV*, pages 843–852, 2017.

[318] Saeid Asgari Taghanaki, Kumar Abhishek, Joseph Paul Cohen, Julien Cohen-Adad, and Ghassan Hamarneh. Deep semantic segmentation of natural and medical images: a review. *Artificial Intelligence Review*, June 2020.

[319] Saeid Asgari Taghanaki, Kumar Abhishek, and Ghassan Hamarneh. Improved inference via deep input transfer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 819–827. Springer, 2019.

[320] Saeid Asgari Taghanaki, Noirin Duggan, Hillgan Ma, Xinchi Hou, Anna Celler, Francois Benard, and Ghassan Hamarneh. Segmentation-free direct tumor volume and metabolic activity estimation from PET scans. *Computerized Medical Imaging and Graphics*, 63:52–66, January 2018.

[321] A. A. Taha and A. Hanbury. Metrics for Evaluating 3D Medical Image Segmentation: Analysis, Selection, and Tool. *BMC Medical Imaging*, 15(1):29, 2015.

[322] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding. Embracing Imperfect Datasets: A Review of Deep Learning Solutions for Medical Image Segmentation. *Medical Image Analysis*, 63:101693, 2020.

[323] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. MnasNet: Platform-Aware Neural Architecture Search for Mobile. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2820–2828, 2019.

[324] Mingxing Tan and Quoc Le. EfficientNet: Rethinking Model Scaling for Convolutional neural Networks. In *Proceedings of the International Conference on Machine Learning*, pages 6105–6114, 2019.

[325] Teck Yan Tan, Li Zhang, Chee Peng Lim, Ben Fielding, Yonghong Yu, and Emma Anderson. Evolving ensemble models for image segmentation using enhanced particle swarm optimization. *IEEE access*, 7:34004–34019, 2019.

[326] Peng Tang, Qiaokang Liang, Xintong Yan, Shao Xiang, Wei Sun, Dan Zhang, and Gianmarc Coppola. Efficient skin lesion segmentation using separable-UNet with stochastic weight averaging. *Computer methods and programs in biomedicine*, 178:289–301, 2019.

[327] Peng Tang, Xintong Yan, Qiaokang Liang, and Dan Zhang. AFLN-DGCL: Adaptive feature learning network with difficulty-guided curriculum learning for skin lesion segmentation. *Applied Soft Computing*, 110:107656, 2021.

[328] Xianlun Tang, Jiangping Peng, Bing Zhong, Jie Li, and Zhenfu Yan. Introducing frequency representation into convolution neural networks for medical image segmentation via twin-kernel fourier convolution. *Computer Methods and Programs in Biomedicine*, 205:106110, 2021.

[329] Yujiao Tang, Feng Yang, Shaofeng Yuan, et al. A multi-stage framework with context information fusion structure for skin lesion segmentation. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1407–1410. IEEE, 2019.

[330] Xiaozhong Tong, Junyu Wei, Bei Sun, Shaojing Su, Zhen Zuo, and Peng Wu. Ascunet: Attention gate, spatial and channel attention u-net for skin lesion segmentation. *Diagnostics*, 11(3):501, 2021.

[331] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1521–1528, 2011.

[332] Hue Tran, Keng Chen, Adrian C Lim, James Jabbour, and Stephen Shumack. Assessing diagnostic skill in dermatology: a comparison between general practitioners and dermatologists. *Australasian journal of dermatology*, 46(4):230–234, 2005.

[333] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 Dataset, a Large Collection of Multi-Source Dermatoscopic Images of Common Pigmented Skin Lesions. *Scientific Data*, page 180161, 2018.

[334] Philipp Tschandl, Christoph Sinz, and Harald Kittler. Domain-specific classification-pretrained fully convolutional network encoders for skin lesion segmentation. *Computers in biology and medicine*, 104:111–116, 2019.

[335] Wenli Tu, Xiaoming Liu, Wei Hu, and Zhifang Pan. Dense-residual network with adversarial learning for skin lesion segmentation. *IEEE Access*, 7:77037–77051, 2019.

[336] Zhuowen Tu and Xiang Bai. Auto-context and its application to high-level vision tasks and 3D brain image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 32(10):1744–1757, 2009.

