Medical Image Analysis 88 (2023) 102863

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Survey paper

A survey on deep learning for skin lesion segmentation

Zahra Mirikharaji [a,2], Kumar Abhishek [a,2], Alceu Bissoto [c], Catarina Barata [b], Sandra Avila [c], Eduardo Valle [d], M. Emre Celebi [e,*,3], Ghassan Hamarneh [a,*,3]

[a] Medical Image Analysis Lab, School of Computing Science, Simon Fraser University, Burnaby V5A 1S6, Canada
[b] Institute for Systems and Robotics, Instituto Superior Técnico, Avenida Rovisco Pais, Lisbon 1049-001, Portugal
[c] RECOD.ai Lab, Institute of Computing, University of Campinas, Av. Albert Einstein 1251, Campinas 13083-852, Brazil
[d] RECOD.ai Lab, School of Electrical and Computing Engineering, University of Campinas, Av. Albert Einstein 400, Campinas 13083-952, Brazil
[e] Department of Computer Science and Engineering, University of Central Arkansas, 201 Donaghey Ave., Conway, AR 72035, USA

ARTICLE INFO

Keywords:
Skin lesion
Deep learning
Segmentation
Survey

ABSTRACT

Skin cancer is a major public health problem that could benefit from computer-aided diagnosis to reduce the burden of this common disease. Skin lesion segmentation from images is an important step toward achieving this goal. However, the presence of natural and artificial artifacts (e.g., hair and air bubbles), intrinsic factors (e.g., lesion shape and contrast), and variations in image acquisition conditions make skin lesion segmentation a challenging task. Recently, various researchers have explored the applicability of deep learning models to skin lesion segmentation. In this survey, we cross-examine 177 research papers that deal with deep learning-based segmentation of skin lesions. We analyze these works along several dimensions, including input data (datasets, preprocessing, and synthetic data generation), model design (architecture, modules, and losses), and evaluation aspects (data annotation requirements and segmentation performance). We discuss these dimensions both from the viewpoint of select seminal works, and from a systematic viewpoint, examining how those choices have influenced current trends, and how their limitations should be addressed. To facilitate comparisons, we summarize all examined works in a comprehensive table as well as an interactive table available online[1].

1. Introduction

Segmentation is a challenging and critical operation in the automated skin lesion analysis workflow. Rule-based skin lesion diagnostic systems, popular in the clinical setting, rely on an accurate lesion segmentation for the estimation of diagnostic criteria such as asymmetry, border irregularity, and lesion size, as needed for implementing the ABCD algorithm (Asymmetry, Border, Color, Diameter of lesions) (Friedman et al., 1985; Nachbar et al., 1994) and its derivatives: ABCDE (ABCD plus Evolution of lesions) (Abbasi et al., 2004) and ABCDEF (ABCDE plus the "ugly duckling" sign) (Jensen and Elewski, 2015). By contrast, in machine learning-based diagnostic systems, restricting the areas within an image, thereby focusing the model on the interior of the lesion, can improve the robustness of the classification. For example, recent studies have shown the utility of segmentation in improving the deep learning (DL)-based classification performance for certain diagnostic categories by regularizing attention maps (Yan et al., 2019), allowing the cropping of lesion images (Yu et al., 2017a; Mahbod et al., 2020; Liu et al., 2020; Singh et al., 2023), tracking the evolution of lesions (Navarro et al., 2018) and the removal of imaging artifacts (Maron et al., 2021a; Bissoto et al., 2022). In a DL-based skin lesion classification framework, presenting the delineated skin lesion to the user can also help with interpreting the DL black box (Jaworek-Korjakowska et al., 2021), and thus may either instill trust, or raise suspicion, in computer-aided diagnosis (CAD) systems for skin cancer.

Lesion detection and segmentation are also useful as preprocessing steps when analyzing wide-field images with multiple lesions (Birkenfeld et al., 2020). Additionally, radiation therapy and image-guided human or robotic surgical lesion excision require localization and delineation of lesions (American Cancer Society, 2023). Ensuring fair diagnosis that is unbiased to minority groups, a pressing issue with the deployment of these models and the trust therein, requires the estimation of lesion-free skin tone, which in turn also relies upon the delineation of skin lesions (Kinyanjui et al., 2020). However, despite the importance of lesion segmentation, manual delineation of skin

* Corresponding authors.
E-mail addresses: ecelebi@uca.edu (M.E. Celebi), hamarneh@sfu.ca (G. Hamarneh).
[2] Joint first authors.
[3] Joint senior authors.
[1] https://github.com/sfu-mial/skin-lesion-segmentation-survey.

https://doi.org/10.1016/j.media.2023.102863
Received 24 May 2022; Received in revised form 1 February 2023; Accepted 31 May 2023
Available online 9 June 2023

lesions remains a laborious task that suffers from significant inter- and intra-observer variability and consequently, a fast, reliable, and automated segmentation algorithm is needed.

Skin cancer and its associated expenses, $8.1 billion annually in U.S. (Guy et al., 2015), have grown into a major public health issue in the past decades. In the USA alone, 97,610 new cases of melanoma are expected in 2023 (Siegel et al., 2023). 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 diagnosis is critical for a good prognosis: melanoma can be cured with a simple outpatient surgery if detected early, but its five-year survival rate drops from over 99% to 32% if it is diagnosed at an advanced stage (American Cancer Society, 2023).

Two imaging modalities are commonly employed in automated skin lesion analysis (Daneshjou et al., 2022): dermoscopic (microscopic) images and clinical (macroscopic) images. While dermoscopic images allow the inspection of lesion properties that are invisible to the naked eye, they are not always accessible even to dermatologists (Engasser and Warshaw, 2010). On the other hand, clinical images acquired using conventional cameras are easily accessible but suffer from lower quality. Dermoscopy is a non-invasive skin imaging technique that aids in the diagnosis of skin lesions by allowing dermatologists to visualize sub-surface structures (Kittler et al., 2002). However, even with dermoscopy, diagnostic accuracy can vary widely, ranging from 24% to 77%, depending on the clinician's level of expertise (Tran et al., 2005). Moreover, dermoscopy may actually lower the diagnostic accuracy in the hands of inexperienced dermatologists (Binder et al., 1995). Therefore, to minimize the diagnostic errors that result from the difficulty and the subjectivity of visual interpretation and to reduce the burden of skin diseases and limited access to dermatologists, the development of CAD systems is crucial.

Segmentation is the partitioning of an image into meaningful regions. Semantic segmentation, in particular, assigns appropriate class labels to each region. For skin lesions, the task is almost always binary, separating the lesion from the surrounding skin. Automated skin lesion segmentation is hindered by illumination and contrast issues, intrinsic inter-class similarities and intra-class variability, occlusions, artifacts, and the diversity of imaging tools used. The lack of large datasets with ground-truth segmentation masks generated by experts compounds the problem, impeding both the training of models and their reliable evaluation. Skin lesion images are occluded by natural artifacts such as hair (Fig. 1(a)), blood vessels (Fig. 1(b)), and artificial ones such as surgical marker annotations (Fig. 1(c)), lens artifacts (dark corners) (Fig. 1(d)), and air bubbles (Fig. 1(e)). Intrinsic factors such as lesion size and shape variation (Figs. 1(f) and 1(g)), different skin colors (Fig. 1(h)), low contrast (Fig. 1(i)), and ambiguous boundaries (Fig. 1(h)) complicate the automated segmentation of skin lesions.

Before the deep learning revolution, segmentation was based on classical image processing and machine learning techniques such as adaptive thresholding (Green et al., 1994; Celebi et al., 2013), active contours (Erkol et al., 2005), region growing (Iyatomi et al., 2006; Celebi et al., 2007a), unsupervised clustering (Gómez et al., 2007), and support vector machines (Zortea et al., 2011). These approaches depend on hand-crafted features, which are difficult to engineer and often limit invariance and discriminative power from the outset. As a result, such conventional segmentation algorithms do not always perform well on larger and more complex datasets. In contrast, DL integrates feature extraction and task-specific decision seamlessly, and does not just cope with, but actually requires larger datasets.

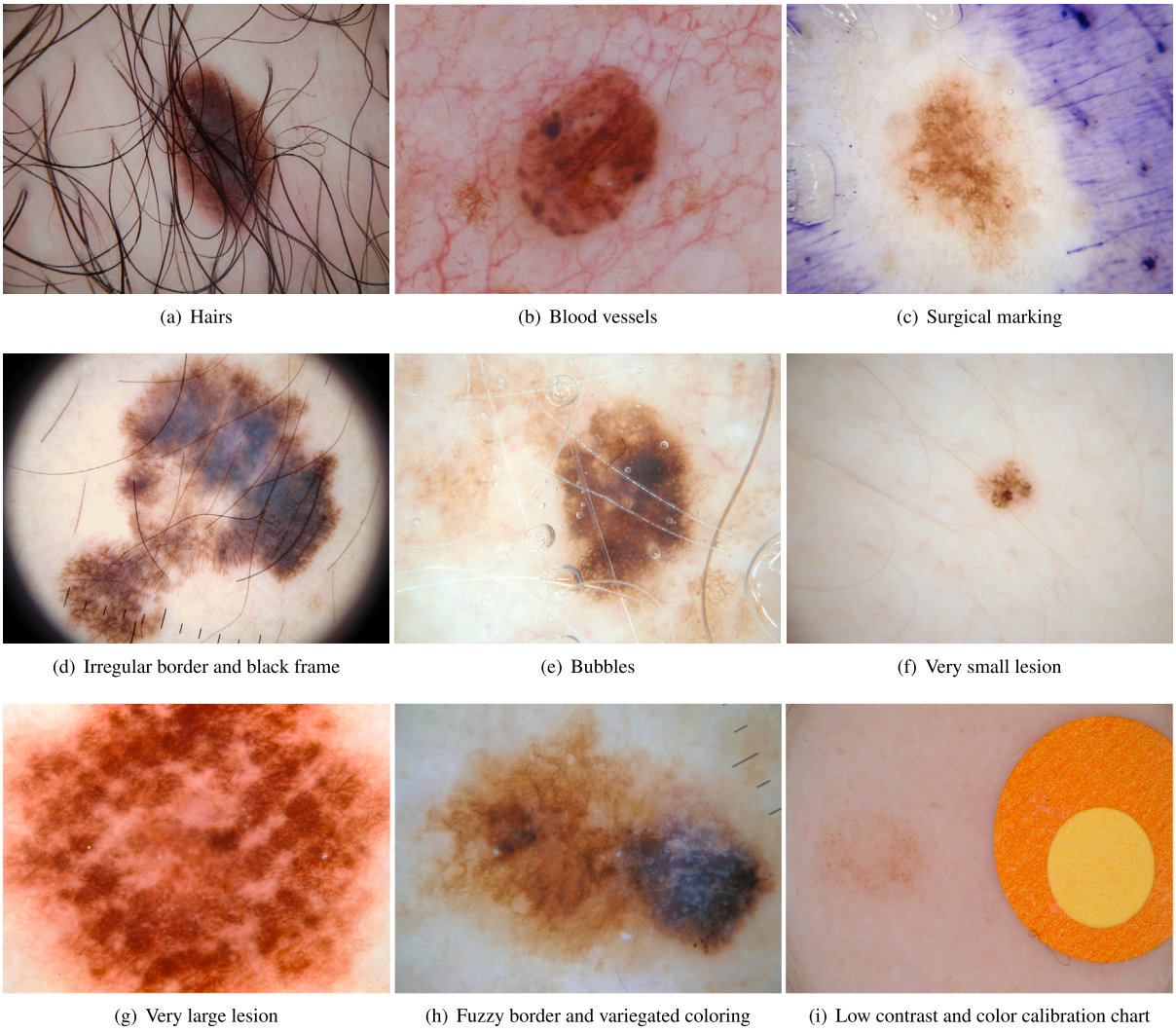
Survey of surveys. Celebi et al. (2009b) reviewed 18 skin lesion segmentation algorithms for dermoscopic images, published between 1998 and 2008, with their required preprocessing and postprocessing steps. Celebi et al. (2015b) later extended their work with 32 additional algorithms published between 2009 and 2014, discussing performance evaluation and computational requirements of each approach, and suggesting guidelines for future works. Both surveys appeared before DL was widely adopted for skin lesion segmentation, but cover all the important works based on classical image processing and machine learning. Adegun and Viriri (2020a) reviewed the literature on DL-based skin image analysis, with an emphasis on the best-performing algorithms in the ISIC (International Skin Imaging Collaboration) Skin Image Analysis Challenges 2018 (Codella et al., 2019) and 2019 (Tschandl et al., 2018; Codella et al., 2018; Combalia et al., 2019). However, since their review focused on the ISIC Challenges 2018 and 2019, it is more general as it covers both lesion classification and segmentation. Consequently, the number of papers surveyed for skin lesion segmentation by Adegun and Viriri (2020a) is almost an order of magnitude smaller than that in this review.

Main contributions. No existing survey approaches the present work in breadth or depth, as we cross-examine 177 research papers that deal with the automated segmentation of skin lesions in clinical and dermoscopic images. We analyze the works along several dimensions, including input data (datasets, preprocessing, and synthetic data generation), model design (architecture, modules, and losses), and evaluation (data annotation and evaluation metrics). We discuss these dimensions both from the viewpoint of select seminal works, and from a systematic viewpoint, examining how those choices have influenced current trends, and how their limitations should be addressed. We summarize all examined works in a comprehensive table to facilitate comparisons.

Search strategy. We searched DBLP and Arxiv Sanity Preserver for all scholarly publications: peer-reviewed journal papers, papers published in the proceedings of conferences or workshops, and non-peer-reviewed preprints from 2014 to 2022. The DBLP search query was `(conv* | trans* | deep | neural | learn*) (skin | derm*) (segment* | delineat* | extract* | localiz*)`, thus restricting our search to DL-based works involving skin and segmentation. We use 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 unpublished preprints, we also searched on Arxiv Sanity Preserver using a similar query. [4] We filtered our search results to remove false positives (31 papers) and included only papers related to skin lesion segmentation. We excluded papers that focused on general skin segmentation and general skin conditions (e.g., psoriasis, acne, or certain sub-types of skin lesions). We also included unpublished preprints from arXiv, which (a) passed minimum quality checks levels and (b) had at least 10 citations, and excluded those that were 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. This led to the filtering out of papers of visibly low quality ((a–c) criteria above; 18 papers) and those with no methodological contribution (20 papers).

The remaining text is organized as follows: in Section 2, we introduce the publicly available datasets and discuss preprocessing and synthetic data generation; in Section 3, we review the various network architectures used in deep segmentation models and discuss how deep models benefit from these networks. We also describe various loss functions designed either for general use or specifically for skin lesion segmentation. In Section 4, we detail segmentation evaluation techniques and measures. Finally, in Section 5, we discuss the open challenges in DL-based skin lesion segmentation and conclude our survey. A visual overview of the structure of this survey is presented in Fig. 2.

[4] **Arxiv Sanity Preserver**: https://www.arxiv-sanity-lite.com/search?q=segmentation+skin+melanoma+deep+learning+convolution+transformer.



(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

Fig. 1. Factors that complicate dermoscopy image segmentation.
Image source: ISIC 2016 dataset (Gutman et al., 2016).

2. Input data

Obtaining data in sufficient quantity and quality is often a significant obstacle to developing effective segmentation models. State-of-the-art segmentation models have a huge number of adjustable parameters that allow them to generalize well, provided they are trained on massive labeled datasets (Sun et al., 2017; Buslaev et al., 2020). Unfortunately, skin lesion datasets—like most medical image datasets (Asgari Taghanaki et al., 2021)—tend to be small (Curiel-Lewandrowski et al., 2019) due to issues such as copyright, patient privacy, acquisition and annotation cost, standardization, and scarcity of many pathologies of interest. The two most common modalities used in the training of skin lesion segmentation models are *clinical images*, which are close-ups of the lesions acquired using conventional cameras, and *dermoscopic images*, which are acquired using dermoscopy, a non-invasive skin imaging through optical magnification, and either liquid immersion and low angle-of-incidence lighting, or cross-polarized lighting. Dermoscopy eliminates skin surface reflections (Kittler et al., 2002), 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 (Menzies et al., 2003).

Annotation is often the greatest barrier for increasing the amount of data. Objective evaluation of segmentation often requires laborious *region-based annotation*, in which an expert manually outlines the region where the lesion (or a clinical feature) appears in the image. By contrast, *textual annotation* may involve 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), and other details (e.g., timestamp, camera model) (Caffery et al., 2018). We extensively discuss the image annotation issue in Section 4.1.

2.1. Datasets

The availability of larger, more diverse, and better-annotated datasets is one of the main driving factors for the advances in skin image analysis in the past decade (Marchetti et al., 2018; Celebi et al., 2019). Works in skin image analysis date back to the 1980s (Vanker and Van Stoecker, 1984; Dhawan et al., 1984), but until the mid-2000s, these works used small, private datasets, containing a few hundred images.

The *Interactive Atlas of Dermoscopy* (sometimes called the *Edra Atlas*, in reference to the publisher) by Argenziano et al. (2000) included a CD-ROM with 1039 dermoscopy images (26% melanomas, 4% carcinomas, 70% nevi) of 1024×683 pixels, acquired by three European

Fig. 2. An overview of the various components of this review. We structure the review based on the different elements of a DL-based segmentation pipeline and conclude it with discussions on future potential research directions.

university hospitals (University of Graz, Austria, University of Naples, Italy, and University of Florence, Italy). The works of Celebi et al. (2007b, 2008) 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) became available. Recently, Kawahara et al. (2019) placed this valuable dataset, along with additional textual annotations based on the 7-point checklist, in public domain under the name *derm7pt*. Shortly after the publication of the Interactive Atlas of Dermoscopy, Menzies et al. (2003) 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 at the Sydney Melanoma Unit, Australia.

The PH^2 dataset, released by Mendonca et al. (2013) and detailed by Mendonca et al. (2015), was the first public dataset to provide region-based annotations with segmentation masks, and masks for the clinically significant colors (white, red, light brown, dark brown, blue-gray, and black) present in the images. The dataset contains 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 (Ballerini et al., 2013) also provides region-based annotations for 1300 clinical images (10 diagnostic categories including melanomas, seborrhoeic keratosis, and basal cell carcinoma) of sizes ranging from 177 × 189 to 2176 × 2549 pixels. The images were acquired with a Canon EOS 350D SLR camera, in controlled lighting and at a consistent distance from the lesions, resulting in a 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" (ISIC, 2023). At the time of writing, the archive contains more than 240,000 images, of which more than 71,000 are publicly available. These images were acquired in leading worldwide clinical centers, using a variety of devices.

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 featured only classification. Each subset is associated with a challenge (year), one or more tasks, and has two (training/test) or three (training/validation/test) splits. The ISIC Challenge 2016 (Gutman et al., 2016) (ISIC 2016, for brevity) contains 1279 images split into 900 for training (19% melanomas, 81% nevi), and 379 for testing (20% melanomas, 80% nevi). There is a large variation in image size, ranging from 0.5 to 12 megapixels. All tasks used the same images. The ISIC 2017 (Codella et al., 2018) dataset more than doubled, with 2750 images split into 2000 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, ranging from 0.5 to 29 megapixels, and all tasks used the same images.

Table 1
Public skin lesion datasets with lesion segmentation annotations. All the datasets contain RGB images of skin lesions.

Dataset	Year	Modality	Size	Training/validation/test	Class distribution	Additional info
DermQuest[a] (DermQuest, 2012)	2012	Clinical	137	–	61 non-melanomas 76 melanomas	Acquired with different cameras under various lighting conditions
DermoFit (Ballerini et al., 2013)	2013	Clinical	1300	–	1224 non-melanomas 76 melanomas	Sizes ranging from 177 × 189 to 2176 × 2549 pixels
Pedro Hispano Hospital (PH^2) (Mendonca et al., 2013)	2013	Dermoscopy	200	–	160 benign nevi 40 melanomas	Sizes ranging from 553 × 763 to 577 × 769 pixels acquired at 20 × magnification
ISIC2016 (Gutman et al., 2016)	2016	Dermoscopy	1279	900/–/379	Training: 727 non-melanomas 173 melanomas Test: 304 non-melanomas 75 melanomas	Sizes ranging from 566 × 679 to 2848 × 4288 pixels
ISIC2017 (Codella et al., 2018)	2017	Dermoscopy	2750	2000/150/600	Training: 1626 non-melanomas 374 melanomas Test: 483 non-melanomas 117 melanomas	Sizes ranging from 540 × 722 to 4499 × 6748 pixels
ISIC2018 (Codella et al., 2019)	2018	Dermoscopy	3694	2594/100/1000	–	Sizes ranging from 540 × 576 to 4499 × 6748 pixels
HAM10000 (Tschandl et al., 2018, 2020; ViDIR Dataverse, 2020)	2020	Dermoscopy	10,015	–	1113 non-melanomas 8902 melanomas	All images of 600 × 450 pixels

[a]DermQuest was deactivated on December 31, 2019. However, 137 of its images are publicly available (Glaister, 2013).

ISIC 2018 provided, for the first time, separate datasets for the tasks, with 2594 training (20% melanomas, 72% nevi, and 8% seborrheic keratoses) and 100/1000 for validation/test images ranging from 0.5 to 29 megapixels, for the tasks of segmentation and feature extraction (Codella et al., 2019), and 10,015/1512 training/test images for the classification task, all with 600 × 450 pixels. The training dataset for classification was the HAM10000 dataset (Tschandl et al., 2018), 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 categories: 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%). As a part of a 2020 study of human–computer collaboration for skin lesion diagnosis involving dermatologists and general practitioners (Tschandl et al., 2020), the lesions in the HAM10000 dataset were segmented by a single dermatologist and consequently released publicly (ViDIR Dataverse, 2020), making this the single largest publicly available skin lesion segmentation dataset (Table 1).