[337] R. Unnikrishnan, C. Pantofaru, and M. Hebert. Toward Objective Evaluation of Image Segmentation Algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):929–944, 2007.

[338] Halil Murat Ünver and Enes Ayan. Skin lesion segmentation in dermoscopic images with combination of YOLO and grabcut algorithm. *Diagnostics*, 9(3):72, 2019.

[339] Arash Vahdat. Toward robustness against label noise in training deep discriminative neural networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 5601–5610, 2017.

[340] Eduardo Valle, Michel Fornaciali, Afonso Menegola, Julia Tavares, Flávia Vasques Bittencourt, Lin Tzy Li, and Sandra Avila. Data, depth, and design: Learning reliable models for skin lesion analysis. *Neurocomputing*, 383:303–313, 2020.

[341] C. J. van Rijsbergen. *Information Retrieval*. Butterworth–Heinemann, Second edition, 1979.

[342] Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. Learning from noisy large-scale datasets with minimal supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 839–847, 2017.

[343] Olga Veksler. Star shape prior for graph-cut image segmentation. In *European Conference on Computer Vision*, pages 454–467. Springer, 2008.

[344] GM Venkatesh, YG Naresh, Suzanne Little, and Noel E O'Connor. A deep residual architecture for skin lesion segmentation. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, pages 277–284. Springer, 2018.

[345] Sulaiman Vesal, Shreyas Malakarjun Patil, Nishant Ravikumar, and Andreas K Maier. A multi-task framework for skin lesion detection and segmentation. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, pages 285–293. Springer, 2018.

[346] Sulaiman Vesal, Nishant Ravikumar, and Andreas Maier. SkinNet: A deep learning framework for skin lesion segmentation. In *2018 IEEE Nuclear Science Symposium and Medical Imaging Conference Proceedings (NSS/MIC)*, pages 1–3. IEEE, 2018.

[347] Hilke Vorwerk, Gabriele Beckmann, Michael Bremer, Maria Degen, Barbara Dietl, Rainer Fietkau, Tammo Gsänger, Robert Michael Hermann, Markus Karl Alfred Herrmann, Ulrike Höller, Michael van Kampen, Wolfgang Körber, Burkhard Maier, Thomas Martin, Michael Metz, Ronald Richter, Birgit Siekmeyer, Martin Steder, Daniela Wagner, Clemens Friedrich Hess, Elisabeth Weiss, and Hans Christiansen. The delineation of target volumes for radiotherapy of lung cancer patients. *Radiotherapy and Oncology*, 91(3):455–460, June 2009.

[348] Nhat Vu and BS Manjunath. Shape prior segmentation of multiple objects with graph cuts. In *Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[349] Huan Wang, Guotai Wang, Ze Sheng, and Shaoting Zhang. Automated segmentation of skin lesion based on pyramid attention network. In *International Workshop on Machine Learning in Medical Imaging*, pages 435–443. Springer, 2019.

[350] Min Wang, Baoyuan Liu, and Hassan Foroosh. Factorized convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 545–553, 2017.

[351] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. Understanding convolution for semantic segmentation. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 1451–1460. IEEE, 2018.

[352] Ruxin Wang, Shuyuan Chen, Jianping Fan, and Ye Li. Cascaded context enhancement for automated skin lesion segmentation. *arXiv preprint arXiv:2004.08107*, 2020.

[353] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[354] Xiaohong Wang, Henghui Ding, and Xudong Jiang. Dermoscopic image segmentation through the enhanced high-level parsing and class weighted loss. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 245–249. IEEE, 2019.

[355] Xiaohong Wang, Xudong Jiang, Henghui Ding, and Jun Liu. Bi-directional dermoscopic feature learning and multi-scale consistent decision fusion for skin lesion segmentation. *IEEE transactions on image processing*, 29:3039–3051, 2019.

[356] Xiaohong Wang, Xudong Jiang, Henghui Ding, Yuqian Zhao, and Jun Liu. Knowledge-aware deep framework for collaborative skin lesion segmentation and melanoma recognition. *Pattern Recognition*, page 108075, 2021.

[357] Yaxiong Wang, Yunchao Wei, Xueming Qian, Li Zhu, and Yi Yang. Donet: Dual objective networks for skin lesion segmentation. *arXiv preprint arXiv:2008.08278*, 2020.

[358] Yisen Wang, Weiyang Liu, Xingjun Ma, James Bailey, Hongyuan Zha, Le Song, and Shu-Tao Xia. Iterative learning with open-set noisy labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8688–8696, 2018.

[359] Simon K Warfield, Kelly H Zou, and William M Wells. Simultaneous Truth and Performance Level Estimation (STAPLE): An Algorithm for the Validation of Image Segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921, 2004.

[360] Lisheng Wei, Kun Ding, and Huosheng Hu. Automatic skin cancer detection in dermoscopy images based on ensemble lightweight deep learning network. *IEEE Access*, 8:99633–99647, 2020.

[361] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1568–1576, 2017.

[362] Zenghui Wei, Hong Song, Lei Chen, Qiang Li, and Guanghui Han. Attention-based DenseUnet network with adversarial training for skin lesion segmentation. *IEEE Access*, 7:136616–136629, 2019.

[363] Adi Wibowo, Satriawan Rasyid Purnama, Panji Wisnu Wirawan, and Hanif Rasyidi. Lightweight encoder-decoder model for automatic skin lesion segmentation. *Informatics in Medicine Unlocked*, page 100640, 2021.

[364] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2691–2699, 2015.

[365] Fengying Xie, Jiawen Yang, Jie Liu, Zhiguo Jiang, Yushan Zheng, and Yukun Wang. Skin lesion segmentation using high-resolution convolutional neural network. *Computer methods and programs in biomedicine*, 186:105241, 2020.

[366] Yutong Xie, Jianpeng Zhang, Yong Xia, and Chunhua Shen. A mutual bootstrapping model for automated skin lesion segmentation and classification. *IEEE Transactions on Medical Imaging*, 2020.

[367] Zhiqiang Xie, Enmei Tu, Hao Zheng, Yun Gu, and Jie Yang. Semi-supervised skin lesion segmentation with learning model confidence. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1135–1139. IEEE, 2021.

[368] Yuan Xue, Tao Xu, and Xiaolei Huang. Adversarial learning with multi-scale loss for skin lesion segmentation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 859–863. IEEE, 2018.

[369] Yiqi Yan, Jeremy Kawahara, and Ghassan Hamarneh. Melanoma recognition via visual attention. In *International Conference on Information Processing in Medical Imaging*, pages 793–804. Springer, 2019.

[370] Xulei Yang, Hangxing Li, Li Wang, Si Yong Yeo, Yi Su, and Zeng Zeng. Skin lesion analysis by multi-target deep neural networks. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1263–1266. IEEE, 2018.

[371] X. Yi, E. Walia, and P. Babyn. Generative Adversarial Network in Medical Imaging: A Review. *Medical Image Analysis*, 58:101552, 2019.

[372] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *international conference on learning representations*, 2016.

[373] Lequan Yu, Hao Chen, Qi Dou, Jing Qin, and Pheng-Ann Heng. Automated Melanoma Recognition in Dermoscopy Images via Very Deep Residual Networks. *IEEE Transactions on Medical Imaging*, 36(4):994–1004, 2017.

[374] Yang Yu, Zhiqiang Gong, Ping Zhong, and Jiaxin Shan. Unsupervised representation learning with deep convolutional neural network for remote sensing images. In *International Conference on Image and Graphics*, pages 97–108, 2017.

[375] Jing Yuan, Wu Qiu, Eranga Ukwatta, Martin Rajchl, Yue Sun, and Aaron Fenster. An efficient convex optimization approach to 3D prostate MRI segmentation with generic star shape prior. *Prostate MR Image Segmentation Challenge, International Conference on Medical image computing and computer-assisted intervention*, 2012.

[376] Yading Yuan. Automatic skin lesion segmentation with fully convolutional-deconvolutional networks. *arXiv preprint arXiv:1703.05165*, 2017.

[377] Yading Yuan, Ming Chao, and Yeh-Chi Lo. Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance. *IEEE transactions on medical imaging*, 36(9):1876–1886, 2017.