ISIC 2019 (Codella et al., 2018; Tschandl et al., 2018; Combalia et al., 2019) contains 25,331 training 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 8238 test images (diagnostic distribution unknown). The images range from 600 × 450 to 1024 × 1024 pixels.

ISIC 2020 (Rotemberg et al., 2021) contains 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 (USA), the Melanoma Institute, the Sydney Melanoma Diagnostic Centre, and the University of Queensland (Australia), the Medical University of Vienna (Austria), the University of Athens (Greece), and the Hospital Clinic Barcelona (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 appearances are atypical for a given patient, and which present an increased risk of malignancy (Gachon et al., 2005).

There is, however, an overlap among these ISIC Challenge datasets. Abhishek (2020) analyzed all the lesion segmentation datasets from the ISIC Challenges (2016–2018) and found considerable overlap between these 3 datasets, with as many as 1940 images shared between at least 2 datasets and 706 images shared between all 3 datasets. In a more recent analysis of the ISIC Challenge datasets for the lesion diagnosis task from 2016 through 2020, Cassidy et al. (2022) found overlap between the datasets as well as the presence of duplicates within the datasets. Using a duplicate removal strategy, they curated a new set of 45,590 training images (8.61% melanomas, 91.39% others) and 11,397 validation images (8.61% melanomas, 91.39% others), leading to a total of 56,987 images. Additionally, since the resulting dataset is highly imbalanced (melanomas versus others in a ratio of 1 : 10.62), the authors also curated a balanced dataset with 7848 training images (50% melanoma, 50% others) and 1962 validation images (50% melanoma, 50% others).

Table 1 shows a list of publicly available skin lesion datasets with pixel-wise annotations, image modality, sample size, original split sizes, and diagnostic distribution. Fig. 3 shows how frequently these datasets appear in the literature. It is also worth noting that several other skin lesion image datasets have not been described in our survey as they do not provide the corresponding skin lesion segmentation annotations. However, these datasets, including SD-198 (Sun et al., 2016), MED-NODE (Giotis et al., 2015), derm7pt (Kawahara et al., 2019), Interactive Dermatology Atlas (Usatine and Madden, 2013), Dermatology Information System (DermIS, 2012), DermWeb (Lui et al., 2009), DermNet New Zealand (Oakley et al., 1995), may still be relevant for skin lesion segmentation research (see Section 5).

Biases in computer vision datasets are a constant source of issues (Torralba and Efros, 2011), which is compounded in medical imaging due to the smaller number of samples, insufficient image resolution, lack of geographical or ethnic diversity, or statistics unrepresentative of clinical practice. All existing skin lesion datasets suffer to a certain extent from one or more of the aforementioned issues, to which we add the specific issue of the availability and reliability of annotations. For lesion classification, many samples lack the gold standard histopathological confirmation, and ground-truth segmentation,

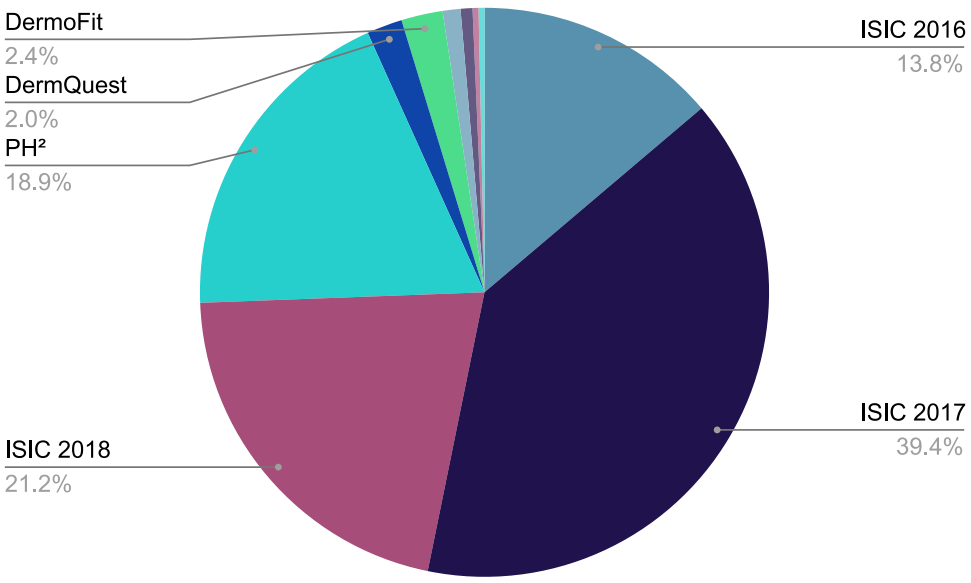


Fig. 3. The frequency of utilization of different skin lesion segmentation datasets in the surveyed studies. We found that 82 papers evaluated on more than 1 dataset, with 36 papers opting for cross-dataset evaluation (CDE in Table 3). ISIC datasets (ISIC 2016, ISIC 2017, ISIC 2018, and ISIC Archive) are used in the majority of papers, with 168 of 177 papers using at least one ISIC dataset and the ISIC 2017 dataset being the most popular (117 papers). The PH^2 dataset is the second most widely used (56 papers) following ISIC datasets.

even when available, is inherently noisy (Section 4.2). The presence of artifacts (Fig. 1) may lead to spurious correlations, an issue that Bissoto et al. (2019) attempted to quantify for classification models.

2.2. Synthetic data generation

Data augmentation—synthesizing new samples from existing ones—is commonly employed in the training of DL models. Augmented data serve as a regularizer, increase the amount and diversity of data (Shorten and Khoshgoftaar, 2019), induce desirable invariances on the model, and 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. Fig. 4 exemplifies the procedure, as applied to a dermoscopic image with Albumentations (Buslaev et al., 2020), a state-of-the-art open-source library for image augmentation.

As mentioned earlier, augmented training data induce invariance on the models: random translations and croppings, for example, help induce a translation-invariant model. This has implications for skin lesion analysis, e.g., data augmentation for generic datasets (such as ImageNet Deng et al., 2009) forgo vertical mirroring and large-angle rotations, because natural scenes have a strong vertical anisotropy, while skin lesion images are isotropic. In addition, augmented *test* data (test-time augmentation) may also improve generalization by combining the predictions of several augmented samples through, for example, average pooling or majority voting (Shorten and Khoshgoftaar, 2019). Perez et al. (2018) have systematically evaluated the effect of several data augmentation schemes for skin lesion classification, finding that the use of both training and test augmentation is critical for performance, surpassing, in some cases, increases of real data without augmentation. Valle et al. (2020) found, in a very large-scale experiment, that test-time augmentation was the second most influential factor for classification performance, after training set size. No systematic study of this kind exists for skin lesion segmentation.

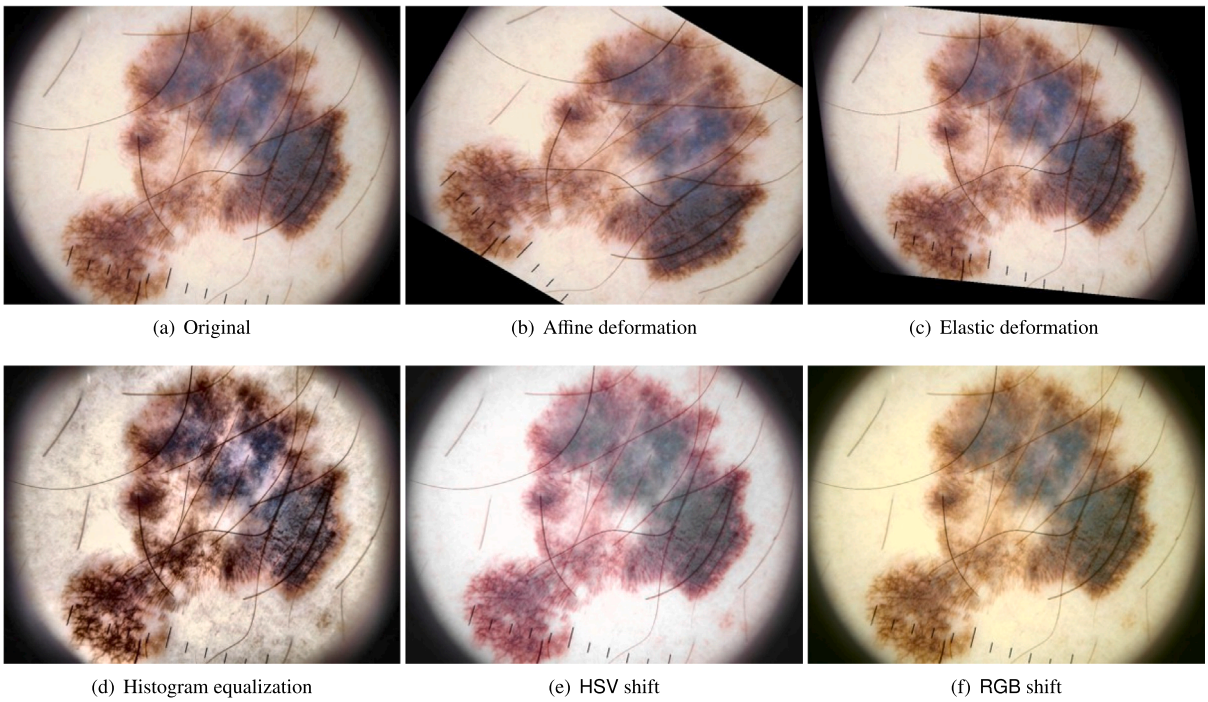
Although traditional data augmentation is crucial for training DL models, it falls short of providing samples that are both diverse and plausible from the same distribution as real data. Thus, modern data augmentation (Tajbakhsh et al., 2020) employs generative modeling, learning the probability distribution of the real data, and sampling from that distribution. Generative adversarial networks (GANs) (Goodfellow et al., 2020) are the most promising approach in this direction (Shorten and Khoshgoftaar, 2019), especially for medical image analysis (Yi et al., 2019; Kazeminia et al., 2020; Shamsolmoali et al., 2021). GANs employ an adversarial training between a generator, which attempts to generate realistic fake samples, and a discriminator, which attempts to differentiate real samples from fake ones. When the procedure converges, the generator output is surprisingly convincing, but GANs are computationally expensive and difficult to train (Creswell et al., 2018).

Synthetic generation of skin lesions has received some recent interest, especially in the context of improving classification. Works can be roughly divided into those that use GANs to create new images from a Gaussian latent variable (Baur et al., 2018; Pollastri et al., 2020; Abdelhalim et al., 2021), and those that implement GANs based on image-to-image translation (Abhishek and Hamarneh, 2019; Bissoto et al., 2018; Ding et al., 2021).

Noise-based GANs, such as DCGAN (Yu et al., 2017b), LAPGAN (Denton et al., 2015), and PGAN (Karras et al., 2018), learn to decode a Gaussian latent variable into an image that belongs to the training set distribution. The main advantage of these 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 of lower quality, and, in the case of segmentation, one needs to generate plausible pairs of images and segmentation masks.

Image-to-image translation GANs, such as pix2pix (Isola et al., 2017) and pix2pixHD (Wang et al., 2018), 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 of higher quality (or more "plausible"). There is no need to generate separate segmentation maps because the generated image is intrinsically compatible with the input segmentation map.

The two seminal papers on GANs for skin lesions (Baur et al., 2018; Bissoto et al., 2018) evaluate several models. Baur et al. (2018)



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

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

Fig. 4. Various data augmentation transformations applied to a dermoscopic image (image source: ISIC 2016 dataset Gutman et al., 2016) using the Albumentations library (Buslaev et al., 2020).

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 gives considerably better results. They further examine the PGAN against a panel of human judges, composed by dermatologists and DL experts, in a "visual Turing test", showing that both had difficulties in distinguishing the fake images from the true ones. Bissoto et al. (2018) adapt the PGAN to be class-conditioned on diagnostic category, and the image-to-image pix2pixHD to employ the semantic annotation provided by the feature extraction task of the ISIC 2018 dataset (Table 1), comparing those to an unmodified DCGAN on 256 × 256-pixel images, and finding the modified pix2pixHD to be qualitatively better. They use the performance improvement on a separate classification network as a quantitative metric, finding that the use of samples from both PGAN and pix2pixHD leads to the best improvements. They also showcase images of size up to 1024 × 1024 pixels generated by the pix2pixHD-derived model.

Pollastri et al. (2020) extended DCGAN and LAPGAN architectures to generate the segmentation masks (in the pairwise scheme explained above), making their work the only noise-based GANs usable for segmentation to date. Bi et al. (2019a) introduced stacked adversarial learning to GANs to learn class-specific skin lesion image generators given the ground-truth segmentations. Abhishek and Hamarneh (2019) employ pix2pix to translate a binary segmentation mask into a dermoscopic image and use the generated image–mask pairs to augment skin lesion segmentation training datasets, improving segmentation performance.

Ding et al. (2021) feed a segmentation mask and an instance mask to a conditional GAN generator, where the instance mask states the diagnostic category to be synthesized. In both cases, the discriminator receives different resolutions of the generated image and is required to make a decision for each of them. Abdelhalim et al. (2021) is a recent work that also conditions PGAN on the class label and uses the generated outputs to augment a melanoma diagnosis dataset.

Recently, Bissoto et al. (2021) cast doubt on the power of GAN-synthesized data augmentation to reliably improve skin lesion classification. Their evaluation, which included four GAN models, four datasets, and several augmentation scenarios, showed improvement only in a severe cross-modality scenario (training on dermoscopic and testing on clinical images). To the best of our knowledge, no corresponding systematic evaluation exists for skin lesion segmentation.

2.3. Supervised, semi-supervised, weakly supervised, self-supervised learning

Although supervised DL has achieved outstanding performance in various medical image analysis applications, its dependency on high-quality annotations limits its applicability, as well as its generalizability to unseen, out-of-distribution data. 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 conjunction with a subset of pixel-level fully-annotated samples.

Since pixel-level annotation of skin lesion images is costly, there is a trade-off between annotation precision 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 Section 4.2.) To deal with label noise, Mirikharaji et al. (2019) learn a model robust to annotation noise, making use of a large set of unreliable annotations and a small set of perfect clean annotations. They propose to learn a spatially adaptive weight map corresponding to each training data, assigning different weights to noisy and clean pixel-level annotations while training the deep model. To remove the dependency on having a set of perfectly clean annotations, Redekop and Chernyavskiy (2021) propose to alter noisy ground-truth masks during training by considering the quantification of aleatoric uncertainty (Der Kiureghian and Ditlevsen, 2009; Gal, 2016; Depeweg et al., 2018; Kwon et al., 2020) 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. (2020) 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.

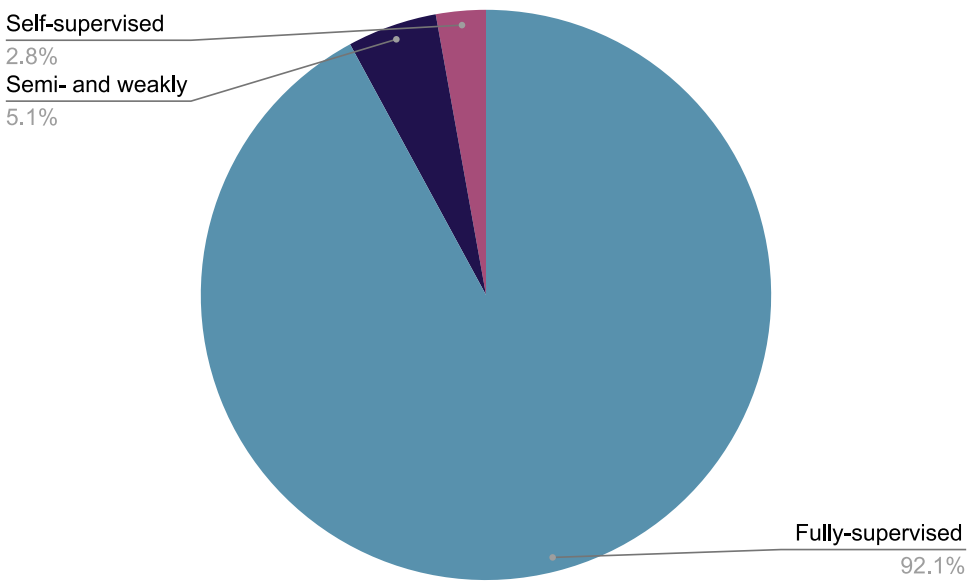


Fig. 5. A breakdown of different levels of supervision used in the 177 surveyed works. Fully supervised models continue to make up the majority of the literature (163 papers), with semi-supervised and weakly supervised methods appearing in only 9 papers. Self-supervision in skin lesion segmentation is fairly new, with all the 5 papers appearing from 2020 onwards.

When there is a labeled dataset, even if the number of labeled samples is far less than that of unlabeled samples, semi- and self-supervision techniques can be applied. Li et al. (2021c) propose a semi-supervised approach, using a transformation-consistent self-ensemble to leverage unlabeled data and to regularize the model. They minimize 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 on 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 (Gidaris et al., 2018). 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. (2020b) 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.

2.4. Image preprocessing

Preprocessing may facilitate the segmentation of skin lesion images. Typical preprocessing operations include:

- **Downsampling**: Dermoscopy is typically a high-resolution technique, resulting in large image sizes, while many convolutional neural network (CNN) architectures, e.g., LeNet, AlexNet, VGG, GoogLeNet, ResNet, etc., 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 (FCNs)) may benefit from downsampling for computational reasons. Downsampling is common in the skin lesion segmentation literature (Codella et al., 2017; Yu et al., 2017a; Yuan et al., 2017; Al-Masni et al., 2018; Zhang et al., 2019b; Pollastri et al., 2020).
- **Color space transformations**: RGB images are expected by most models, but some works (Codella et al., 2017; Al-Masni et al., 2018; Yuan and Lo, 2019; Pollastri et al., 2020; Pour and Seker, 2020) employ alternative color spaces (Busin et al., 2008), such as CIELAB, CIELUV, and HSV. Often, one or more channels of the transformed space are combined with the RGB channels for reasons including, but not limited to, increasing the class separability, decoupling luminance and chromaticity, ensuring (approximate) perceptual uniformity, achieving invariance to illumination or viewpoint, and eliminating highlights.
- **Additional inputs**: In addition to color space transformations, recent works incorporate more focused and domain-specific inputs to the segmentation models, such as Fourier domain representation using the discrete Fourier transform (Tang et al., 2021b) and inputs based on the physics of skin illumination and imaging (Abhishek et al., 2020).
- **Contrast enhancement**: Insufficient contrast (Fig. 1(i)) is a prime reason for segmentation failures (Bogo et al., 2015), leading some works (Saba et al., 2019; Schaefer et al., 2011) to enhance the image contrast prior to segmentation.
- **Color normalization**: Varying illumination (Barata et al., 2015a,b) may lead to inconsistencies in skin lesion segmentation. This problem can be addressed by color normalization (Goyal et al., 2019b).
- **Artifact removal**: Dermoscopic images often present artifacts, among which hair (Fig. 1(g)) is the most distracting (Abbas et al., 2011), leading some studies (Ünver and Ayan, 2019; Zafar et al., 2020; Li et al., 2021b) to remove it prior to segmentation.

Classical machine learning models (e.g., nearest neighbors, decision trees, support vector machines Celebi et al., 2007b, 2008; Iyatomi et al., 2008; Barata et al., 2014; Shimizu et al., 2015), which rely on hand-crafted features (Barata et al., 2019), tend to benefit more from preprocessing than DL models, which, when properly trained, can learn from the data how to bypass input issues (Celebi et al., 2015a; Valle et al., 2020). However, preprocessing may still be helpful when dealing with small or noisy datasets.

3. Model design and training

Multi-layer perceptrons (MLPs) for pixel-level classification (Gish and Blanz, 1989; Katz and Merickel, 1989) appeared soon after the publication of the seminal backpropagation paper (Rumelhart et al., 1986), but these shallow feed-forward networks had many drawbacks (LeCun

et al., 1998), including an excessive number of parameters, lack of invariance, and disregard for the inherent structure present in images.

CNNs are deep feedforward neural networks designed to extract progressively more abstract features from multidimensional signals (1-D signals, 2-D images, 3-D video, etc.) (LeCun et al., 2015). Therefore, in addition to addressing the aforementioned problems of MLPs, CNNs automate *feature engineering* (Bengio et al., 2013), that is, the design of algorithms that can transform raw signal values to discriminative features. Another advantage of CNNs over traditional machine learning classifiers is that they require minimal preprocessing of the input data. Due to their significant advantages, CNNs have become the method of choice in many medical image analysis applications over the past decade (Litjens et al., 2017). The key enablers in this deep learning revolution were: (i) the availability of massive data sets; (ii) the availability of powerful and inexpensive graphics processing units; (iii) the development of better network architectures, learning algorithms, and regularization techniques; and (iv) the development of open-source deep learning frameworks.

Semantic segmentation may be understood as the attempt to answer the parallel and complementary questions "what" and "where" in a given image. 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 first appeared in the mid-2000s (Ning et al., 2005), but their use accelerated after the seminal paper on FCNs by Long et al. (2015), which, along with U-Net (Ronneberger et al., 2015), have become 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.

3.1. Architecture

An ideal skin lesion segmentation algorithm is accurate, computationally inexpensive, invariant to noise and input transformations, requires little training data and is easy to implement and train. Unfortunately, no algorithm has, so far, been able to achieve these conflicting goals. DL-based segmentation tends towards accuracy and invariance at the cost of computation and training data. Ease of implementation is debatable: on the one hand, the algorithms often forgo cumbersome preprocessing, postprocessing, and feature engineering steps. On the other hand, tuning and optimizing them is often a painstaking task.