[378] Yading Yuan and Yeh-Chi Lo. Improving Dermoscopic Image Segmentation with Enhanced Convolutional-Deconvolutional Networks. *IEEE Journal of Biomedical and Health Informatics*, 23(2):519–526, 2019.

[379] Kashan Zafar, Syed Omer Gilani, Asim Waris, Ali Ahmed, Mohsin Jamil, Muhammad Nasir Khan, and Amer Sohail Kashif. Skin lesion segmentation from dermoscopic images using convolutional neural network. *Sensors*, 20(6):1601, 2020.

[380] Guodong Zeng and Guoyan Zheng. Multi-scale fully convolutional densenets for automated skin lesion segmentation in dermoscopy images. In *International Conference Image Analysis and Recognition*, pages 513–521. Springer, 2018.

[381] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. 2017.

[382] Guokai Zhang, Xiaoang Shen, Sirui Chen, Lipeng Liang, Ye Luo, Jie Yu, and Jianwei Lu. DSM: A deep supervised multi-scale network learning for skin cancer segmentation. *IEEE Access*, 7:140936–140945, 2019.

[383] H. Zhang, J. E. Fritts, and S. A. Goldman. Image Segmentation Evaluation: A Survey of Unsupervised Methods. *Computer Vision and Image Understanding*, 110(2):260–280, 2008.

[384] Jing Zhang, Caroline Petitjean, and Samia Ainouz. Kappa loss for skin lesion segmentation in fully convolutional network. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 2001–2004. IEEE, 2020.

[385] Lei Zhang, Guang Yang, and Xujiong Ye. Automatic skin lesion segmentation by coupling deep fully convolutional networks and shallow network with textons. *Journal of Medical Imaging*, 6(2):024001, 2019.

[386] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6848–6856, 2018.

[387] Yu Zhang and Qiang Yang. A survey on multi-task learning. *arXiv preprint arXiv:1707.08114*, 2017.

[388] Chen Zhao, Renjun Shuai, Li Ma, Wenjia Liu, and Menglin Wu. Segmentation of dermoscopy images based on deformable 3D convolution and ResU-NeXt++. *Medical & Biological Engineering & Computing*, 59(9):1815–1832, 2021.

[389] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid Scene Parsing Network. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2881–2890, 2017.

[390] Mengliu Zhao, Jeremy Kawahara, Kumar Abhishek, Sajjad Shamanian, and Ghassan Hamarneh. Skin3d: Detection and longitudinal tracking of pigmented skin lesions in 3d total-body textured meshes. *Medical Image Analysis*, 77:102329, 2022.

[391] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1529–1537, 2015.

[392] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[393] Liangjiu Zhu, Shuanglang Feng, Weifang Zhu, and Xinjian Chen. ASNet: An adaptive scale network for skin lesion segmentation in dermoscopy images. In *Medical Imaging 2020: Biomedical Applications in Molecular, Structural, and Functional Imaging*, volume 11317, pages 226–231. International Society for Optics and Photonics, SPIE, 2020.

[394] Qiuming Zhu. On the Performance of Matthews Correlation Coefficient (MCC) for Imbalanced Dataset. *Pattern Recognition Letters*, 136:71–80, 2020.

[395] Alex P Zijdenbos, Benoit M Dawant, Richard A Margolin, and Andrew C Palmer. Morphometric Analysis of White Matter Lesions in MR Images: Method and Validation. *IEEE Transactions on Medical Imaging*, 13(4):716–724, 1994.

[396] Maciel Zortea, Stein Olav Skrøvseth, Thomas R Schopf, Herbert M Kirchesch, and Fred Godtliebsen. Automatic segmentation of dermoscopic images by iterative classification. *International journal of biomedical imaging*, 2011, 2011.

[397] Kelly H Zou, Simon K Warfield, Aditya Bharatha, Clare MC Tempany, Michael R Kaus, Steven J Haker, William M Wells III, Ferenc A Jolesz, and Ron Kikinis. Statistical Validation of Image Segmentation Quality Based on a Spatial Overlap Index. *Academic Radiology*, 11(2):178–189, 2004.

[398] Hasib Zunair and A Ben Hamza. Sharp U-Net: Depthwise convolutional network for biomedical image segmentation. *Computers in Biology and Medicine*, 136:104699, 2021.