As shown in Fig. 6, we have classified the existing literature into single-network models, multiple-network models, hybrid-feature models, and Transformer models. 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 performing simultaneous classification and segmentation), and GANs. Hybrid-feature models combine DL with hand-crafted features. Transformer models, as the name suggests, employ Transformers either with or without CNNs for segmentation, and have started being used for skin lesion segmentation only recently. We classified works according to their most relevant feature, but the architectural improvements discussed in Section 3.1.1 also appear in the models listed in the other sections. In Fig. 7, we show how frequently different architectural modules appear in the 177 surveyed works, grouped by our taxonomy of model architectures (Fig. 6).

Table 3 summarizes all the 177 surveyed works in this review, with the following attributes for each work: type of publication, datasets, architectural modules, loss functions, and augmentations used, reported Jaccard index, whether the paper performed cross-dataset evaluation (CDE) and postprocessing (PP), and whether the corresponding code was released publicly. For papers that reported segmentation results on more than 1 dataset, we list all of them and list the performance on only one dataset, formatting that particular dataset in bold. Since ISIC 2017 is the most popular dataset (Fig. 3), wherever reported, we note the performance (Jaccard index) on ISIC 2017. For papers that do not report the Jaccard index and instead report the Dice score, we compute the former based on the latter and report this computed score denoted by an asterisk. Cross-dataset evaluation (CDE) refers to when a paper trained model(s) on one dataset but evaluated on another.

3.1.1. Single network models

The approaches in this section employ a single DL model, usually an FCN, following an *encoder–decoder* structure, where the encoder extracts increasingly abstract features, and the decoder outputs the segmentation mask. In this section, we discuss these architectural choices for designing deep models for skin lesion segmentation.

Earlier DL-based skin lesion segmentation works adopted either FCN (Long et al., 2015) or U-Net (Ronneberger et al., 2015). FCN originally comprised a backbone of VGG16 (Simonyan and Zisserman, 2014) 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 (Ronneberger et al., 2015), originally proposed for segmenting electron microscopy images, was rapidly adopted in the medical image segmentation literature. As its name suggests, it is a U-shaped model, with an encoder stacking convolutional layers that double 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 producing detailed segmentation boundaries.

U-Net frequently appears in the skin lesion segmentation literature both in its original form (Codella et al., 2017; Pollastri et al., 2020; Ramani and Ranjani, 2019) and modified forms (Tang et al., 2019a; Alom et al., 2019; Hasan et al., 2020), discussed below. Some works introduce their own models (Yuan et al., 2017; Al-Masni et al., 2018).

3.1.1.1. 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, facilitating 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 (He et al., 2016) alleviates gradient degradation in very deep networks. It provides a direct path for the gradient to flow through 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 (Sarker et al., 2018; Baghersalimi et al., 2019; Yu et al., 2017a) or both encoder and decoder (He et al., 2017; Venkatesh et al., 2018; Li et al., 2018a; Tu et al., 2019; Zhang et al., 2019a; He et al., 2018; Xue et al., 2018). Residual connections have also appeared in recurrent units (Alom et al., 2019, 2020), dense blocks (Song et al., 2019), chained pooling (He et al., 2017; Li et al., 2018a; He et al., 2018), and 1-D factorized convolutions (Singh et al., 2019).

Skip connections appear in encoder–decoder architectures, connecting high-resolution features from the encoder's contracting path to the semantic features on the decoder's expanding path (Ronneberger et al., 2015). These connections help preserve localization, especially near region boundaries, and combine multi-scale features, resulting in sharper boundaries in the predicted segmentation. Skip connections are very popular in skin lesion segmentation because they are effective and easy to implement (Zhang et al., 2019a; Baghersalimi et al., 2019; Song et al., 2019; Wei et al., 2019; Venkatesh et al., 2018; Azad et al., 2019; He et al., 2017; Alom et al., 2019; Sarker et al., 2018; Zeng and Zheng, 2018; Li et al., 2018a; Tu et al., 2019; Yu et al., 2017a; Singh et al., 2019; He et al., 2018; Xue et al., 2018; Alom et al., 2020; Vesal et al., 2018b; Liu et al., 2019b).

Dense connections expand the convolutional layers by connecting each layer to all its subsequent layers, concatenating their features (Huang

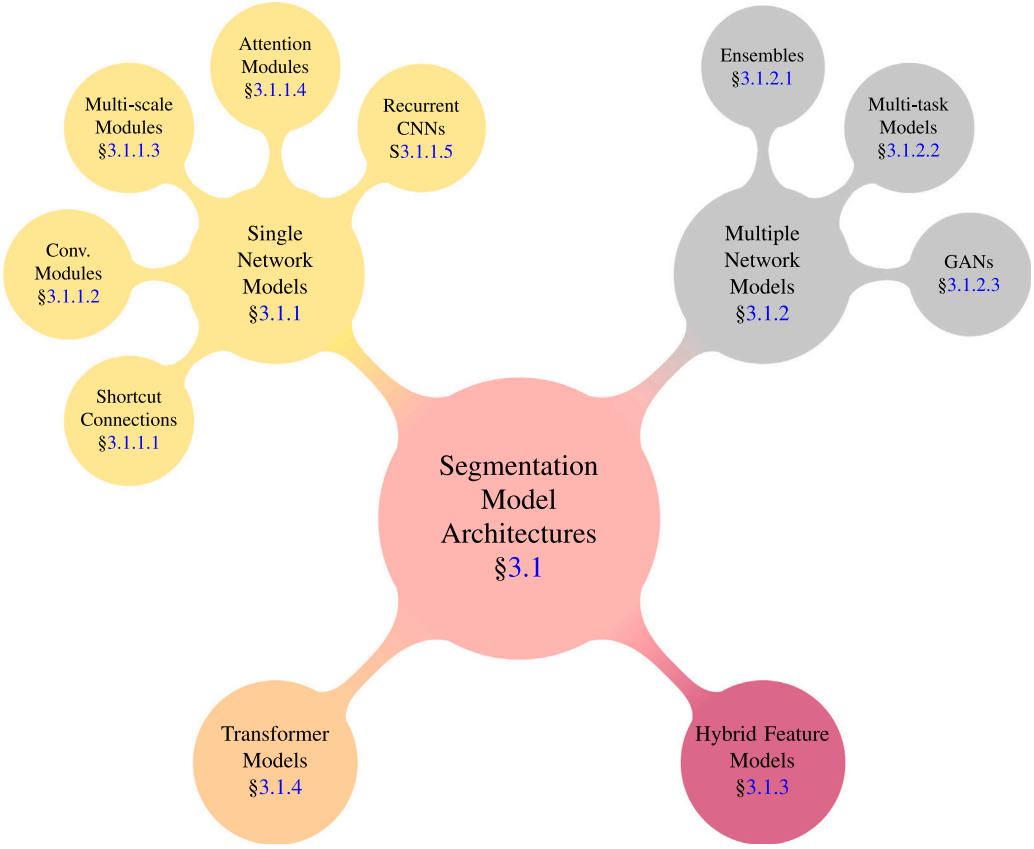


Fig. 6. Taxonomy of DL-based skin lesion segmentation model architectures.

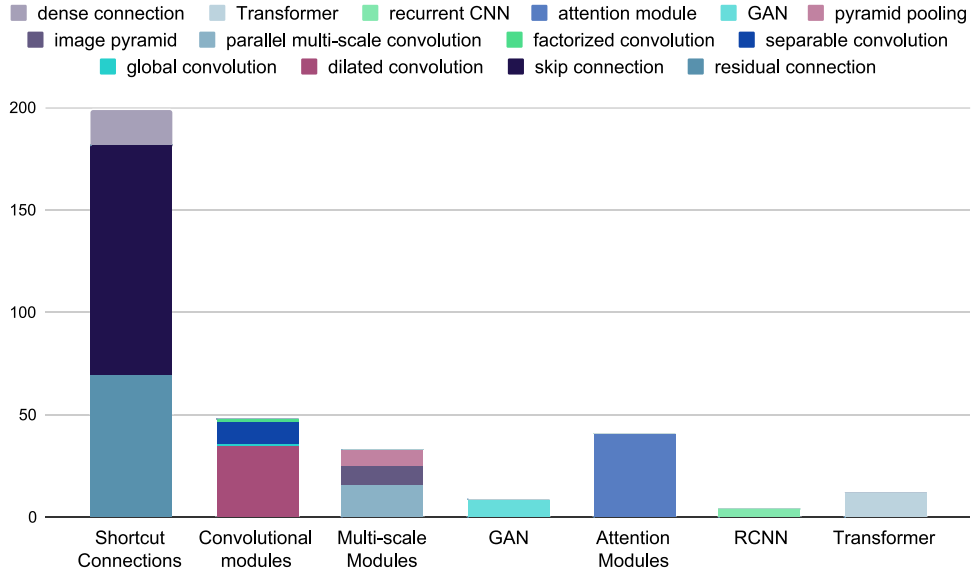


Fig. 7. The frequency of utilization of different architectural modules in the surveyed studies. Shortcut connections, particularly, skip connections (112 papers) and residual connections (70 papers) are the two most frequent components in DL-based skin lesion segmentation models. Attention mechanisms learn dependencies between elements in sequences, either spatially or channel-wise, and are therefore used by several encoder–decoder-style segmentation models (41 papers). Dilated convolutions help expand the receptive field of CNN-models without any additional parameters, which is why they are the most popular variant of convolution in the surveyed studies (35 papers). Finally, papers using Transformers (12 papers) started appearing from 2021 onwards and are on the rise.

et al., 2017). Iterative reuse of features in dense connections maximizes information flow forward and backward. Similar to deep supervision (Section 3.2.5), the gradient is propagated backwards directly through all previous layers. Several works (Zeng and Zheng, 2018; Song et al., 2019; Li et al., 2021c; Tu et al., 2019; Vesal et al., 2018b) integrated dense blocks in both the encoder and the decoder. Baghersalimi et al. (2019), Hasan et al. (2020) and Wei et al. (2019) used multiple dense blocks iteratively in only the encoder, while Li et al. (2018a) proposed dense deconvolutional blocks to reuse features from the previous layers. Azad et al. (2019) encoded densely connected convolutions into the bottleneck of their encoder–decoder to obtain better features.

3.1.1.2. Convolutional modules. As mentioned earlier, 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 convolution 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 have been adopted in deep convolutional neural networks (DCNNs) to increase the receptive field and 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 (Yu and Koltun, 2016). Dilated convolutions are convolutional modules with upsampled filters containing zeros between consecutive filter values. Sarker et al. (2018) and Jiang et al. (2019) 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 (Vesal et al., 2018b) used dilated convolutions at the lower level of the network to enlarge the field-of-view and capture non-local information. Liu et al. (2019b) introduced dilated convolutions to the U-Net architecture, significantly improving the segmentation performance. Furthermore, different versions of the DeepLab architecture (Chen et al., 2017a,b, 2018a), which replace standard convolutions with dilated ones, have been used in skin lesion segmentation (Goyal et al., 2019a,b; Cui et al., 2019; Chen et al., 2018b; Canalini et al., 2019).

Separable convolution: Separable convolution or depth-wise separable convolution (Chollet, 2017) is a spatial convolution operation that convolves each input channel with its corresponding kernel. This 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 maintaining the accuracy. DSNet (Hasan et al., 2020) and separable-Unet (Tang et al., 2019a) 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. (2019b), Cui et al. (2019) and, Canalini et al. (2019) incorporated depth-wise separable convolutions in conjunction with dilated convolution to improve the speed and 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, by keeping high-resolution feature maps, segmentation models become more suitable for localization and, in contrast, less suitable 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, such kernels have a large number of parameters, which renders them computationally expensive. To tackle this, global convolutional network (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 (Peng et al., 2017b). SeGAN (Xue et al., 2018) employed GCN modules with large kernels 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 (Wang et al., 2017) are designed to reduce the number of convolution filter parameters as well as the computation time through kernel decomposition when a high-dimensional kernel is substituted with a sequence of lower-dimensional convolutions. Additionally, by adding non-linearity between the composited kernels, the network's capacity may improve. FCA-Net (Singh et al., 2019) and MobileGAN (Sarker et al., 2019) 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.

3.1.1.3. Multi-scale modules. In FCNs, taking semantic context into account when assigning per-pixel labels leads to a more accurate prediction (Long et al., 2015). Exploiting multi-scale contextual information, effectively combining them as well as encoding them in deep semantic segmentation have been widely explored.

Image Pyramid: RefineNet (He et al., 2017) and its extension (He et al., 2018), MSFCDN (Zeng and Zheng, 2018), FCA-Net (Singh et al., 2019), and Abraham and Khan (2019) fed a pyramid of multi-resolution skin lesion images as input to their deep segmentation network to extract multi-scale discriminative features. RefineNet (He et al., 2017, 2018), Factorized channel attention network (FCA-Net Singh et al., 2019) and Abraham and Khan (2019) 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. Multi-scale fully convolutional DenseNets (MSFCDN Zeng and Zheng, 2018) gradually integrated multi-scale features extracted from the image pyramid into the encoder's down-sampling path. Also, Jafari et al. (2016, 2017) 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 at 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 (Zhang et al., 2019a; Wang et al., 2019a; Jahanifar et al., 2018) or multiple dilated convolutions with different dilation rates (Goyal et al., 2019a,b; Cui et al., 2019; Chen et al., 2018b; Canalini et al., 2019) can be adopted in parallel paths to extract multi-scale contextual features from images. DSM (Zhang et al., 2019a) integrated multi-scale convolutional blocks into the skip connections of an encoder–decoder structure to handle different lesion sizes. Wang et al. (2019a) utilized multi-scale convolutional branches in the bottleneck of an encoder–decoder architecture, followed by attention modules to selectively aggregate the extracted multi-scale features.

Pyramid pooling: Another way of incorporating multi-scale information into deep segmentation models is to integrate a pyramid pooling (PP) module in the network architecture (Zhao et al., 2017). PP fuses a hierarchy of features extracted from different sub-regions by adopting parallel pooling kernels of various sizes, followed by up-sampling and concatenation to create the final feature maps. Sarker et al. (2018) and Jahanifar et al. (2018) utilized PP in the decoder to benefit from coarse-to-fine features extracted by different receptive fields from skin lesion images.

Dilated convolutions and skip connections are two other types of multi-scale information extraction techniques, which are explained in Sections 3.1.1.2 and 3.1.1.1, respectively.

3.1.1.4. Attention modules. An explicit way to exploit contextual dependencies in the pixel-wise labeling task is the self-attention mechanism (Hu et al., 2018; Fu et al., 2019). 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. (2019a) and Sarker et al. (2019) leveraged both spatial and channel attention modules to recalibrate the feature maps by examining the feature similarity between pairs of positions or channels and updating each feature value by a weighted sum of all other features. Singh et al. (2019) 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 (Oktay et al.,

2018), multiple works (Abraham and Khan, 2019; Song et al., 2019; Wei et al., 2019) integrated a spatial attention gate in an encoder–decoder architecture to combine coarse semantic feature maps and fine localization feature maps. Kaul et al. (2019) proposed FocusNet which utilizes squeeze-and-excitation blocks into a hybrid encoder–decoder architecture. Squeeze-and-excitation blocks model the channel-wise interdependencies to re-weight feature maps and improve their representation power. Experimental results demonstrate that attention modules help the network focus on the lesions and suppress irrelevant feature responses in the background.

3.1.1.5. Recurrent convolutional neural networks. Recurrent convolutional neural networks (RCNN) integrate recurrent connections into convolutional layers by evolving the recurrent input over time (Pinheiro and Collobert, 2014). 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. (2017) utilized recurrent layers in the decoder to capture spatial dependencies between deep-encoded features and recover segmentation maps at the original resolution. ∇^N-Net (Alom et al., 2020), RU-Net, and R2U-Net (Alom et al., 2019) 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. (2019) 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 convolutional LSTM modules (ConvLSTMs) which process the feature maps into two directions of backward and forward paths and concatenate their outputs to obtain the final output. Modifications to the traditional pooling layers were also proposed, using a dense pooling strategy (Nasr-Esfahani et al., 2019).

3.1.2. Multiple network models

Motivations for models comprising more than one DL sub-model are diverse, ranging from alleviating training noise and exploiting a diversity of features learned by different models to exploring synergies between multi-task learners. After examining the literature (Fig. 6), 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 models, in a separate category.

3.1.2.1. 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 (Sagi and Rokach, 2018). Ensemble performance is contingent on the quality and diversity of the component models, which can be combined at the feature level (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 (Tang et al., 2019a), which explores various learning-rate decay schemes, and builds a single model by averaging the weights learned at different epochs to bypass poor local minima during training. Since the weights correspond to features learned by the convolution filters, this approach can be interpreted as feature fusion.

Most works employ a single DL architecture with multiple training routines, varying configurations more or less during training (Canalini et al., 2019). The changes between component models may involve network hyperparameters: number of filters per block and their size (Codella et al., 2017); optimization and regularization hyperparameters: learning rate, weight decay (Tan et al., 2019b); the training set: multiple splits of a training set (Yuan et al., 2017; Yuan and Lo, 2019), separate models per class (Bi et al., 2019b); preprocessing: different color spaces (Pollastri et al., 2020); different pretraining strategies to initialize feature extractors (Canalini et al., 2019); or different ways to initialize the network parameters (Cui et al., 2019). Test-time augmentation may also be seen as a form of inference-time ensembling (Chen et al., 2018b; Liu et al., 2019b; Jahanifar et al., 2018) that combines the outputs of multiple augmented images to generate a more reliable prediction.

Bi et al. (2019b) trained a separate DL model for each class, as well as a separate classification model. For inference, the classification model output is used to weight the outputs of the category-specific segmentation networks. In contrast, Soudani and Barhoumi (2019) trained a 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 on a very small test set (10% of ISIC 2017 test set).

Several works also ensemble different model architectures for skin lesion segmentation. Goyal et al. (2019b) investigate multiple fusion approaches to avoid severe errors from individual models, comparing the average-, maximum- and minimum-pooling of their outputs. A common assumption is that the component models of the ensemble are trained independently, but Bi et al. (2017b) 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 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.

3.1.2.2. Multi-task models. Multi-task models jointly address more than one goal, in the hope that synergies among the tasks will improve overall performance (Zhang and Yang, 2022). This can be particularly helpful in medical image analysis, wherein aggregating tasks may alleviate the issue of insufficient data or annotations. For skin lesions, a few multi-task models exploiting segmentation and classification have been proposed (Chen et al., 2018b; Li and Shen, 2018; Yang et al., 2018; Xie et al., 2020b; Jin et al., 2021).

The synergy between tasks may appear when their models share common relevant features. Li and Shen (2018) assume that all features are shareable between the tasks, and train a single fully convolutional residual network to assign class probabilities at the pixel level. They aggregate the class probability maps to estimate both lesion region and class by weighted averaging of probabilities for different classes inside the lesion area. Yang et al. (2018) learn an end-to-end model formed by a shared convolutional feature extractor followed by three task-specific branches (one to segment skin lesions, one to classify them as melanoma versus non-melanoma, and one to classify them as seborrheic keratosis versus non-seborrheic keratosis.) Similarly, Chen et al. (2018b) add a common feature extractor and separate task heads, and introduce a learnable gate function that controls the flow of information between the tasks to model the latent relationship between two tasks.

Instead of using a single architecture for classification and segmentation, Xie et al. (2020b) and Jin et al. (2021) use three CNNs in sequence to perform a coarse segmentation, followed by classification and, finally, a fine segmentation. Instead of shared features, these works exploit sequential guidance, in which the output of each task improves the learning of the next. While Xie et al. (2020b) feed the output of each network to the next, assuming that the classification network is a diagnostic category and a class activation map (Zhou et al., 2016), Jin et al. (2021) introduce feature entanglement modules, which aggregate features learned by different networks.

All multi-task models discussed so far have results suggesting complementarity between classification and segmentation, but there is no clear advantage among these models. The segmentation of dermoscopic features (e.g., networks, globules, regression areas) combined with the other tasks is a promising avenue of research, 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 (Yu et al., 2017a; Codella et al., 2017; Gonzalez-Diaz, 2018; Al-Masni et al., 2020), since without mutual influence (sharing of losses or features) or feedback between the two tasks, there is no opportunity for synergy.

Vesal et al. (2018a) stressed the importance of object localization as an ancillary task for lesion delineation, in particular deploying Faster-RCNN (Ren et al., 2015) to regress a bounding box to crop the lesions before training a SkinNet segmentation model. While this two-stage approach considerably improves the results, it is computationally expensive (a fast non-DL-based bounding box detection algorithm was proposed earlier by Celebi et al. (2009a)). Goyal et al. (2019a) 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.

3.1.2.3. Generative adversarial models. We discussed GANs for synthesizing new samples, their main use in skin lesion analysis, in Section 2.2. In this section, we are interested in GANs not for generating additional 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 class labels and ground-truth segmentation masks.

Peng et al. (2019), Tu et al. (2019), Lei et al. (2020), and Izadi et al. (2018) 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. (2019) use the same scheme, with a dual discriminator. Lei et al. (2020) also employ a second discriminator that takes as input only segmentations (unpaired from input images).

Since 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. (2019) and Tu et al. (2019) address this by pre-multiplying both the generated and real segmentations with the (normalized) input images before feeding them to the discriminator.

We discuss adversarial loss functions further in Section 3.2.8.

3.1.3. 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 combine the best of both worlds, with strategies ranging from employing pre- or postprocessing to enforce prior knowledge to adding hand-crafted features Providing the model with prior knowledge about the expected shape of skin lesions—which is missing from CNNs—may improve the performance. Mirikharaji and Hamarneh (2018) encode shape information into an additional regularization loss, which penalizes segmentation maps that deviate from a star-shaped prior (Section 3.2.6).

Conditional random fields (CRFs) use pixel-level color information models to refine the segmentation masks output by the CNN. While both Tschandl et al. (2019) and Adegun and Viriri (2020b) consider a single CNN, Qiu et al. (2020) combine the outputs of multiple CNNs into a single mask, before feeding it together with the input image to the CRFs. Ünver and Ayan (2019) use GrabCut (Rother et al., 2004) to obtain the segmentation mask given the dermoscopy image and a region proposal obtained by the YOLO (Redmon et al., 2016) network. These 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 features with CNNs follow two distinct approaches. The first consists of pre-filtering the input images to increase the contrast between the lesion and the surrounding skin. Techniques explored include local binary patterns (LBPs) (Ross-Howe and Tizhoosh, 2018; Jayapriya and Jacob, 2020), wavelets (Ross-Howe and Tizhoosh, 2018), Laplacian pyramids (Pour and Seker, 2020), and Laplacian filtering (Saba et al., 2019). The second approach consists of predicting an additional segmentation mask to combine with the one generated by the CNN. Zhang et al. (2019b), for example, use LBPs to consider the textural patterns of skin lesions and guide the networks towards more refined segmentations. Bozorgtabar et al. (2017b) 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. The score mask is then 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 so far indicate that this may be a promising research direction.

3.1.4. Transformer models

Initially proposed for natural language processing (Vaswani et al., 2017), Transformers have proliferated in the last couple of years in other areas, including computer vision applications, especially with improvements made over the years for optimizing the computational cost of self-attention (Parmar et al., 2018; Hu et al., 2019; Ramachandran et al., 2019; Cordonnier et al., 2019; Zhao et al., 2020; Dosovitskiy et al., 2020), and have consequently also been adapted for semantic segmentation tasks (Ranftl et al., 2021; Strudel et al., 2021; Zheng et al., 2021). For medical image segmentation, TransUNet (Chen et al., 2021) was one of the first works to use Transformers along with CNNs in the encoder of a U-Net-like encoder–decoder architecture, and Gulzar and Khan (2022) showed that TransUNet outperforms several CNN-only models for skin lesion segmentation. To reduce the computational complexity involved with high-resolution medical images, Cao et al. (2021) proposed the Swin-Unet architecture that uses self-attention within shifted windows (Liu et al., 2021b). For a comprehensive review of the literature of Transformers in general medical image analysis, we refer the interested readers to the surveys by He et al. (2022) and Shamshad et al. (2022).

Zhang et al. (2021b) propose TransFuse which parallelly computes features from CNN and Transformer modules, with the former capturing low-level spatial information and the latter responsible for modeling global context, and these features are then combined using a self-attention-based fusion module. Evaluation on the ISIC 2017 dataset shows superior segmentation performance and faster convergence. The multi-compound Transformer (Ji et al., 2021) leverages Transformer-based self-attention and cross-attention modules between the encoder and the decoder components of U-Net to learn rich features from multi-scale CNN features. Wang et al. (2021a) incorporate boundary-wise prior knowledge in segmentation models using a boundary-aware Transformer (BAT) to deal with the ambiguous boundaries in skin lesion images. More recently, Wu et al. (2022a) introduce a feature-adaptive Transformer network (FAT-Net) that comprised of a dual CNN-Transformer encoder, a light-weight trainable feature-adaptation module, and a memory-efficient decoder using a squeeze-and-excitation module. The resulting segmentation model is more accurate at segmenting skin lesions while also being faster (fewer parameters and computation) than several CNN-only models.

3.2. Loss functions

A segmentation model f may be formalized as a function $\hat{y} = f_\theta(x)$, which maps an input image x to an estimated segmentation map $\hat{y}$ parameterized by a (large) set of parameters θ. For skin lesions, $\hat{y}$ is a binary mask separating the lesion from the surrounding skin. Given a training set of images x_i and their corresponding ground-truth masks y_i $\{(x_i, y_i); i = 1, \ldots, N\}$, training a segmentation model consists of finding the model parameters θ that maximize the likelihood of observing those data:

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

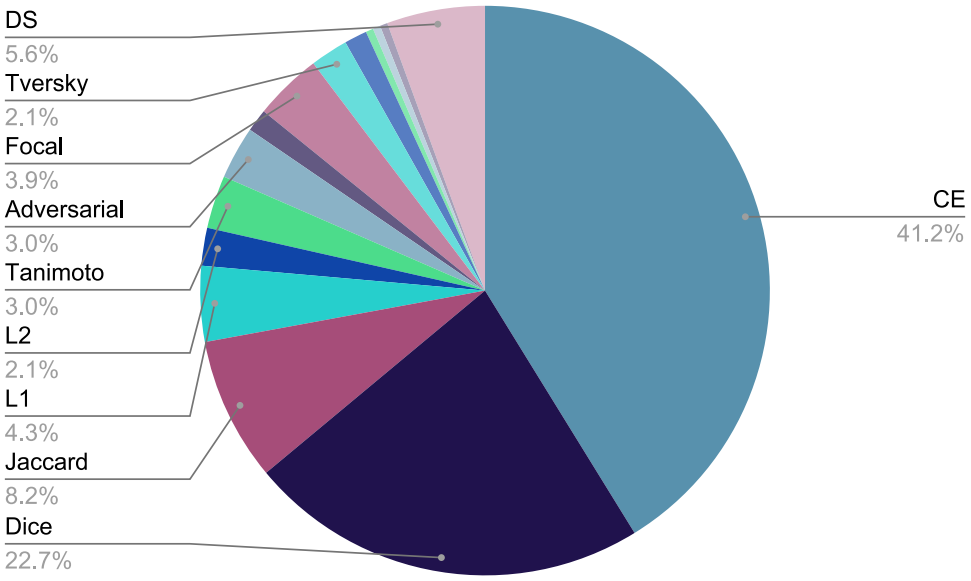


Fig. 8. The distribution of loss functions used by the surveyed works in DL-based skin lesion segmentation. Cross-entropy loss is the most popular loss function (96 papers), followed by Dice (53 papers) and Jaccard (19 papers) losses. Of the 177 surveyed papers, 65 use a combination of losses, with CE + Dice (27 papers) and CE + Jaccard (11 papers) being the most popular combinations.

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

$$\theta^* = \arg\min_{\theta} \sum_{i=1}^{N} \mathcal{L}(\hat{y}_i|y_i) = \arg\min_{\theta} \sum_{i=1}^{N} \mathcal{L}(f_\theta(x_i)|y_i). \tag{2}$$

The choice of the loss function is thus critical, as it encodes not only the main optimization objective, but also much of the prior information needed to guide the learning and constrain the search space. As can been in Table 3, many skin lesion segmentation models employ a combination of losses to enhance generalization (see Fig. 8).

3.2.1. Losses based on p-norms

Losses based on p-norms are the simplest ones, and comprise the mean squared error (MSE) (for $p = 2$) and the mean absolute error (MAE) (for $p = 1$).

$$\text{MSE}(X, Y; \theta) = -\sum_{i=1}^{N} \|y_i - \hat{y}_i\|_2, \tag{3}$$

$$\text{MAE}(X, Y; \theta) = -\sum_{i=1}^{N} \|y_i - \hat{y}_i\|_1. \tag{4}$$

In GANs, to regularize the segmentations produced by the generator, it is common to utilize hybrid losses containing MSE (ℓ_2 loss) (Peng et al., 2019) or MAE (ℓ_1 loss) (Peng et al., 2019; Tu et al., 2019; Lei et al., 2020). The MSE has also been used as a regularizer to match attention and ground-truth maps (Xie et al., 2020a).

3.2.2. Cross-entropy loss

Semantic segmentation may be viewed as classification at the pixel level, i.e., as assigning a class label to each pixel. From this perspective, minimizing the negative log-likelihoods of pixel-wise predictions (i.e., maximizing their likelihood) may be achieved by minimizing a cross-entropy loss $\mathcal{L}_{ce}$:

$$\begin{aligned}\mathcal{L}_{ce}(X, Y; \theta) &= -\sum_{i=1}^{N} \sum_{p \in \Omega_i} y_{ip} \log \hat{y}_{ip} + (1 - y_{ip}) \log(1 - \hat{y}_{ip}), \quad \hat{y}_{ip} \\ &= P(y_{ip} = 1 | X(i); \theta),\end{aligned} \tag{5}$$

where Ω_i is the set of all image i pixels, P is the probability, x_{ip} is pth image pixel in ith image and, $y_{ip} \in \{0, 1\}$ and $\hat{y}_{ip} \in [0, 1]$ are respectively the true and the predicted labels of x_{ip}. The cross-entropy loss appears in the majority of deep skin lesion segmentation works, e.g., Song et al. (2019), Singh et al. (2019), and Zhang et al. (2019a).

Since the gradient of the cross-entropy loss function is inversely proportional to the predicted probabilities, hard-to-predict samples are weighted more in the parameter update equations, leading to faster convergence. A variant, the weighted cross-entropy loss, penalizes pixels and class labels differently. Nasr-Esfahani et al. (2019) used pixel weights inversely proportional to their distance to lesion boundaries to enforce sharper boundaries. Class weighting may also mitigate the class imbalance, which, left uncorrected, tends to bias models towards the background class, since lesions tend to occupy a relatively small portion of images. Chen et al. (2018b), Goyal et al. (2019a), and Wang et al. (2019b) apply such a correction, using class weights inversely proportional to the class pixel frequency. Mirikharaji et al. (2019) weighted the pixels according to annotation noise estimated using a set of cleanly annotated data. All the aforementioned losses treat pixels independently without enforcing spatial coherence, which motivates their combination with other consistency-seeking losses.

3.2.3. Dice and Jaccard loss

The Dice score and the Jaccard index are two popular metrics for segmentation evaluation (Section 4.3), measuring the overlap between predicted segmentation and ground-truth. Models may employ differentiable approximations of these metrics, known as soft Dice (He et al., 2017; Kaul et al., 2019; He et al., 2018; Wang et al., 2019a) and soft Jaccard (Venkatesh et al., 2018; Hasan et al., 2020; Sarker et al., 2019) to optimize an objective directly related to the evaluation metric.

For two classes, these losses are defined as follows:

$$\mathcal{L}_{dice}(X, Y; \theta) = 1 - \frac{1}{N} \sum_{i=1}^{N} \frac{2 \sum_{p \in \Omega} y_{ip} \hat{y}_{ip}}{\sum_{p \in \Omega} y_{ip} + \hat{y}_{ip}}, \tag{6}$$

$$\mathcal{L}_{jacc}(X, Y; \theta) = 1 - \frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{p \in \Omega} y_{ip} \hat{y}_{ip}}{\sum_{p \in \Omega} y_{ip} + \hat{y}_{ip} - y_{ip} \hat{y}_{ip}}. \tag{7}$$

Different variations of overlap-based loss functions address the class imbalance problem in medical image segmentation tasks. The Tanimoto distance loss, $\mathcal{L}_{td}$ is a modified Jaccard loss optimized in some models (Canalini et al., 2019; Baghersalimi et al., 2019; Yuan et al., 2017):

$$\mathcal{L}_{td}(X, Y; \theta) = 1 - \frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{p \in \Omega} y_{ip} \hat{y}_{ip}}{\sum_{p \in \Omega} y_{ip}^2 + \hat{y}_{ip}^2 - y_{ip} \hat{y}_{ip}}, \tag{8}$$

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

The Tversky loss (Abraham and Khan, 2019), inspired by the Tversky index, is another Jaccard variant that penalizes false positives and false negatives differently to address the class imbalance problem:

$$\mathcal{L}_{tv}(X,Y;\theta) = 1 - \frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{p\in\Omega} y_{ip}\hat{y}_{ip}}{\sum_{p\in\Omega} y_{ip}\hat{y}_{ip} + \alpha y_{ip}(1-\hat{y}_{ip}) + \beta(1-y_{ip})\hat{y}_{ip}}, \quad (9)$$

where α and β tune the contributions of false negatives and false positives with $\alpha + \beta = 1$.

Abraham and Khan (2019) combined the Tvserky and focal losses (Lin et al., 2017), the latter encouraging the algorithm to focus on the hard-to-predict pixels:

$$\mathcal{L}_{ftv} = \mathcal{L}_{tv}^{\frac{1}{\gamma}}, \quad (10)$$

where γ controls the relative importance of the hard-to-predict samples.

3.2.4. 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 (Abhishek and Hamarneh, 2021). In contrast to the overlap-based losses discussed in Section 3.2.3, MCC considers misclassifying the background pixels by penalizing false negative labels, making it more effective in the presence of skewed class distributions. MCC loss is defined as:

$$\mathcal{L}_{MCC}(X,Y;\theta) = 1 - \frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{p\in\Omega}\hat{y}_{ip}y_{ip} - \frac{\sum_{p\in\Omega}\hat{y}_{ip}\sum_{p\in\Omega}y_{ip}}{M_i}}{f(\hat{y}_i y_i)}, \quad (11)$$

$$f(\hat{y}_i, y_i) = \sqrt{\sum_{p\in\Omega}\hat{y}_{ip}\sum_{p\in\Omega}y_{ip} - \frac{\sum_{p\in\Omega}\hat{y}_{ip}(\sum_{p\in\Omega}y_{ip})^2}{M_i} - \frac{(\sum_{p\in\Omega}\hat{y}_{ip})^2\sum_{p\in\Omega}y_{ip}}{M_i} + (\frac{\sum_{p\in\Omega}\hat{y}_{ip}\sum_{p\in\Omega}y_{ip}}{M_i})^2}, \quad (12)$$

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

3.2.5. Deep supervision loss

In DL models, the loss may apply not only to the final decision layer, but also to the intermediate hidden layers. The 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 and improves segmentation performance by constraining the feature space. Deep supervision loss appears in several skin lesion segmentation works (He et al., 2017; Zeng and Zheng, 2018; Li et al., 2018a,b; He et al., 2018; Zhang et al., 2019a; Tang et al., 2019b), 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:

$$\mathcal{L}_{ds}(X,Y;\theta) = \sum_{l=1}^{m}\gamma_l\mathcal{L}_l(X,Y;\theta), \quad (13)$$

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

3.2.6. Star-shape loss

In contrast to pixel-wise losses which act on pixels independently and cannot enforce spatial constraints, the star-shape loss (Mirikharaji and Hamarneh, 2018) aims to capture class label dependencies and preserve the target object structure in the predicted segmentation masks. Based upon prior knowledge about the shape of skin lesions, the star-shape loss, $\mathcal{L}_{ssh}$ penalizes discontinuous decisions in the estimated output as follows:

$$\mathcal{L}_{ssh}(X,Y;\theta) = \sum_{i=1}^{N}\sum_{p\in\Omega}\sum_{q\in\ell_{pc}}\mathbb{1}_{y_{ip}=y_{iq}}\times|y_{ip}-\hat{y}_{ip}|\times|\hat{y}_{ip}-\hat{y}_{iq}|, \quad (14)$$

where c is the lesion center, ℓ_{pc} is the line segment connecting pixels p and c and, q is any pixel lying on ℓ_{pc}. This loss encourages all pixels lying between p and q on ℓ_{pc} to be assigned the same estimator whenever p and q have the same ground-truth label. The result is a radial spatial coherence from the lesion center.

3.2.7. End-point error loss

Many authors consider the lesion boundary the most challenging region to segment. The end-point error loss (Sarker et al., 2018; Singh et al., 2019) underscores borders by using the first derivative of the segmentation masks instead of their raw values:

$$\mathcal{L}_{epe}(X,Y;\theta) = \sum_{i=1}^{N}\sum_{p\in\Omega}\sqrt{(\hat{y}^0_{ip}-y^0_{ip})^2 + (\hat{y}^1_{ip}-y^1_{ip})^2}, \quad (15)$$

where $\hat{y}^0_{ip}$ and $\hat{y}^1_{ip}$ are the directional first derivatives of the estimated segmentation map in the x and y spatial directions, respectively and, similarly, y^0_{ip} and y^1_{ip} for the ground-truth derivatives. Thus, this loss function encourages the magnitude and orientation of edges of estimation and ground-truth to match, thereby mitigating vague boundaries in skin lesion segmentation.

3.2.8. Adversarial loss

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

$$\mathcal{L}_{adv}(X,Y;\theta,\theta_a) = \mathcal{L}_s(X,Y;\theta) - \lambda[\mathcal{L}_{ce}(Y,1;\theta_a) + \mathcal{L}_{ce}(\hat{Y},0;\theta,\theta_a)], \quad (16)$$

where θ_a are the adversarial model parameters. The adversarial loss employs a binary cross-entropy loss to encourage the segmentation model to produce indistinguishable prediction maps from ground-truth maps. The adversarial objective (Eq. (16)) is optimized in a mini-max game by simultaneously minimizing it with respect to θ and maximizing it with respect to θ_a.

Pixel-wise losses, such as cross-entropy (Izadi et al., 2018; Singh et al., 2019; Jiang et al., 2019), soft Jaccard (Sarker et al., 2019; Tu et al., 2019; Wei et al., 2019), end-point error (Tu et al., 2019; Singh et al., 2019), MSE (Peng et al., 2019) and MAE (Sarker et al., 2019; Singh et al., 2019; Jiang et al., 2019) losses have all been incorporated in adversarial learning of skin lesion segmentation. In addition, Xue et al. (2018) and Tu et al. (2019) presented a multi-scale adversarial term to match a hierarchy of local and global contextual features in the predicted maps and ground-truths. In particular, they minimize the MAE of multi-scale features extracted from different layers of the adversarial model.

3.2.9. Rank loss

Assuming that hard-to-predict pixels lead to larger prediction errors while training the model, rank loss (Xie et al., 2020b) is proposed to encourage learning more discriminative information for harder pixels. The image pixels are ranked based on their prediction errors, and the top K pixels with the largest prediction errors from the lesion or background areas are selected. Let $\hat{y}^0_{ij}$ and $\hat{y}^1_{il}$ are respectively the selected jth hard-to-predict pixel of background and lth hard-to-predict pixel of lesion in the image i, we have:

$$\mathcal{L}_{rank}(X,Y;\theta) = \sum_{i=1}^{N}\sum_{j=1}^{K}\sum_{l=1}^{K}\max\{0,\hat{y}^0_{ij} - \hat{y}^1_{il} + margin\}, \quad (17)$$

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

Similar to rank loss, narrowband suppression loss (Deng et al., 2020) also adds a constraint between hard-to-predict 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 the largest prediction errors.

4. Evaluation

Evaluation is one of the main challenges for any image segmentation task, skin lesions included (Celebi et al., 2015b). Segmentation evaluation may be subjective or objective (Zhang et al., 2008), the former involving the visual assessment of the results by a panel of human experts, and the latter involving the comparison of the results with ground-truth segmentations using quantitative evaluation metrics.

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 limited settings. In objective assessment, experts are consulted once, to provide the ground-truth segmentations, and that knowledge can then be reused indefinitely. However, due to intra- and inter-annotator variations, it raises the question of whether any individual ground-truth segmentation reflects the ideal "true" segmentation, an issue we address in Section 4.2. It also raises the issue of choosing one or more evaluation metrics (Section 4.3).

4.1. Segmentation annotation

Obtaining ground-truth segmentations is paramount for the objective evaluation of segmentation algorithms. For synthetically generated images (Section 2.2), ground-truth segmentations 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 with their segmentation masks.

For images obtained from real patients, however, human experts have to provide the ground-truth segmentations. Various workflows have been proposed to reconcile 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 (Bogo et al., 2015). 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 (Codella et al., 2019), with an initial border generated by a segmentation algorithm, which is then refined by the expert using an annotation software, by adjusting the parameters of the segmentation algorithm manually. This method is fast and easy to correct, but there might be a learning curve, and its accuracy depends 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 on a computer. The skin lesion image file may be opened either in a raster graphics editor (e.g., GNU Image Manipulation Program (GIMP) or Adobe Photoshop), or in a dedicated annotation software (Ferreira et al., 2012), 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 Codella et al., 2019) or smooth curve segments (e.g., cubic B-splines Celebi et al., 2007a). This method provides a good compromise, being easy to implement, fast, and accurate to perform, after an acceptable learning period for the annotator.

4.2. Inter-annotator agreement

Formally, dataset ground-truths can be viewed as samples of an estimator of the true label, which can never be directly observed (Smyth et al., 1995). This problem 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 (Zijdenbos et al., 1994; Chalana and Kim, 1997; Guillod et al., 2002; Grau et al., 2004; Bogo et al., 2015; Lampert et al., 2016), the latter appearing as inter-annotator (different experts) and intra-annotator (same expert at different times) variability.

In the largest study of its kind to date, Fortina et al. (2012) 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 (Section 4.3) between annotators was $\approx 15\%$, and that in 10% of cases, this 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 with more than 90% of any given ground-truth on the same image (Fortina et al., 2012).

Due to the aforementioned variability 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 (Chalana and Kim, 1997). 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 (Celebi et al., 2007a).
- Measuring and reporting the results for each annotator separately (Celebi et al., 2008), which might require non-trivial multivariate analyses if the aim is to rank the algorithms.
- Measuring each automated segmentation against all corresponding ground-truths and reporting the average result (Schaefer et al., 2011).
- Measuring each automated segmentation against an *ensemble ground-truth* formed by combining the corresponding ground-truths pixel-wise using a bitwise OR (Garnavi et al., 2011a; Garnavi and Aldeen, 2011), bitwise AND (Garnavi et al., 2011b), or a majority voting (Iyatomi et al., 2006, 2008; Norton et al., 2012).

The ground-truth ensembling process can be generalized using a *thresholded probability map* (Biancardi et al., 2010). 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 $(n-\epsilon)/2n$, 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 (Smyth et al., 1995; Lampert et al., 2016).

The abovementioned approaches fail to consider the differences of experience or performance of the annotators (Warfield et al., 2004). More elaborate ground-truth fusion alternatives include shape averaging (Rohlfing and Maurer, 2006), border averaging (Chen and Parent, 1989; Chalana and Kim, 1997), binary label fusion algorithms such as STAPLE (Warfield et al., 2004), TESD (Biancardi et al., 2010), and SIMPLE (Langerak et al., 2010), as well as other more recent algorithms (Peng and Li, 2013; Peng et al., 2016, 2017a).

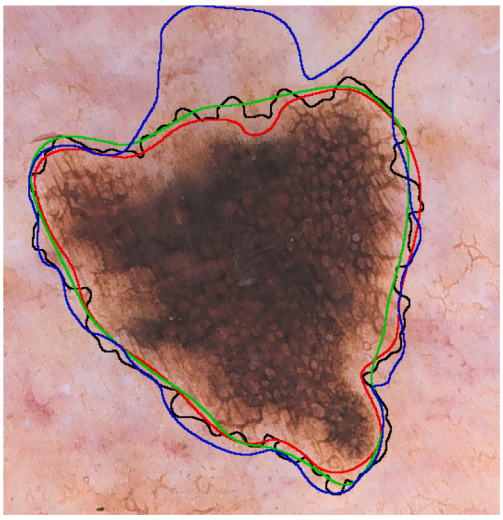


Fig. 9. Sample segmentation results demonstrating inter-annotator disagreements. Note how annotator preferences can affect the manual segmentations, e.g., smooth lesion borders (green), jagged lesion borders (black), oversegmented lesion (blue), etc.
Image source: Figure taken from Celebi et al. (2009c) with permission.

STAPLE (Simultaneous Truth And Performance Level Estimation) has been very influential in medical image segmentation evaluation, 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 only a few annotators or when their performances vary too much (Langerak et al., 2010; Lampert et al., 2016), a situation addressed by SIMPLE (Selective and Iterative Method for Performance Level Estimation) (Langerak et al., 2010) by iteratively discarding poor quality ground-truths.

Instead of attempting to fuse multiple ground-truths into a single one before employing conventional evaluation metrics, the metrics themselves may be modified to take into account annotation variability. Celebi et al. (2009c) proposed the *normalized probabilistic rand index* (NPRI) (Unnikrishnan et al., 2007), a generalization of the *rand index* (Rand, 1971). It penalizes segmentation results more (less) in regions where the ground-truths agree (disagree). Fig. 9 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 (Celebi et al., 2009c). Despite its many desirable qualities, NPRI has a subtle flaw: it is non-monotonic with the fraction of misclassified pixels (Peserico and Silletti, 2010). Consequently, this index might be unsuitable for comparing poor segmentation algorithms.

4.3. Evaluation metrics

We can frame the skin lesion 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. 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
Definitions of true positive, false negative, false positive, and true negative pixels in the context of skin lesion segmentation.

		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 (aka a contingency table Pearson, 1904; Miller and Nicely, 1955) $\mathsf{C} = \left(\begin{smallmatrix} \mathsf{TP} & \mathsf{FN} \\ \mathsf{FP} & \mathsf{TN} \end{smallmatrix}\right)$, where TP, FN, FP, and TN denote the numbers of true positives, false negatives, false positives, and true negatives, respectively. Clearly, we have $N = \mathsf{TP}+\mathsf{FN}+\mathsf{FP}+\mathsf{TN}$, 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 (Baldi et al., 2000; Japkowicz and Shah, 2011; Taha and Hanbury, 2015):

- Sensitivity (SE) and Specificity (SP) (Kahn, 1942; Yerushalmy, 1947; Binney et al., 2021): $\mathsf{SE} = \dfrac{\mathsf{TP}}{\mathsf{TP}+\mathsf{FN}}$ & $\mathsf{SP} = \dfrac{\mathsf{TN}}{\mathsf{TN}+\mathsf{FP}}$
- Precision (PR) and Recall (RE) (Kent et al., 1955): $\mathsf{PR} = \dfrac{\mathsf{TP}}{\mathsf{TP}+\mathsf{FP}}$ & $\mathsf{RE} = \dfrac{\mathsf{TP}}{\mathsf{TP}+\mathsf{FN}}$
- Accuracy (AC) $= \dfrac{\mathsf{TP}+\mathsf{TN}}{\mathsf{TP}+\mathsf{FN}+\mathsf{FP}+\mathsf{TN}}$
- F-measure (F) (van Rijsbergen, 1979) $= \dfrac{2|\mathsf{AS}\cap\mathsf{MS}|}{|\mathsf{AS}|+|\mathsf{MS}|} = \dfrac{2\cdot\mathsf{PR}\cdot\mathsf{RE}}{\mathsf{PR}+\mathsf{RE}} = \dfrac{2\mathsf{TP}}{2\mathsf{TP}+\mathsf{FP}+\mathsf{FN}}$
- G-mean (GM) (Kubat et al., 1998) $= \sqrt{\mathsf{SE}\cdot\mathsf{SP}}$
- Balanced Accuracy (BA) (Chou and Fasman, 1978) $= \dfrac{\mathsf{SE}+\mathsf{SP}}{2}$
- Jaccard index (J) (Jaccard, 1901) $= \dfrac{|\mathsf{AS}\cap\mathsf{MS}|}{|\mathsf{AS}\cup\mathsf{MS}|} = \dfrac{\mathsf{TP}}{\mathsf{TP}+\mathsf{FN}+\mathsf{FP}}$
- Matthews Correlation Coefficient (MCC) (Matthews, 1975) $= \dfrac{\mathsf{TP}\cdot\mathsf{TN}-\mathsf{FP}\cdot\mathsf{FN}}{\sqrt{(\mathsf{TP}+\mathsf{FP})(\mathsf{TP}+\mathsf{FN})(\mathsf{TN}+\mathsf{FP})(\mathsf{TN}+\mathsf{FN})}}$

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 (Hance et al., 1996) defined as follows:

$$\mathsf{XOR} = \frac{|\mathsf{AS}\oplus\mathsf{MS}|}{|\mathsf{MS}|} = \frac{|\,(\mathsf{AS}\cup\mathsf{MS})-(\mathsf{AS}\cap\mathsf{MS})\,|}{|\mathsf{MS}|} = \frac{\mathsf{FP}+\mathsf{FN}}{\mathsf{TP}+\mathsf{FN}}. \tag{18}$$

It is essential to notice that different evaluation measures capture different aspects of a segmentation algorithm's performance on a given dataset, and thus there is no universally applicable evaluation measure (Japkowicz and Shah, 2011). 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, Garnavi et al. (2011a) argue that true positives have the highest importance in the segmentation of skin lesion 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

Table 3
DL models for skin lesion segmentation. Performance measure reported is the Jaccard index computed on the dataset, shown in boldface. 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, J: Jaccard index, CDE : cross-data evaluation. the highlighted dataset and PP: postprocessing, 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. Please see the corresponding sections for more details: Section 3.1 for model architectures, Section 3.2 for loss functions, and Section 4 for model evaluation. An interactive version of this table is available online at https://github.com/sfu-mial/skin-lesion-segmentation-survey.

Ref.	Venue	Data	Arch. modules	Seg. loss	J	CDE	Augmentation	PP	code
Jafari et al. (2016)	Peer-reviewed conference	Der-mQuest	Image pyramid	–	–	✗	–	✓	✗
He et al. (2017)	Peer-reviewed conference	ISIC2016 **ISIC2017**	residual con. skip con. image pyramid	Dice CE DS	75.80%	✗	Rotation	✓	✗
Bozorgtabar et al. (2017b)	Peer-reviewed journal	ISIC2016	–	–	80.60%	✗	Rotation	✗	✗
Ramachandram and Taylor (2017)	Peer-reviewed journal	ISIC2017	–	CE	79.20%	✗	Rotation, flipping color jittering	✗	✗
Yu et al. (2017a)	Peer-reviewed journal	ISIC2016	skip con. residual con.	–	82.90%	✗	Rotation, translation random noise cropping	✗	✓
Bi et al. (2017b)	Peer-reviewed journal	**ISIC2016** PH^2	–	CE	84.64%	✓	Flipping, cropping	✓	✗
Jafari et al. (2017)	Peer-reviewed journal	Der-mQuest	image pyramid	–	–	✗	–	✓	✗
Yuan et al. (2017)	Peer-reviewed journal	**ISIC2016** PH^2	–	Tanimoto	84.7%	✓	Flipping, rotation scaling, shifting contrast norm.	✓	✗
Ramachandram and DeVries (2017)	Non Peer-reviewed technical report	**ISIC2017**	dilated conv.	CE	64.20%	✗	Rotation flipping	✓	✗
Bozorgtabar et al. (2017a)	Peer-reviewed conference	**ISIC2016**	–	CE	82.90%	✗	Rotations	✓	✗
Bi et al. (2017a)	Peer-reviewed conference	**ISIC2016**	parallel m. s.	–	86.36%	✗	Crops, flipping	✓	✗
Attia et al. (2017)	Peer-reviewed conference	**ISIC2016**	recurrent net.	–	93.00%	✗	–	✗	✗
Deng et al. (2017)	Peer-reviewed conference	**ISIC2016**	parallel m. s.	–	84.1%	✗	–	✗	✗
Mishra and Daescu (2017)	Peer-reviewed conference	**ISIC2017**	skip con.	Dice	84.2%	✗	Rotation flipping	✓	✗
Goyal et al. (2017)	Peer-reviewed conference	**ISIC2017**	–	CE Dice	–	✗	–	✗	✗
Vesal et al. (2018a)	Peer-reviewed conference	**ISIC2017** PH^2	dilated conv. dense con. skip con.	Dice	88.00%	✓	–	✗	✗
Venkatesh et al. (2018)	Peer-reviewed conference	ISIC2017	residual con. skip con.	Jaccard	76.40%	✗	Rotation, flipping translation, scaling	✓	✗
Yang et al. (2018)	Peer-reviewed conference	ISIC2017	skip con. parallel m.s. conv.	–	74.10%	✗	Rotation, flipping	✗	✗
Sarker et al. (2018)	Peer-reviewed conference	ISIC2016 **ISIC2017**	skip con. residual con. dilated conv. pyramid pooling	CE EPE	78.20%	✗	Rotation, scaling	✗	✓
Al-Masni et al. (2018)	Peer-reviewed journal	**ISIC2017** PH^2	–	CE	77.10%	✓	Rotation	✗	✗
Li et al. (2018b)	Peer-reviewed conference	ISIC2017	skip con. residual con.	DS	77.23%	✗	Flipping, rotation	✗	✓

(*continued on next page*)

Table 3 (*continued*).

Zeng and Zheng (2018)	Peer-reviewed conference	ISIC2017	dense con. skip con. image pyramid	CE ℓ_2 DS	78.50%	✗	Flipping, rotation	✓	✗
DeVries and Taylor (2018)	Non Peer-reviewed technical report	ISIC2017	skip con.	CE	73.00%	✗	Flipping, rotation	✗	✗
Izadi et al. (2018)	Peer-reviewed conference	DermoFit	skip con.	CE ADV	81.20%	✗	Flipping, rotation elastic deformation	✗	✓
Li et al. (2018a)	Peer-reviewed journal	ISIC2016 **ISIC2017**	skip con. residual con. dense con.	Jaccard DS	76.50%	✗	–	✗	✗
Mirikharaji and Hamarneh (2018)	Peer-reviewed conference	ISIC2017	residual con.	CE Star shape	77.30%	✗	–	✗	✗
Pollastri et al. (2018)	Peer-reviewed conference	ISIC2017	–	Jaccard ℓ_1	78.10%	✗	GAN	✓	✗
Vesal et al. (2018b)	Abstract	ISIC2017	dilated conv. dense con. skip con.	Dice	76.67%	✗	Rotation, flipping, translation, scaling, color shift	✗	✗
Chen et al. (2018b)	Peer-reviewed conference	ISIC2017	residual con. dilated conv. parallel m.s. conv.	WCE	78.70%	✗	Rotation, flipping cropping, zooming Gaussian noise	✓	✗
Jahanifar et al. (2018)	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	✓	✗
Mirikharaji et al. (2018)	Peer-reviewed conference	ISIC2016	skip con.	CE	83.30%	✗	Flipping, rotation	✗	✗
Bi et al. (2018)	Non Peer-reviewed technical report	ISIC2018	residual con.	CE	83.12%	✗	GAN	✗	✗
He et al. (2018)	Peer-reviewed journal	ISIC2016 **ISIC2017**	skip con. residual con. image pyramid	CE Dice DS	76.10%	✗	Rotation	✓	✗
Xue et al. (2018)	Peer-reviewed conference	ISIC2017	skip con. residual con. global conv. GAN	ℓ_1 DS ADV	78.50%	✗	Cropping color jittering	✗	✗
Ebenezer and Rajapakse (2018)	Non Peer-reviewed technical report	**ISIC 2018**	skip con.	Dice	75.6%	✗	Rotation flipping zooming	✓	✓
Goyal et al. (2019b)	Peer-reviewed journal	**ISIC2017** PH^2	dilated conv. parallel m.s. conv. separable conv.	–	79.34%	✓	–	✓	✗
Azad et al. (2019)	Peer-reviewed conference	ISIC2018	skip con. dense con. recurrent CNN	CE	74.00%	✗	–	✗	✓
Alom et al. (2019)	Peer-reviewed journal	ISIC2017	skip con. residual con. recurrent CNN	CE	75.68%	✗	–	✗	✗
Yuan and Lo (2019)	Peer-reviewed journal	ISIC2017	–	Tanimoto	76.50%	✗	Rotation, flipping shifting, scaling random normaliz.	✓	✗
Goyal et al. (2019a)	Peer-reviewed conference	**ISIC2017** PH^2	dilated conv. parallel m.s. conv.	WCE	82.20%	✓	–	✗	✗

(*continued on next page*)

Table 3 (*continued*).

Bi et al. (2019b)	Peer-reviewed journal	ISIC2016 **ISIC2017** PH^2	skip con. residual con.	CE	77.73%	✓	Flipping, cropping	✓	✗
Tschandl et al. (2019)	Peer-reviewed journal	ISIC2017	skip con.	CE Jaccard	76.80%	✗	Flipping, rotation	✓	✗
Li et al. (2021c)	Peer-reviewed journal	ISIC2017	skip con. dense con. semi-supervised ensemble	CE ℓ_1	79.80%	✗	Flipping, rotating scaling	✓	✗
Zhang et al. (2019b)	Peer-reviewed journal	ISIC2016 **ISIC2017**	skip con.	CE	72.94%	✗	–	✗	✗
Baghersalimi et al. (2019)	Peer-reviewed journal	ISIC2016 **ISIC2017** PH^2	skip con. residual con. dense con.	Tanimoto	78.30%	✓	Flipping, cropping	✗	✗
Jiang et al. (2019)	Peer-reviewed conference	ISIC2017	residual con. dilated conv. GAN	ADV ℓ_2	76.90%	✗	Rotation, flipping	✗	✗
Tang et al. (2019b)	Peer-reviewed conference	ISIC2016	skip con.	Tanimoto DS	85.34%	✗	Rotation, flipping	✗	✗
Bi et al. (2019a)	Peer-reviewed conference	ISIC2017	residual con.	CE	77.14%	✗	GAN	✗	✗
Abraham and Khan (2019)	Peer-reviewed conference	ISIC2018	skip con. image pyramid attention	TV Focal	74.80%	✗	–	✗	✓
Cui et al. (2019)	Peer-reviewed conference	ISIC2018	dilated conv. parallel m.s. conv. separable conv.	–	83.00%	✗	–	✗	✗
Song et al. (2019)	Peer-reviewed conference	ISIC2017	skip con. residual con. dense con. attention mod.	CE Jaccard	76.50%	✗	–	✗	✗
Singh et al. (2019)	Peer-reviewed journal	ISIC2016 **ISIC2017** ISIC2018	skip con. residual con. factorized conv. attention mod. GAN	CE ℓ_1 EPE	78.65%	✗	–	✗	✓
Tan et al. (2019b)	Peer-reviewed journal	**ISIC2017** DermoFit PH^2	dilated conv.	Dice	62.29%*	✓	–	✓	✗
Kaul et al. (2019)	Peer-reviewed conference	ISIC2017	skip con. residual con. attention mod.	Dice	75.60%	✗	Channel shift	✗	✗
De Angelo et al. (2019)	Peer-reviewed conference	**ISIC2017** Private	skip con.	CE Dice	76.07%	✗	Flipping, shifting rotation color jittering	✓	✗
Zhang et al. (2019a)	Peer-reviewed journal	**ISIC2017** PH^2	skip con. residual con. parallel m.s. conv.	CE Dice DS	78.50%	✓	Flipping, rotation whitening contrast enhance.	✓	✗
Soudani and Barhoumi (2019)	Peer-reviewed journal	ISIC2017	residual con.	CE	78.60%	✗	Rotation, flipping	✗	✗
Mirikharaji et al. (2019)	Peer-reviewed conference	ISIC2017	skip con.	WCE	68.91%*	✗	–	✗	✗
Nasr-Esfahani et al. (2019)	Peer-reviewed journal	Der-mQuest	dense con.	WCE	85.20%	✗	Rotation, flipping cropping	✗	✗

(*continued on next page*)

Table 3 (*continued*).

Wang et al. (2019a)	Peer-reviewed conference	**ISIC2017** ISIC2018	skip con. residual con. parallel m.s. conv. attention mod.	WDice	77.60%	✗	Copping, flipping	✗	✗
Sarker et al. (2019)	Non Peer-reviewed technical report	**ISIC2017** ISIC2018	factorized conv. attention mod. GAN	CE Jaccard ℓ_1, ADV	77.98%	✗	Flipping gamma reconst. contrast adjust.	✗	✗
Tu et al. (2019)	Peer-reviewed journal	**ISIC2017** PH^2	skip con. residual con. dense con. GAN	Jaccard EPE, ℓ_1 DS, ADV	76.80%	✓	Flipping	✗	✗
Wei et al. (2019)	Peer-reviewed journal	ISIC2016 **ISIC2017** PH^2	skip con. residual con. attention mod. GAN	Jaccard ℓ_1 ADV	80.45%	✓	Rotation, flipping color jittering	✗	✗
Ünver and Ayan (2019)	Peer-reviewed journal	**ISIC2017** PH^2	–	ℓ_2	74.81%	✓	–	✓	✗
Al-masni et al. (2019)	Peer-reviewed conference	ISIC2017	–	–	77.11%	✗	Rotation, flipping	✗	✗
Canalini et al. (2019)	Peer-reviewed conference	ISIC2017	dilated conv. parallel m.s. conv. separable conv.	CE Tanimoto	85.00%	✗	Rotating, flipping shifting, shearing scaling color jittering	✓	✗
Wang et al. (2019b)	Peer-reviewed conference	ISIC2017	residual con.	WCE	78.10%	✗	Flipping, scaling	✗	✗
Alom et al. (2020)	Peer-reviewed conference	ISIC2018	skip con. residual con. recurrent CNN	CE	88.83%	✗	Flipping	✗	✗
Pollastri et al. (2020)	Peer-reviewed journal	ISIC2017	–	Tanimoto	78.90%	✗	GAN flipping, rotation shifting, scaling color jittering	✗	✗
Liu et al. (2019b)	Peer-reviewed conference	ISIC2017	skip con. dilated conv.	CE	75.20%	✗	Scaling, cropping rotation, flipping image deformation	✗	✗
Abhishek and Hamarneh (2019)	Peer-reviewed conference	**ISIC2017** PH^2	skip con.	–	68.69%*	✓	Rotation, flipping GAN	✗	✓
Shahin et al. (2019)	Peer-reviewed conference	**ISIC2018**	skip con. image pyramid	Generalized Dice	73.8%	✗	Rotation flipping zooming	✗	✗
Adegun and Viriri (2019)	Peer-reviewed conference	**ISIC2017**	–	Dice	83.0%	✗	Elastic	✗	✗
Taghanaki et al. (2019)	Peer-reviewed conference	ISIC 2017	skip con.	Dice ℓ_1 SSIM	69.35%*	✗	Rotation flipping gradient-based perturbation	✗	✗
Saini et al. (2019)	Peer-reviewed conference	**ISIC 2017** ISIC 2018 PH2	skip con. multi-task	Dice	84.9%	✗	Rotation, flipping shearing, stretch crop, contrast	✗	✗
Wang et al. (2019c)	Peer-reviewed journal	ISIC2016 **ISIC2017**	skip con. residual con. dilated conv.	WCE	81.47%	✗	Flipping, scaling	✗	✗
Kamalakannan et al. (2019)	Peer-reviewed journal	ISIC Archive	skip con.	CE	–	✗	–	✗	✗
Hasan et al. (2020)	Peer-reviewed journal	**ISIC2017** PH^2	skip con. dense con. separable conv.	CE Jaccard	77.50%	✓	Rotation, zooming shifting, flipping	✗	✓

(*continued on next page*)

Table 3 (*continued*).

Al Nazi and Abir (2020)	Peer-reviewed conference	**ISIC2018** PH^2	skip con.	Dice	80.00%	✓	Rotation, zooming flipping, elastic dist. Gaussian dist. histogram equal. color jittering	✗	✓
Deng et al. (2020)	Peer-reviewed conference	**ISIC2017** PH^2	dilated conv. parallel m.s. conv. separable conv. semi-supervised	Dice Narrowband suppression	83.9%	✓	Rotation	✓	✗
Xie et al. (2020b)	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	✗	✓
Zhang et al. (2020a)	Peer-reviewed conference	SCD ISIC2016 **ISIC2017** ISIC2018	skip con.	Kappa Loss	84.00%*	✗	Rotation, shifting shearing, zooming flipping	✗	✓
Saha et al. (2020)	Peer-reviewed conference	ISIC2017 **ISIC2018**	skip con. dense con.	CE	81.9%	✗	Color jittering rotation flipping translation	✗	✗
Henry et al. (2020)	Peer-reviewed conference	**ISIC2018**	skip con. parallel m. s. conv. attention mod.	–	78.04%	✗	Color jittering rotation, cropping flipping, shift	✗	✓
Jafari et al. (2020)	Peer-reviewed conference	**ISIC2018**	skip con. residual con. dense con.	CE	75.5%	✗	–	✗	✓
Li et al. (2020a)	Peer-reviewed conference	**ISIC2018**	skip con. residual con. ensemble semi-supervised	CE Dice	75.5%	✗	–	✗	✗
Guo et al. (2020)	Peer-reviewed conference	**ISIC2018**	skip con. dilated conv. parallel m. s. conv.	Focal Jaccard	77.60%	✗	–	✗	✓
Li et al. (2020b)	Peer-reviewed conference	**ISIC2018**	skip con. residual con. self-supervised	MSE KLD	87.74%*	✗	–	✗	✗
Jiang et al. (2020)	Peer-reviewed journal	**ISIC2017** PH^2	skip con. residual con. attention mod.	CE	73.35%	✗	Flipping	✗	✗
Qiu et al. (2020)	Peer-reviewed journal	**ISIC2017** PH^2	ensemble	–	80.02%	✗	Translation rotation shearing	✓	✗
Xie et al. (2020a)	Peer-reviewed journal	ISIC2016 **ISIC2017** PH^2	attention mod.	CE	78.3%	✗	Rotation flipping	✗	✗
Zafar et al. (2020)	Peer-reviewed journal	**ISIC2017** PH^2	skip con. residual con.	CE	77.2%	✗	Rotation	✗	✗
Azad et al. (2020)	Peer-reviewed conference	**ISIC 2017** ISIC 2018 PH2	dilated conv. attention mod.	–	96.98%	✗	–	✗	✓
Nathan and Kansal (2020)	Non Peer-reviewed technical report	ISIC 2016 **ISIC 2017** ISIC 2018 PH2	skip con. residual con.	CE Dice	78.28%	✗	Rotation, flipping shearing, zoom	✗	✗

(*continued on next page*)

Table 3 (*continued*).

Mirikharaji et al. (2021)	Peer-reviewed conference	**ISIC Archive** PH2 DermoFit	skip con. residual con. ensemble	CE	72.11%	✗	–	✗	✗
Öztürk and Özkaya (2020)	Peer-reviewed journal	**ISIC 2017** PH2	residual con.	–	78.34%	✓	–	✗	✗
Abhishek et al. (2020)	Peer-reviewed conference	**ISIC 2017** DermoFit PH2	skip con.	Dice	75.70%	✓	Rotation flipping	✗	✓
Kaymak et al. (2020)	Peer-reviewed journal	ISIC 2017	–	–	72.5%	✗	–	✗	✗
Bagheri et al. (2020)	Peer-reviewed journal	**ISIC2017** Der-mQuest	dilated conv. parallel m.s. conv. separable conv.	–	79.05%	✓	Rotation, flipping brightness change resizing	✗	✗
Jayapriya and Jacob (2020)	Peer-reviewed journal	ISIC2016	skip con. parallel m.s. conv.	–	92.42%	✗	–	✗	✗
Wang et al. (2020a)	Non Peer-reviewed technical report	ISIC2016 **ISIC2017** PH^2	residual con. dilated conv. attention mod.	CE Dice DS	80.30%	✓	Flipping, rotation cropping	✗	✗
Wang et al. (2020b)	Non Peer-reviewed technical report	**ISIC2018** PH^2	attention mod. skip con. parallel m.s. conv. recurrent CNN	Dice Focal Tversky	80.6%	✗	Rotation flipping cropping	✗	✗
Ribeiro et al. (2020)	Peer-reviewed conference	ISIC Archive PH^2 DermoFit	skip con. residual con. dilated conv.	Soft Jaccard CE	–	✓	Gaussian noise color jittering	✓	✓
Zhu et al. (2020)	Peer-reviewed conference	ISIC2018	skip con. residual con. dilated conv. attention mod.	CE Dice	82.15%	✗	Flipping	✗	✗
Gu et al. (2020)	Peer-reviewed journal	ISIC 2018	residual con. skip con. attention mod.	Dice	85.32%*	✗	Cropping, flipping rotation	✗	✓
Lei et al. (2020)	Peer-reviewed journal	**ISIC 2017** ISIC 2018	skip con. dense con. dilated conv. GAN	CE ℓ_1 ADV	77.1%	✓	Flipping, rotation	✗	✗
Andrade et al. (2020)	Peer-reviewed journal	**DermoFit** SMART-SKINS	residual con. dilated conv. GAN	Dice	81.03%	✗	Flipping, brightness saturation, contrast, hue Gaussian hue	✗	✗
Wu et al. (2020)	Peer-reviewed journal	**ISIC 2017** ISIC 2018	residual con. attention mod. multi-scale	CE Dice	82.55%	✗	Flipping, rotation scaling, cropping sharpening, color distribution adj., noise	✗	✗
Arora et al. (2021)	Peer-reviewed journal	ISIC 2018	skip con. attention mod.	Dice Tversky Focal Tversky	83%	✗	Flipping	✓	✗
Jin et al. (2021)	Peer-reviewed journal	**ISIC2017** ISIC2018	skip con. residual con. attention mod.	Dice Focal	80.00%	✗	Flipping, rotation affine trans. scaling, cropping	✗	✓
Hasan et al. (2021)	Peer-reviewed journal	ISIC 2016 **ISIC 2017**	skip con. residual con. separable conv.	Dice CE	66.66%*	✗	Flipping, rotation shifting, zooming intensity adjust.	✗	✗

(*continued on next page*)

Table 3 (*continued*).

Kosgiker et al. (2021)	Peer-reviewed journal	**ISIC 2017** PH^2	–	MSE CE	90.25%	✗	–	✗	✗
Bagheri et al. (2021a)	Peer-reviewed journal	ISIC2016 ISIC2017 **ISIC2018** PH^2 DermQuest	parallel m.s. conv. dilated conv.	Dice CE	85.04%	✓	Rotation flipping color jittering	✗	✗
Saini et al. (2021)	Peer-reviewed conference	**ISIC2017** ISIC2018 PH^2	pyramid pooling residual con. skip con. dilated conv. attention mod.	Dice	85.00%	✓	Rotation, shearing color jittering	✗	✗
Tong et al. (2021)	Peer-reviewed journal	ISIC2016 ISIC2017 **PH^2**	skip con. attention mod.	CE	84.2%	✓	Flipping	✗	✗
Bagheri et al. (2021b)	Peer-reviewed journal	**DermQuest** ISIC2017 PH^2	ensemble	CE Focal	86.53%	✓	Rotation flipping color jittering	✓	✗
Ren et al. (2021)	Peer-reviewed journal	ISIC2017	dense con. dilated conv. separable conv. attention mod.	Dice CE	76.92%	✗	Flipping, rotation	✗	✗
Liu et al. (2021a)	Peer-reviewed journal	ISIC2017	residual con. dilated conv. pyramid pooling	WCE	79.46%	✗	Flipping, cropping rotation image deformation	✗	✗
Khan et al. (2021)	Peer-reviewed journal	ISIC2018	skip con. image pyramid	Dice	85.10%	✗	–	✗	✓
Redekop and Chernyavskiy (2021)	Peer-reviewed conference	ISIC2017	–	–	68.77%*	✗	–	✗	✗
Kaul et al. (2021)	Peer-reviewed conference	ISIC2018	skip con. residual con. attention mod.	CE Tversky adaptive logarithmic	82.71%	✗	–	✗	✓
Abhishek and Hamarneh (2021)	Peer-reviewed conference	**ISIC2017** PH^2 DermoFit	skip con.	MCC	75.18%	✗	Flipping, rotation	✗	✓
Tang et al. (2021b)	Peer-reviewed journal	ISIC2018	skip con.	CE	78.25%	✗	–	✗	✗
Xie et al. (2021)	Peer-reviewed conference	ISIC2018	dilated conv. semi-supervised	CE KL div.	82.37%	✗	Scaling, rotation elastic transformation	✗	✗
Poudel and Lee (2021)	Peer-reviewed journal	ISIC2017	skip con. attention mod.	CE	87.44%	✗	Scaling, flipping rotation Gaussian noise median blur	✗	✗
Şahin et al. (2021)	Peer-reviewed journal	ISIC2016 **ISIC 2017**	skip con. Gaussian process	–	74.51%	✗	Resize rotation reflection	✓	✗
Sarker et al. (2021)	Peer-reviewed journal	**ISIC 2017** ISIC 2018	parallel m.s. conv. attention mod. GAN	ℓ_1 Jaccard	81.98%	✗	Flipping, contrast gamma reconstruction	✗	✗

(*continued on next page*)

Table 3 (*continued*).

Wang et al. (2021b)	Peer-reviewed journal	ISIC 2016 **ISIC 2017**	residual con. skip con. lesion-based pooling feature fusion	CE	82.4%	✗	Flipping, scaling cropping	✗	✗
Sachin et al. (2021)	book chapter	ISIC 2018	residual con. skip con.	–	75.96%	✗	Flipping, scaling color jittering	✗	✗
Wibowo et al. (2021)	Peer-reviewed journal	**ISIC 2017** ISIC 2018 PH2	BConvLSTM separable conv. residual con. skip con.	Jaccard	80.25%	✗	Distortion, blur color jittering contrast gamma sharpen	✓	✓
Gudhe et al. (2021)	Peer-reviewed journal	ISIC 2018	dilated conv. residual con. skip con.	CE	91%	✗	Flipping, scaling shearing, color jittering Gaussian blur Gaussian noise	✗	✓
Khouloud et al. (2021)	Peer-reviewed journal	ISIC 2016 **ISIC 2017** ISIC 2018 PH2	Feature pyramid residual con. skip con. attention mod.	–	86.92%*	✗	–	✗	✗
Gu et al. (2021)	Peer-reviewed conference	ISIC 2017	asymmetric conv. skip con.	DS	79.4%	✗	Cropping, flipping rotation	✗	✗
Zhao et al. (2021)	Peer-reviewed journal	ISIC 2018	pyramid pooling attention mod. residual con. skip con.	CE Dice	86.84%	✗	Cropping	✗	✗
Tang et al. (2021a)	Peer-reviewed journal	ISIC 2016 **ISIC 2017** ISIC 2018	attention mod. residual con. skip con. ensemble pyramid pooling	Focal	80.7%	✗	Copying	✗	✗
Zunair and Hamza (2021)	Peer-reviewed journal	ISIC 2018	sharpening kernel residual con.	CE	79.78%	✗	–	✗	✓
Li et al. (2021a)	Peer-reviewed conference	ISIC 2017	skip con.	CE KL div.	71.12%*	✗	–	✗	✓
Zhang et al. (2021a)	Peer-reviewed conference	ISIC 2016	skip con. residual con. feature fusion semi-supervised self-supervised	CE Dice	80.49%	✗	Flipping, rotation zooming, cropping	✗	✓
Xu et al. (2021)	Peer-reviewed conference	ISIC 2018	Transformer multi-scale	Dice	89.6%	✗	Flipping, rotation	✗	✗
Ahn et al. (2021)	Peer-reviewed conference	PH^2	self-supervised clustering	CE Spatial loss Consistency loss	71.53%*	✗	–	✗	✓
Zhang et al. (2021b)	Peer-reviewed conference	ISIC 2017	skip con. feature fusion Transformer	CE Jaccard	79.5%	✗	Rotation, flipping color jittering	✗	✓

(*continued on next page*)

Table 3 (*continued*).

Ji et al. (2021)	Peer-reviewed conference	ISIC 2018	skip con. multi-scale Transformer	CE Dice	82.4%*	✗	Flipping	✗	✓
Wang et al. (2021a)	Peer-reviewed conference	ISIC 2016 **ISIC 2018** PH^2	multi-scale Transformer	CE Dice	84.3%*	✓	Flipping, scaling	✗	✓
Yang et al. (2021)	Peer-reviewed journal	**ISIC 2018** PH^2	skip con. multi-scale feature fusion	CE Dice	94.0%	✗	Rotation, flipping cropping, HSC manipulation, luminance and contrast shift	✗	✗
Tao et al. (2021)	Peer-reviewed journal	**ISIC 2017** PH^2	skip con. dense con. attention mod. multi-scale	–	78.85%	✗	Rotation	✗	✗
Kim and Lee (2021)	Peer-reviewed journal	**ISIC 2016** PH^2	residual con. skip con.	boundary aware loss	84.33%*	✗	–	✗	✗
Dai et al. (2022)	Peer-reviewed journal	**ISIC2018** PH2	residual con. skip con. dilated conv. image pyramid attention mod.	CE Dice SoftDice	83.45%	✓	Cropping, flipping rotation	✗	✗
Bi et al. (2022)	Peer-reviewed journal	ISIC2016 **ISIC2017** PH2	residual con. skip con. attention mod. feature fusion	CE	83.70%	✓	Cropping, flipping	✗	✗
Lin et al. (2022)	Peer-reviewed conference	**ISIC 2017** ISIC 2018	attention mod. Transformer	CE Jaccard DS	77.81%*	✗	Flipping, rotation	✗	✗
Wu et al. (2022b)	Peer-reviewed conference	PH^2	skip con. Transformer multi-scale	CE	70.0%*	✗	–	✗	✗
Valanarasu and Patel (2022)	Peer-reviewed conference	ISIC 2018	skip con.	CE Dice	81.7%	✗	–	✗	✓
Basak et al. (2022)	Peer-reviewed journal	**ISIC 2017** PH^2 HAM10000	residual con. multi-scale attention mod.	CE Jaccard DS	97.4%	✗	–	✗	✓
Wu et al. (2022a)	Peer-reviewed journal	ISIC 2016 **ISIC 2017** ISIC 2018 PH^2	skip con. residual con. attention mod. Transformer	CE Dice	76.53%	✗	Flipping, rotation brightness change contrast change in H, S, V	✗	✓
Liu et al. (2022a)	Peer-reviewed journal	ISIC 2017	skip con. residual con. dilated conv. attention mod.	CE Dice	78.62%	✗	Flipping, rotation	✗	✗
Wang et al. (2022b)	Peer-reviewed journal	ISIC 2017	skip con. residual con. Transformer	–	84.52%	✗	Flipping, rotation	✗	✓
Zhang et al. (2022a)	Peer-reviewed conference	ISIC 2017	skip con. feature fusion	Dice Focal	74.54%	✗	Flipping	✗	✗
Wang et al. (2022d)	Peer-reviewed conference	**ISIC 2017** PH^2	skip con. residual con. self-supervised	Dice	76.5%	✓	Rotation, flipping color jittering	✗	✗

(*continued on next page*)

Table 3 (*continued*).

Dong et al. (2022)	Peer-reviewed journal	ISIC 2016 **ISIC 2017** ISIC 2018	residual con. skip con. Transformer feature fusion	CE Dice	74.55%	✗	–	✗	✗
Chen et al. (2022)	Peer-reviewed journal	**ISIC 2017** PH^2	skip con. attention mod. recurrent net.	CE	80.36%	✓	Flipping, rotation affine trans. masking, mesh distortion	✗	✗
Kaur et al. (2022b)	Peer-reviewed journal	ISIC 2016 **ISIC 2017** ISIC 2018 PH^2	dilated conv.	CE	81.7%	✓	Scaling, rotation translation	✗	✗
Badshah and Ahmad (2022)	Peer-reviewed journal	ISIC 2018	residual con. BConvLSTM	–	94.5%	✗	–	✗	✗
Alam et al. (2022)	Peer-reviewed journal	HAM10000	residual con. separable conv.	Dice	91.1%	✗	–	✗	✓
Yu et al. (2022)	Peer-reviewed journal	ISIC 2018	skip con. attention mod. multi-scale	–	87.89%	✗	–	✗	✗
Jiang et al. (2022)	Peer-reviewed journal	**ISIC 2017** ISIC 2018	skip con. attention mod. ConvLSTM	CE Jaccard	80.5%	✗	–	✗	✗
Ramadan et al. (2022)	Peer-reviewed journal	ISIC 2018	skip con. attention mod.	CE Dice sens.-spec. loss	91.4%	✗	–	✗	✗
Zhang et al. (2022b)	Peer-reviewed journal	**ISIC 2017** ISIC 2018	skip con. dense con. semi-supervised	CE contrastive loss	73.89%	✗	Scaling, flipping color distortion	✗	✗
Tran and Pham (2022)	Peer-reviewed journal	**ISIC 2017** PH^2	skip con. attention mod.	Focal Tversky fuzzy loss	79.2%	✗	Rotation, zooming flipping	✗	✗
Wang and Wang (2022)	Peer-reviewed journal	ISIC 2017	skip con. residual con. attention mod.	CE Jaccard	78.28%	✗	Rotation, zooming resizing, shifting	✗	✗
Zhao et al. (2022b)	Peer-reviewed conference	ISIC 2017	skip con. self-supervised	CE Dice	67.08%*	✗	–	✗	✗
Wang et al. (2022c)	Peer-reviewed conference	PH^2	few shot mask avg. pooling	Dice	86.97%*	✗	–	✗	✗
Wang et al. (2022a)	Peer-reviewed conference	**ISIC 2017** ISIC 2018	residual con. dilated conv. multi-scale feature fusion Transformer	CE Jaccard	78.76%	✗	–	✗	✗
Liu et al. (2022b)	Peer-reviewed conference	**ISIC 2017** ISIC 2018	skip con. dilated conv. multi-scale pyramid pooling Transformer	CE	80.19%	✗	–	✗	✗
Gu et al. (2022)	Peer-reviewed journal	ISIC 2017	skip con. global adaptive pooling	CE ℓ_2	80.53%	✗	Scaling, rotation flipping	✗	✗

(*continued on next page*)

Table 3 (*continued*).

Khan et al. (2022)	Peer-reviewed journal	**ISIC 2017** PH^2	residual con. attention mod. ensemble	CE	79.2%	✗	–	✗	✗
Alahmadi and Alghamdi (2022)	Peer-reviewed journal	**ISIC 2017** ISIC 2018 PH^2	skip con. feature fusion semi-supervised Transformer	CE Dice ℓ_2	82.78%*	✗	–	✗	✗
Li et al. (2022)	Peer-reviewed journal	ISIC 2018	skip con. residual con. dilated conv. attention mod. pyramid pooling multi-scale	CE Dice	88.92%	✗	Flipping, rotation	✗	✗
Kaur et al. (2022a)	Peer-reviewed journal	ISIC 2016 **ISIC 2017** ISIC 2018 PH^2	–	Tversky	77.8%	✓	Rotation, scaling	✗	✗

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 and Aldeen (2011) parameterize the weights of TP, FN, FP, and TN in their weighted performance index and then use a constrained non-linear 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.

We make the following key observations about the popular evaluation metrics and how they have been used in the skin lesion segmentation literature:

- 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, tracing their usage in serologic test reports in the early 1900s (Binney et al., 2021). 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 (Luque et al., 2020).
- 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 (Zou et al., 2004; Chicco and Jurman, 2020; Luque et al., 2020).
- MCC is equivalent to the *phi coefficient*, which is simply the *Pearson correlation coefficient* applied to binary data (Chicco and Jurman, 2020). 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 (Matthews, 1975). Although it is biased to a certain extent (Luque et al., 2020; Zhu, 2020), this measure appears to be suitable for class-imbalanced domains (Boughorbel et al., 2017; Chicco and Jurman, 2020; Luque et al., 2020).
- J (aka *Intersection over Union* Jaccard, 1912) and F (aka *Dice coefficient* aka *Sørensen–Dice coefficient* Dice, 1945; Sørensen, 1948) are highly popular in medical image segmentation (Crum et al., 2006). 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). (ii) It can be shown (Zijdenbos et al., 1994) that if TN is sufficiently large compared to TP, FN, and FP, which is common in skin lesion segmentation, F becomes equivalent to *Cohen's kappa* (Cohen, 1960), 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, they are 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 (Celebi et al., 2009c). Nevertheless, owing to its intuitive formulation, XOR was popular in skin lesion segmentation until about 2015 (Celebi et al., 2015b).
- The 2016 and 2017 ISIC Challenges (Gutman et al., 2016; Codella et al., 2018) adopted five measures: AC, SE, SP, F, and J, with the participants ranked based on the last measure. The 2018 ISIC Challenge (Codella et al., 2019) 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 challenge organizers set the threshold equal to 0.65 based on an earlier study (Codella et al., 2017) that determined the average pairwise J similarities among the manual segmentations outlined by three expert dermatologists. Since the

majority of papers in this survey (168 out of 177 papers) use the ISIC datasets (Fig. 3), we list the J for all the papers in Table 3 wherever it has been reported in the corresponding papers. For papers that did not report J and instead reported F, we list the computed J based on F and denote it with an asterisk.

- Some of the aforementioned measures (i.e., GM and BA) have *not* been used in a skin lesion segmentation study yet.
- The evaluation measures discussed above are all region-based and thus fairly insensitive to border irregularities (Lee et al., 2003), i.e., indentations, and protrusions along the border. Boundary-based evaluation measures (Taha and Hanbury, 2015) have *not* been used in the skin lesion segmentation literature much except for the symmetric Hausdorff metric (Silveira et al., 2009), which is known to be sensitive to noise (Huttenlocher et al., 1993) and biased in favor of small lesions (Bogo et al., 2015).

5. Discussion and future research

In this paper, we presented an overview of DL-based skin lesion segmentation algorithms. A lot of work has been done in this field since the first application of CNNs on these images in 2015 (Codella et al., 2015). In fact, the number of skin lesion segmentation papers published over the past 8 years (2015–2022) is more than thrice those published over the previous 17 years (1998–2014) (Celebi et al., 2015b).

However, despite the large body of work, skin lesion segmentation remains an open problem, as evidenced by the ISIC 2018 Skin Lesion Segmentation Live Leaderboard (ISIC, 2018). The live leaderboard has been open and accepting submissions since 2018, and even after the permitted usage of external data, the best thresholded Jaccard index (the metric used to rank submissions) is 83.6%. Additionally, the release of the HAM10000 lesion segmentations (Tschandl et al., 2020; ViDIR Dataverse, 2020) in 2020 shows that progressively larger skin lesion segmentation datasets continue to be released. We believe that the following aspects of skin lesion segmentation via deep learning are worthy of future work:

- Mobile dermoscopic image analysis: With the availability of various inexpensive dermoscopes designed for smartphones, mobile dermoscopic image analysis is of great interest worldwide, especially in regions where access to dermatologists is limited. Typical DL-based image segmentation algorithms have millions of weights. In addition, classical CNN architectures are known to exhibit difficulty dealing with certain image distortions such as noise and blur (Dodge and Karam, 2016), and DL-based skin lesion diagnosis models have been demonstrated to be susceptible to similar artifacts: various kinds of noise and blur, brightness and contrast changes, dark corners (Maron et al., 2021b), bubbles, rulers, ink markings, etc. (Katsch et al., 2022). Therefore, the current dermoscopic image segmentation algorithms may not be ideal for execution on typically resource-constrained mobile and edge devices, needed for patient privacy so that uploading skin images to remote servers is avoided. Leaner DL architectures, e.g., MobileNet (Howard et al., 2019), ShuffleNet (Zhang et al., 2018), EfficientNet (Tan and Le, 2019), MnasNet (Tan et al., 2019a), and UNeXt (Valanarasu and Patel, 2022), should be investigated in addition to the robustness of such architectures with respect to image noise and blur.
- Datasets: To train more accurate and robust deep neural segmentation architectures, we need larger, more diverse, and more representative skin lesion datasets with multiple manual segmentations per image. Additionally, as mentioned in Section 2.1, several skin lesion image classification datasets do not have the corresponding lesion mask annotations, and given their popularity in skin image analysis tasks, they may be good targets for manual delineations. For example, the PAD-UFES-20 dataset (Pacheco et al., 2020) consists of clinical images of skin lesions captured using smartphones, and obtaining ground-truth segmentations on this dataset would help advance skin image analysis on mobile devices. Additionally, a recent study conducted by Daneshjou et al. (2021a) found that as little as 10% of the AI-based studies for dermatological diagnosis included skin tone information for at least one dataset used, and that several studies included little to no images of darker skin tones, underlining the need to curate datasets with diverse skin tones.
- Collecting segmentation annotations: At the time of this writing, the ISIC Archive contains over 71,000 publicly available images. Considering that the largest public dermoscopic image set contained a little over 1000 images about six 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 (Kovashka et al., 2016) could be explored to collect annotations from non-experts. Experts could then revise these initial annotations, or methods that tackle the problem of annotation noise (Mirikharaji et al., 2019; Karimi et al., 2020; Li et al., 2021a) could be explored. Note that the utility of crowdsourcing in medical image annotation has been demonstrated in multiple studies (Foncubierta-Rodriguez and Muller, 2012; Gurari et al., 2015; Sharma et al., 2017; Goel et al., 2020). Additionally, keeping in mind the time-consuming nature of manual supervised annotation, an alternative is to use weakly-supervised annotation, e.g., bounding-box annotations (Dai et al., 2015; Papandreou et al., 2015), which are much less time-consuming to collect. For example, for several large skin lesion image datasets that do not have any lesion mask annotations (see Section 2.1), bounding-box lesion annotations can be obtained more easily than dense pixel-level segmentation annotations. In addition, weakly-supervised annotation (Bearman et al., 2016; Tajbakhsh et al., 2020; Roth et al., 2021; En and Guo, 2022) is more amenable to crowdsourcing (Maier-Hein et al., 2014; Rajchl et al., 2016; Papadopoulos et al., 2017; Lin et al., 2019), especially for non-experts.
- Handling multiple annotations per image: If the skin lesion image dataset at hand contains multiple manual segmentations per image, one should consider either using an algorithm such as STAPLE (Warfield et al., 2004) for fusing the manual segmentations (see Section 4), or relying on learning-based approaches, either through variants of STAPLE adapted for DL-based segmentation (Kats et al., 2019; Zhang et al., 2020b), or other methods (Mirikharaji et al., 2021; Lemay et al., 2022). 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 skin image analysis literature (see Section 4.3) are often region-based, pair-counting measures. Other region-based measures, such as information-theoretic measures (e.g., mutual information, variation of information, etc.) as well as boundary-based measures e.g., Hausdorff distance (Taha and Hanbury, 2015) should be explored as well.
- Unsupervised segmentation and unsupervised segmentation evaluation: Current DL-based skin lesion segmentation algorithms are mostly based on supervised learning, as shown in a supervision-level breakdown of the surveyed works (Fig. 5), meaning that these algorithms require manual segmentations for training segmentation prediction models. 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 skin lesion images, it may be beneficial to investigate unsupervised DL (Ji et al., 2019) as well as unsupervised segmentation evaluation (Chabrier et al., 2006; Zhang et al., 2008).

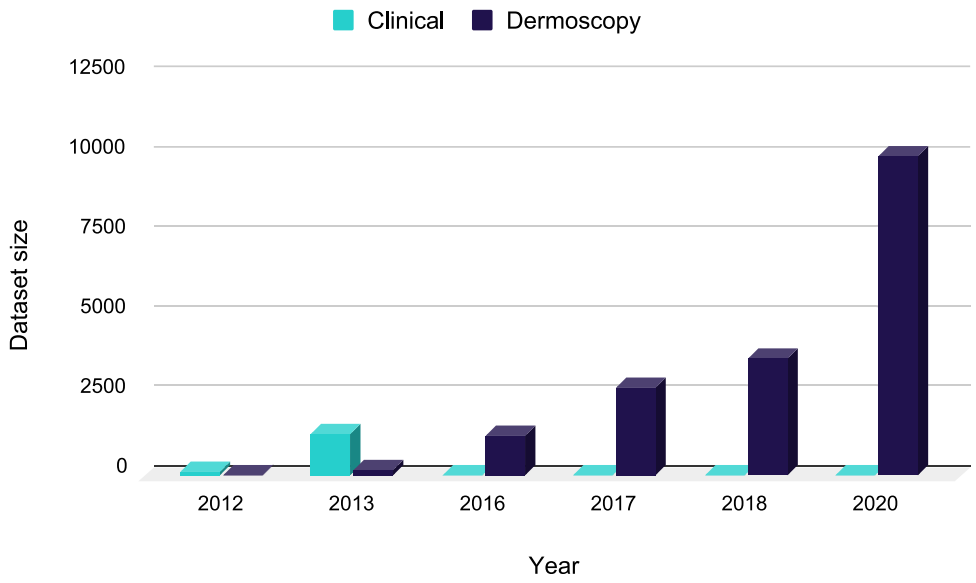


Fig. 10. Number of skin lesion images with ground-truth segmentation maps per year categorized based on modality. It is evident that while the number of dermoscopic skin lesion images has been constantly on the rise, the number of clinical images has remained unchanged for the past several years.

- Systematic evaluations: Systematic evaluations that have been performed for skin lesion classification (Valle et al., 2020; Bissoto et al., 2021; Perez et al., 2018) are, so far, nonexistent in the skin lesion segmentation literature. For example, statistical significance analysis are conducted on the results of a few prior studies in skin lesion segmentation, e.g., Fortina et al. (2012).
- Fusion of hand-crafted and deep features: Can we integrate the deep features extracted by DL models and hand-crafted features synergistically? For example, exploration of shape and appearance priors of skin lesions that may be beneficial to incorporate, via loss terms (Nosrati and Hamarneh, 2016; El Jurdi et al., 2021; Ma et al., 2021), in deep learning models for skin lesion segmentation, similar to star-shape (Mirikharaji and Hamarneh, 2018) and boundary priors (Wang et al., 2021a).
- 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) (Long et al., 2015), atrous (or dilated) convolution (Chen et al., 2017a), and conditional random fields (Krahenbuhl and Koltun, 2011). More research needs to be conducted to determine appropriate strategies for skin lesion segmentation that effectively minimize or avoid the loss of spatial resolution.
- 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 (Bengio, 2012) and tuning these hyperparameters systematically is a laborious undertaking. *Neural architecture search* is an active area of research (Elsken et al., 2019), and some of these model selection approaches have already been applied to semantic segmentation (Liu et al., 2019a) and medical image segmentation (Weng et al., 2019).
- Reproducibility of results: Kapoor and Narayanan (2022) define research in ML-based science to be reproducible if the associated datasets and the code are publicly available and if there are no problems with the data analysis, where problems include the lack of well-defined training and testing partitions of the dataset, leakage across dataset partitions, features selection using the entire dataset instead of only the training partition, etc. Since several skin lesion segmentation datasets come with standardized partitions (Table 1), sharing of the code can lead to more reproducible research (Colliot et al., 2022), with the added benefit to researchers who release their code to be cited significantly more (Vandewalle, 2012). In our analysis, we found that only 38 of the 177 surveyed papers (21.47%) had publicly accessible code (Table 3), a proportion similar to a smaller-scale analysis by Renard et al. (2020) for medical image segmentation. Another potential assessment of a method's generalization performance is its evaluation on a common held-out test set, where the ground truth segmentation masks are private, and users submit their test predictions to receive a performance assessment. For example, the ISIC 2018 dataset's test partition is available through a live leaderboard (ISIC, 2018), but it is rarely used. We found that out of 71 papers published in 2021 and 2022 included in this survey, 36 papers reported results on the ISIC 2018 dataset, but only 1 paper (Saini et al., 2021) used the online submission platform for evaluation.
- Research on clinical images: Another limitation is the limited number of benchmark datasets of clinical skin lesion images with expert pixel-level annotations. Fig. 10 shows that while the number of dermoscopic image datasets with ground-truth segmentation masks has been increasing over the last few years, only a few datasets with clinical images are available. In contrast to dermoscopic images requiring a special tool that is not always utilized even by dermatologists (Engasser and Warshaw, 2010), clinical images captured by digital cameras or smartphones have the advantage of easy accessibility, which can be utilized to evaluate the priority of patients by their lesion severity level, i.e., triage patients. As shown in Fig. 3 and Table 3, most of the deep skin lesion segmentation models are trained and evaluated on dermoscopic images, primarily because of the lack of large-scale clinical skin lesion image segmentation datasets (Table 1), leaving the need to develop automated tools for non-specialists unmet.
- Research on total body images: While there has been some research towards detecting and tracking skin lesions over time in 2D wide-field images (Mirzaalian et al., 2016; Li et al., 2017; Korotkov et al., 2019; Soenksen et al., 2021; Huang et al., 2022) and in 3D total body images (Bogo et al., 2014; Zhao et al., 2022a; Ahmedt-Aristizabal et al., 2023), simultaneous segmentation of

skin lesions from total body images (Sinha et al., 2023) would help with early detection of melanoma (Halpern, 2003; Hornung et al., 2021), thus improving patient outcomes.

- Effect on downstream tasks: End-to-end systems have been proposed for skin images analysis tasks that directly learn the final tasks (e.g., predicting the diagnosis Kawahara et al., 2019 or the clinical management decisions Abhishek et al., 2021 of skin lesions), and these approaches present a number of advantages such as computational efficiency and ease of optimization. On the other hand, skin lesion diagnosis pipelines have been shown to benefit from the incorporation of prior knowledge, specifically lesion segmentation masks (Yan et al., 2019). Therefore, it is worth investigating how lesion segmentation, often an intermediate step in the skin image analysis pipeline, affects the downstream dermatological tasks.
- From binary to multi-class segmentation: While the existing work in skin lesion segmentation is mainly binary segmentation, future work may explore multi-class settings. For example, automated detection and delineation of clinical dermoscopic features (e.g., globules, streaks, pigment networks) within a skin lesion may lead to superior classification performance. Further, dermoscopic feature extraction, a task in the ISIC 2016 (Gutman et al., 2016) and 2017 (Codella et al., 2018) challenges, can be formulated as a multi-class segmentation problem (Kawahara and Hamarneh, 2018). The multiclass formulation can then be addressed by DL models, and can be used either as an intermediate step for improving skin lesion diagnosis or used directly in diagnosis models for regularizing attention maps (Yan et al., 2019). Similarly, multi-class segmentation scenarios may also include multiple skin pathologies on one subject, especially in images with large fields of view, or segmentation of the skin, the lesion(s), and the background, especially in in-the-wild images with diverse backgrounds, such as those in the Fitzpatrick17k dataset (Groh et al., 2021).
- Transferability of models: As the majority of skin lesion datasets are from fair-skinned patients, the generalizability of deep models to populations with diverse skin complexions is questionable. With the emergence of dermatological datasets with diverse skin tones (Groh et al., 2021; Daneshjou et al., 2021b) and methods for diagnosing pathologies fairly (Bevan and Atapour-Abarghouei, 2022; Wu et al., 2022c; Pakzad et al., 2023; Du et al., 2023), it is important to assess the transferability of DL-based skin lesion segmentation models to datasets with diverse skin tones.

CRediT authorship contribution statement

Zahra Mirikharaji: Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Kumar Abhishek:** Conceptualization, Investigation, Writing – original draft, Writing – review & editing. **Alceu Bissoto:** Investigation, Writing – original draft. **Catarina Barata:** Conceptualization, Writing – original draft, Writing – review & editing. **Sandra Avila:** Writing – review & editing. **Eduardo Valle:** Writing – review & editing. **M. Emre Celebi:** Writing – original draft, Writing – review & editing. **Ghassan Hamarneh:** Conceptualization, Methodology, Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Table (HTML and Google Sheets) of surveyed literature is publicly available online.

Acknowledgments

The authors would like to acknowledge Ben Cardoen and Aditi Jain for help with proofreading the manuscript and with creating the interactive table, respectively. Z. Mirikharaji, K. Abhishek, and G. Hamarneh are partially funded by the BC Cancer Foundation - BrainCare BC Fund, Canada, the Natural Sciences and Engineering Research Council of Canada, Canada (NSERC RGPIN-06752), and the Canadian Institutes of Health Research, Canada (CIHR OQI-137993). A. Bissoto is partially funded by FAPESP, Brazil 2019/19619-7. E. Valle is partially funded by CNPq, Brazil 315168/2020-0. S. Avila is partially funded by CNPq, Brazil PQ-2 315231/2020-3, and FAPESP, Brazil 2013/08293-7. A. Bissoto and S. Avila are also partially funded by Google LARA 2020, USA. The RECOD.ai lab is supported by projects from FAPESP, CNPq, and CAPES. C. Barata is funded by FCT project and multi-year funding, Portugal [CEECIND/00326/2017] and LARSyS - FCT Plurianual funding, Portugal 2020–2023. M. E. Celebi was supported by the US National Science Foundation, USA under Award No. OIA-1946391. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Abbas, Q., Celebi, M.E., Garcia, I.F., 2011. Hair removal methods: A comparative study for dermoscopy images. Biomed. Signal Process. Control 6 (4), 395–404.

Abbasi, N.R., Shaw, H.M., Rigel, D.S., Friedman, R.J., McCarthy, W.H., Osman, I., Kopf, A.W., Polsky, D., 2004. Early diagnosis of cutaneous melanoma: Revisiting the ABCD criteria. JAMA 292 (22), 2771–2776.

Abdelhalim, I.S.A., Mohamed, M.F., Mahdy, Y.B., 2021. Data augmentation for skin lesion using self-attention based progressive generative adversarial network. Expert Syst. Appl. 165, 113922.

Abhishek, K., 2020. Input Space Augmentation for Skin Lesion Segmentation in Dermoscopic Images (Master's thesis). Applied Sciences: School of Computing Science, Simon Fraser University, https://summit.sfu.ca/item/20247.

Abhishek, K., Hamarneh, G., 2019. Mask2Lesion: Mask-constrained adversarial skin lesion image synthesis. In: International Workshop on Simulation and Synthesis in Medical Imaging. Springer, pp. 71–80.

Abhishek, K., Hamarneh, G., 2021. Matthews correlation coefficient loss for deep convolutional networks: Application to skin lesion segmentation. In: 2021 IEEE 18th International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 225–229.

Abhishek, K., Hamarneh, G., Drew, M.S., 2020. Illumination-based transformations improve skin lesion segmentation in dermoscopic images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 728–729.

Abhishek, K., Kawahara, J., Hamarneh, G., 2021. Predicting the clinical management of skin lesions using deep learning. Sci. Rep. 11 (1), 1–14.

Abraham, N., Khan, N.M., 2019. 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). IEEE, pp. 683–687.

Adegun, A., Viriri, S., 2019. An enhanced deep learning framework for skin lesions segmentation. In: International Conference on Computational Collective Intelligence. Springer, pp. 414–425.

Adegun, A., Viriri, S., 2020a. Deep learning techniques for skin lesion analysis and melanoma cancer detection: a survey of state-of-the-art. Artif. Intell. Rev. 1–31. http://dx.doi.org/10.1007/s10462-020-09865-y.

Adegun, A.A., Viriri, S., 2020b. FCN-based DenseNet framework for automated detection and classification of skin lesions in dermoscopy images. IEEE Access 8, 150377–150396.

Ahmedt-Aristizabal, D., Nguyen, C., Tychsen-Smith, L., Stacey, A., Li, S., Pathikulangara, J., Petersson, L., Wang, D., 2023. Monitoring of pigmented skin lesions using 3D whole body imaging. Comput. Methods Programs Biomed. 232, 107451.

Ahn, E., Feng, D., Kim, J., 2021. A spatial guided self-supervised clustering network for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 379–388.

Al-Masni, M.A., Al-antari, M.A., Choi, M.-T., Han, S.-M., Kim, T.-S., 2018. Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks. Comput. Methods Programs Biomed. 162, 221–231.

Al-masni, M.A., Al-antari, M.A., Park, H.M., Park, N.H., Kim, T.-S., 2019. 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, IEEE, pp. 95–98.

Al-Masni, M.A., Kim, D.-H., Kim, T.-S., 2020. Multiple skin lesions diagnostics via integrated deep convolutional networks for segmentation and classification. Comput. Methods Programs Biomed. 190, 105351.

Al Nazi, Z., Abir, T.A., 2020. 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. Springer, pp. 371–381.

Alahmadi, M.D., Alghamdi, W., 2022. Semi-supervised skin lesion segmentation with coupling CNN and transformer features. IEEE Access 10, 122560–122569.

Alam, M.J., Mohammad, M.S., Hossain, M.A.F., Showmik, I.A., Raihan, M.S., Ahmed, S., Mahmud, T.I., 2022. S2C-DeLeNet: A parameter transfer based segmentation-classification integration for detecting skin cancer lesions from dermoscopic images. Comput. Biol. Med. 150, 106148.

Alom, M.Z., Aspiras, T., Taha, T.M., Asari, V.K., 2020. Skin cancer segmentation and classification with improved deep convolutional neural network. In: Proceedings of SPIE Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications, Vol. 11318. 1131814.

Alom, M.Z., Yakopcic, C., Hasan, M., Taha, T.M., Asari, V.K., 2019. Recurrent residual U-Net for medical image segmentation. J. Med. Imaging 6 (1), 014006.

American Cancer Society, 2023. Cancer Facts and Figures 2023. American Cancer Society, https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2023/2023-cancer-facts-and-figures.pdf.

Andrade, C., Teixeira, L.F., Vasconcelos, M.J.M., Rosado, L., 2020. Data augmentation using adversarial image-to-image translation for the segmentation of mobile-acquired dermatological images. J. Imaging 7 (1), 2.

Argenziano, G., Soyer, H.P., De Giorgio, V., Piccolo, D., Carli, P., Delfino, M., Ferrari, A., Hofmann-Wellenhof, R., Massi, D., Mazzocchetti, G., Scalvenzi, M., Wolf, I.H., 2000. Interactive Atlas of Dermoscopy. Edra Medical Publishing and New Media.

Arora, R., Raman, B., Nayyar, K., Awasthi, R., 2021. Automated skin lesion segmentation using attention-based deep convolutional neural network. Biomed. Signal Process. Control 65, 102358.

Asgari Taghanaki, S., Abhishek, K., Cohen, J.P., Cohen-Adad, J., Hamarneh, G., 2021. Deep semantic segmentation of natural and medical images: a review. Artif. Intell. Rev. 54 (1), 137–178.

Attia, M., Hossny, M., Nahavandi, S., Yazdabadi, A., 2017. Skin melanoma segmentation using recurrent and convolutional neural networks. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). IEEE, pp. 292–296.

Azad, R., Asadi-Aghbolaghi, M., Fathy, M., Escalera, S., 2019. Bi-directional ConvLSTM U-Net with densley connected convolutions. In: Proceedings of the IEEE International Conference on Computer Vision Workshops.

Azad, R., Asadi-Aghbolaghi, M., Fathy, M., Escalera, S., 2020. Attention Deeplabv3+: Multi-level context attention mechanism for skin lesion segmentation. In: European Conference on Computer Vision Workshops. Springer, pp. 251–266.

Badshah, N., Ahmad, A., 2022. ResBCU-Net: Deep learning approach for segmentation of skin images. Biomed. Signal Process. Control 71, 103137.

Bagheri, F., Tarokh, M.J., Ziaratban, M., 2020. Two-stage skin lesion segmentation from dermoscopic images by using deep neural networks. Jorjani Biomed. J. 8 (2), 58–72.

Bagheri, F., Tarokh, M.J., Ziaratban, M., 2021a. Skin lesion segmentation based on mask RCNN, Multi Atrous Full-CNN, and a geodesic method. Int. J. Imaging Syst. Technol..

Bagheri, F., Tarokh, M.J., Ziaratban, M., 2021b. Skin lesion segmentation from dermoscopic images by using Mask R-CNN, Retina-Deeplab, and graph-based methods. Biomed. Signal Process. Control 67, 102533.

Baghersalimi, S., Bozorgtabar, B., Schmid-Saugeon, P., Ekenel, H.K., Thiran, J.-P., 2019. DermoNet: densely linked convolutional neural network for efficient skin lesion segmentation. EURASIP J. Image Video Process. 2019 (1), 71.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A., Nielsen, H., 2000. Assessing the accuracy of prediction algorithms for classification: An overview. Bioinformatics 16 (5), 412–424.

Ballerini, L., Fisher, R.B., Aldridge, B., Rees, J., 2013. A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions. In: Celebi, M.E., Schaefer, G. (Eds.), Color Medical Image Analysis. Springer, pp. 63–86.

Barata, C., Celebi, M.E., Marques, J.S., 2015a. Improving dermoscopy image classification using color constancy. IEEE J. Biomed. Health Inf. 19 (3), 1146–1152.

Barata, C., Celebi, M.E., Marques, J.S., 2015b. Toward a robust analysis of dermoscopy images acquired under different conditions. In: Celebi, M.E., Mendonca, T., Marques, J.S. (Eds.), Dermoscopy Image Analysis. CRC Press, pp. 1–22.

Barata, C., Celebi, M.E., Marques, J.S., 2019. A survey of feature extraction in dermoscopy image analysis of skin cancer. IEEE J. Biomed. Health Inf. 23 (3), 1096–1109.

Barata, C., Ruela, M., Francisco, M., Mendonca, T., Marques, J.S., 2014. Two systems for the detection of melanomas in dermoscopy images using texture and color features. IEEE Syst. J. 8 (3), 965–979.

Basak, H., Kundu, R., Sarkar, R., 2022. MFSNet: A multi focus segmentation network for skin lesion segmentation. Pattern Recognit. 128, 108673.

Baur, C., Albarqouni, S., Navab, N., 2018. Generating highly realistic images of skin lesions with GANs. In: Proceedings of the Third ISIC Workshop on Skin Image Analysis. pp. 260–267.

Bearman, A., Russakovsky, O., Ferrari, V., Fei-Fei, L., 2016. What's the point: Semantic segmentation with point supervision. In: European Conference on Computer Vision. Springer, pp. 549–565.

Bengio, Y., 2012. Practical recommendations for gradient-based training of deep architectures. In: Montavon, G., Orr, G., Muller, K.R. (Eds.), Neural Networks: Tricks of the Trade, second ed. Springer, pp. 437–478.

Bengio, Y., Courville, A., Vincent, P., 2013. Representation learning: A review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. 35 (8), 1798–1828.

Bevan, P.J., Atapour-Abarghouei, A., 2022. Detecting melanoma fairly: Skin tone detection and debiasing for skin lesion classification. arXiv preprint arXiv:2202.02832.

Bi, L., Feng, D., Fulham, M., Kim, J., 2019a. Improving skin lesion segmentation via stacked adversarial learning. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). IEEE, pp. 1100–1103.

Bi, L., Feng, D., Kim, J., 2018. Improving automatic skin lesion segmentation using adversarial learning based data augmentation. arXiv preprint arXiv:1807.08392.

Bi, L., Fulham, M., Kim, J., 2022. Hyper-fusion network for semi-automatic segmentation of skin lesions. Med. Image Anal. 76, 102334.

Bi, L., Kim, J., Ahn, E., Feng, D., Fulham, M., 2017a. Semi-automatic skin lesion segmentation via fully convolutional networks. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). IEEE, pp. 561–564.

Bi, L., Kim, J., Ahn, E., Kumar, A., Feng, D., Fulham, M., 2019b. Step-wise integration of deep class-specific learning for dermoscopic image segmentation. Pattern Recognit. 85, 78–89.

Bi, L., Kim, J., Ahn, E., Kumar, A., Fulham, M., Feng, D., 2017b. Dermoscopic image segmentation via multistage fully convolutional networks. IEEE Trans. Biomed. Eng. 64 (9), 2065–2074.

Biancardi, A.M., Jirapatnakul, A.C., Reeves, A.P., 2010. A comparison of ground truth estimation methods. Int. J. Comput. Assist. Radiol. Surg. 5 (3), 295–305.

Binder, M., Schwarz, M., Winkler, A., Steiner, A., Kaider, A., Wolff, K., Pehamberger, H., 1995. Epiluminescence microscopy. A useful tool for the diagnosis of pigmented skin lesions for formally trained dermatologists. Arch. Dermatol. 131 (3), 286–291.

Binney, N., Hyde, C., Bossuyt, P.M., 2021. On the origin of sensitivity and specificity. Ann. Intern. Med. 174 (3), 401–407.

Birkenfeld, J.S., Tucker-Schwartz, J.M., Soenksen, L.R., Avilés-Izquierdo, J.A., Marti-Fuster, B., 2020. Computer-aided classification of suspicious pigmented lesions using wide-field images. Comput. Methods Programs Biomed. 195, 105631.

Bissoto, A., Barata, C., Valle, E., Avila, S., 2022. Artifact-based domain generalization of skin lesion models. arXiv preprint arXiv:2208.09756.

Bissoto, A., Fornaciali, M., Valle, E., Avila, S., 2019. (De)constructing bias on skin lesion datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.

Bissoto, A., Perez, F., Valle, E., Avila, S., 2018. 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. pp. 294–302.

Bissoto, A., Valle, E., Avila, S., 2021. 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. pp. 1847–1856.

Bogo, F., Peruch, F., Fortina, A.B., Peserico, E., 2015. Where's the lesion? Variability in human and automated segmentation of dermoscopy images of melanocytic skin lesions. In: Celebi, M.E., Mendonca, T., Marques, J.S. (Eds.), Dermoscopy Image Analysis. CRC Press, pp. 67–95.

Bogo, F., Romero, J., Peserico, E., Black, M.J., 2014. Automated detection of new or evolving melanocytic lesions using a 3D body model. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 593–600.

Boughorbel, S., Jarray, F., El-Anbari, M., 2017. Optimal classifier for imbalanced data using matthews correlation coefficient metric. PLoS One 12 (6), e0177678.

Bozorgtabar, B., Ge, Z., Chakravorty, R., Abedini, M., Demyanov, S., Garnavi, R., 2017a. Investigating deep side layers for skin lesion segmentation. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). IEEE, pp. 256–260.

Bozorgtabar, B., Sedai, S., Roy, P.K., Garnavi, R., 2017b. Skin lesion segmentation using deep convolution networks guided by local unsupervised learning. IBM J. Res. Dev. 61 (4/5), 6–1.

Busin, L., Vandenbroucke, N., Macaire, L., 2008. Color spaces and image segmentation. In: Hawkes, P.W. (Ed.), Advances in Imaging and Electron Physics, Vol. 151. Academic Press, pp. 65–168.

Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A.A., 2020. Albumentations: Fast and flexible image augmentations. Information 11 (2), 125.

Caffery, L.J., Clunie, D., Curiel-Lewandrowski, C., Malvehy, J., Soyer, H.P., Halpern, A.C., 2018. Transforming dermatologic imaging for the digital era: Metadata and standards. J. Digit. Imaging 31, 568–577.

Canalini, L., Pollastri, F., Bolelli, F., Cancilla, M., Allegretti, S., Grana, C., 2019. Skin lesion segmentation ensemble with diverse training strategies. In: International Conference on Computer Analysis of Images and Patterns. Springer, pp. 89–101.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M., 2021. Swin-Unet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537.

Cassidy, B., Kendrick, C., Brodzicki, A., Jaworek-Korjakowska, J., Yap, M.H., 2022. Analysis of the ISIC image datasets: Usage, benchmarks and recommendations. Med. Image Anal. 75, 102305.

Celebi, M.E., Aslandogan, A., Stoecker, W.V., 2007a. Unsupervised border detection in dermoscopy images. Skin Res. Technol. 13 (4), 454–462.

Celebi, M.E., Codella, N., Halpern, A., 2019. Dermoscopy image analysis: Overview and future directions. IEEE J. Biomed. Health Inf. 23 (2), 474–478.

Celebi, M.E., Iyatomi, H., Schaefer, G., Stoecker, W.V., 2009a. Approximate lesion localization in dermoscopy images. Skin Res. Technol. 15 (3), 314–322.

Celebi, M.E., Iyatomi, H., Schaefer, G., Stoecker, W.V., 2009b. Lesion border detection in dermoscopy images. Comput. Med. Imaging Graph. 33 (2), 148–153.

Celebi, M.E., Iyatomi, H., Stoecker, W.V., Moss, R.H., Rabinovitz, H.S., Argenziano, G., Soyer, H.P., 2008. Automatic detection of blue-white veil and related structures in dermoscopy images. Comput. Med. Imaging Graph. 32 (8), 670–677.

Celebi, M.E., Kingravi, H., Uddin, B., Iyatomi, H., Aslandogan, A., Stoecker, W.V., Moss, R.H., 2007b. A methodological approach to the classification of dermoscopy images. Comput. Med. Imaging Graph. 31 (6), 362–373.

Celebi, M.E., Mendonca, T., Marques, J.S. (Eds.), 2015a. Dermoscopy Image Analysis. CRC Press.

Celebi, M.E., Schaefer, G., Iyatomi, H., Stoecker, W.V., Malters, J.M., Grichnik, J.M., 2009c. An improved objective evaluation measure for border detection in dermoscopy images. Skin Res. Technol. 15 (4), 444–450.

Celebi, M.E., Wen, Q., Hwang, S., Iyatomi, H., Schaefer, G., 2013. Lesion border detection in dermoscopy images using ensembles of thresholding methods. Skin Res. Technol. 19 (1), e252–e258.

Celebi, M.E., Wen, Q., Iyatomi, H., Shimizu, K., Zhou, H., Schaefer, G., 2015b. A state-of-the-art survey on lesion border detection in dermoscopy images. In: Celebi, M.E., Mendonca, T., Marques, J.S. (Eds.), Dermoscopy Image Analysis. CRC Press, pp. 97–129.

Chabrier, S., Emile, B., Rosenberger, C., Laurent, H., 2006. Unsupervised performance evaluation of image segmentation. EURASIP J. Adv. Signal Process. 2006, 1–12.

Chalana, V., Kim, Y., 1997. A methodology for evaluation of boundary detection algorithms on medical images. IEEE Trans. Med. Imaging 16 (5), 642–652.

Chen, P., Huang, S., Yue, Q., 2022. Skin lesion segmentation using recurrent attentional convolutional networks. IEEE Access 10, 94007–94018.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y., 2021. TransUNet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.

Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017a. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. 40 (4), 834–848.

Chen, L.-C., Papandreou, G., Schroff, F., Adam, H., 2017b. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587.

Chen, S.E., Parent, R.E., 1989. Shape averaging and its applications to industrial design. IEEE Comput. Graph. Appl. 9 (1), 47–54.

Chen, S., Wang, Z., Shi, J., Liu, B., Yu, N., 2018b. 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). IEEE, pp. 1126–1129.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018a. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 801–818.

Chicco, D., Jurman, G., 2020. The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 21 (1).

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1251–1258.

Chou, P.Y., Fasman, G.D., 1978. Prediction of the secondary structure of proteins from their amino acid sequence. In: Meister, A. (Ed.), Advances in Enzymology and Related Areas of Molecular Biology, Vol. 47. John Wiley & Sons, pp. 45–148.

Codella, N., Cai, J., Abedini, M., Garnavi, R., Halpern, A., Smith, J.R., 2015. Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images. In: Proceedings of the International Workshop on Machine Learning in Medical Imaging. pp. 118–126.

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

Codella, N.C., Nguyen, Q.B., Pankanti, S., Gutman, D.A., Helba, B., Halpern, A.C., Smith, J.R., 2017. Deep learning ensembles for melanoma recognition in dermoscopy images. IBM J. Res. Dev. 61 (4/5), 5:1–5:15.

Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., Kittler, H., Halpern, A., 2019. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). https://arxiv.org/abs/1902.03368.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20 (1), 37–46.

Colliot, O., Thibeau-Sutre, E., Burgos, N., 2022. Reproducibility in machine learning for medical imaging. arXiv preprint arXiv:2209.05097.

Combalia, M., Codella, N.C., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Carrera, C., Barreiro, A., Halpern, A.C., Puig, S., Malvehy, J., 2019. BCN20000: Dermoscopic lesions in the wild. arXiv preprint arXiv:1908.02288.

Cordonnier, J.-B., Loukas, A., Jaggi, M., 2019. On the relationship between self-attention and convolutional layers. arXiv preprint arXiv:1911.03584.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., Bharath, A.A., 2018. Generative adversarial networks: An overview. IEEE Signal Process. Mag. 35 (1), 53–65.

Crum, W.R., Camara, O., Hill, D.L., 2006. Generalized overlap measures for evaluation and validation in medical image analysis. IEEE Trans. Med. Imaging 25 (11), 1451–1461.

Cui, Z., Wu, L., Wang, R., Zheng, W.-S., 2019. Ensemble transductive learning for skin lesion segmentation. In: Chinese Conference on Pattern Recognition and Computer Vision. PRCV, Springer, pp. 572–581.

Curiel-Lewandrowski, C., Novoa, R.A., Berry, E., Celebi, M.E., Codella, N., Giuste, F., Gutman, D., Halpern, A., Leachman, S., Liu, Y., Liu, Y., Reiter, O., Tschandl, P., 2019. Artificial intelligence approach in melanoma. In: Fisher, D.E., Bastian, B.C. (Eds.), Melanoma. Spriner, pp. 599–628.

Dai, D., Dong, C., Xu, S., Yan, Q., Li, Z., Zhang, C., Luo, N., 2022. Ms RED: A novel multi-scale residual encoding and decoding network for skin lesion segmentation. Med. Image Anal. 75, 102293.

Dai, J., He, K., Sun, J., 2015. BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1635–1643.

Daneshjou, R., Barata, C., Betz-Stablein, B., Celebi, M.E., Codella, N., Combalia, M., Guitera, P., Gutman, D., Halpern, A., Helba, B., Kittler, H., Kose, K., Liopyris, K., Malvehy, J., Seog, H.S., Soyer, H.P., Tkaczyk, E.R., Tschandl, P., Rotemberg, V., 2022. Evaluation of image-based AI artificial intelligence reports in dermatology: CLEAR derm consensus guidelines from the international skin imaging collaboration artificial intelligence working group. JAMA Dermatol. 158 (1), 90–96.

Daneshjou, R., Smith, M.P., Sun, M.D., Rotemberg, V., Zou, J., 2021a. Lack of transparency and potential bias in artificial intelligence data sets and algorithms: A scoping review. JAMA Dermatol. 157 (11), 1362–1369.

Daneshjou, R., Vodrahalli, K., Liang, W., Novoa, R.A., Jenkins, M., Rotemberg, V., Ko, J., Swetter, S.M., Bailey, E.E., Gevaert, O., Mukherjee, P., Phung, M., Yekrang, K., Fong, B., Sahasrabudhe, R., Zou, J., Chiou, A., 2021b. Disparities in dermatology AI: Assessments using diverse clinical images. arXiv preprint arXiv:2111.08006.

De Angelo, G.G., Pacheco, A.G., Krohling, R.A., 2019. Skin lesion segmentation using deep learning for images acquired from smartphones. In: 2019 International Joint Conference on Neural Networks. IJCNN, IEEE, pp. 1–8.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 248–255.

Deng, Z., Fan, H., Xie, F., Cui, Y., Liu, J., 2017. Segmentation of dermoscopy images based on fully convolutional neural network. In: 2017 IEEE International Conference on Image Processing. ICIP, IEEE, pp. 1732–1736.

Deng, Z., Xin, Y., Qiu, X., Chen, Y., 2020. Weakly and semi-supervised deep level set network for automated skin lesion segmentation. In: Innovation in Medicine and Healthcare. Springer, pp. 145–155.

Denton, E., Chintala, S., Szlam, A., Fergus, R., 2015. 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. pp. 1486–1494.

Depeweg, S., Hernandez-Lobato, J.-M., Doshi-Velez, F., Udluft, S., 2018. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In: International Conference on Machine Learning. PMLR, pp. 1184–1193.

Der Kiureghian, A., Ditlevsen, O., 2009. Aleatory or epistemic? Does it matter? Struct. Saf. 31 (2), 105–112.

DermIS, 2012. Dermatology information system. https://www.dermis.net/. [Online. Accessed January 26, 2022].

DermQuest, 2012. DermQuest. http://www.dermquest.com. Cited: 2020-04-28.

DeVries, T., Taylor, G.W., 2018. Leveraging uncertainty estimates for predicting segmentation quality. arXiv preprint arXiv:1807.00502.

Dhawan, A.P., Gordon, R., Rangayyan, R.M., 1984. Nevoscopy: Three-dimensional computed tomography of nevi and melanomas in situ by transillumination. IEEE Trans. Med. Imaging 3 (2), 54–61.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26 (3), 297–302.

Ding, S., Zheng, J., Liu, Z., Zheng, Y., Chen, Y., Xu, X., Lu, J., Xie, J., 2021. High-resolution dermoscopy image synthesis with conditional generative adversarial networks. Biomed. Signal Process. Control 64, 102224.

Dodge, S., Karam, L., 2016. Understanding how image quality affects deep neural networks. In: Proceedings of the 2016 International Conference on Quality of Multimedia Experience. pp. 1–6.

Dong, Y., Wang, L., Li, Y., 2022. TC-Net: Dual coding network of transformer and CNN for skin lesion segmentation. Plos One 17 (11), e0277578.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Du, S., Hers, B., Bayasi, N., Hamarneh, G., Garbi, R., 2023. FairDisCo: Fairer AI in dermatology via disentanglement contrastive learning. In: Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV. Springer, pp. 185–202.

Ebenezer, J.P., Rajapakse, J.C., 2018. Automatic segmentation of skin lesions using deep learning. arXiv preprint arXiv:1807.04893.

El Jurdi, R., Petitjean, C., Honeine, P., Cheplygina, V., Abdallah, F., 2021. High-level prior-based loss functions for medical image segmentation: A survey. Comput. Vis. Image Underst. 210, 103248.

Elsken, T., Metzen, J.H., Hutter, F., 2019. Neural architecture search: A survey. J. Mach. Learn. Res. 20, 1–21.

En, Q., Guo, Y., 2022. Annotation by clicks: A point-supervised contrastive variance method for medical semantic segmentation. arXiv preprint arXiv:2212.08774.

Engasser, H.C., Warshaw, E.M., 2010. Dermatoscopy use by US dermatologists: a cross-sectional survey. J. Am. Acad. Dermatol. 63 (3), 412–419.

Erkol, B., Moss, R.H., Stanley, R.J., Stoecker, W.V., Hvatum, E., 2005. Automatic lesion boundary detection in dermoscopy images using gradient vector flow snakes. Skin Res. Technol. 11 (1), 17–26.

Ferreira, P.M., Mendonca, T., Rozeira, J., Rocha, P., 2012. 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. pp. 1–6.

Foncubierta-Rodriguez, A., Muller, H., 2012. Ground truth generation in medical imaging: A crowdsourcing-based iterative approach. In: Proceedings of the ACM Multimedia 2012 Workshop on Crowdsourcing for Multimedia. pp. 9–14.

Fortina, A.B., Peserico, E., Silletti, A., Zattra, E., 2012. Where's the naevus? Inter-operator variability in the localization of melanocytic lesion border. Skin Res. Technol. 18 (3), 311–315.

Friedman, R.J., Rigel, D.S., Kopf, A.W., 1985. Early detection of malignant melanoma: The role of physician examination and self-examination of the skin. CA: Cancer J. Clin. 35 (3), 130–151.

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H., 2019. Dual attention network for scene segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3146–3154.

Gachon, J., Beaulieu, P., Sei, J.F., Gouvernet, J., Claudel, J.P., Lemaitre, M., Richard, M.A., Grob, J.J., 2005. First prospective study of the recognition process of melanoma in dermatological practice. Arch. Dermatol. 141 (4), 434–438.

Gal, Y., 2016. Uncertainty in Deep Learning (Ph.D. thesis). Department of Engineering, University of Cambridge, https://mlg.eng.cam.ac.uk/yarin/thesis/thesis.pdf.

Garnavi, R., Aldeen, M., 2011. Optimized weighted performance index for objective evaluation of border-detection methods in dermoscopy images. IEEE Trans. Inf. Technol. Biomed. 15 (6), 908–917.

Garnavi, R., Aldeen, M., Celebi, M.E., 2011a. Weighted performance index for objective evaluation of borderdetection methods in dermoscopy images. Skin Res. Technol. 17 (1), 35–44.

Garnavi, R., Aldeen, M., Celebi, M.E., Varigos, G., Finch, S., 2011b. Border detection in dermoscopy images using hybrid thresholding on optimized color channels. Comput. Med. Imaging Graph. 35 (2), 105–115.

Gidaris, S., Singh, P., Komodakis, N., 2018. Unsupervised representation learning by predicting image rotations. In: International Conference on Learning Representations. ICLR, pp. 1–16, URL: https://openreview.net/forum?id=S1v4N2l0-.

Giotis, I., Molders, N., Land, S., Biehl, M., Jonkman, M.F., Petkov, N., 2015. MED-NODE: A computer-assisted melanoma diagnosis system using non-dermoscopic images. Expert Syst. Appl. 42 (19), 6578–6585.

Gish, S.L., Blanz, W.E., 1989. 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. pp. 614–621.

Glaister, J.L., 2013. Automatic Segmentation of Skin Lesions from Dermatological Photographs. University of Waterloo, https://uwaterloo.ca/vision-image-processing-lab/research-demos/skin-cancer-detection. Cited: 2022-1-31.

Goel, S., Sharma, Y., Jauer, M.L., Deserno, T.M., 2020. WeLineation: Crowdsourcing delineations for reliable ground truth estimation. In: Proceedings of the Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications. 113180C–1–113180C–8.

Gómez, D.D., Butakoff, C., Ersboll, B.K., Stoecker, W., 2007. Independent histogram pursuit for segmentation of skin lesions. IEEE Trans. Biomed. Eng. 55 (1), 157–161.

Gonzalez-Diaz, I., 2018. Dermaknet: Incorporating the knowledge of dermatologists to convolutional neural networks for skin lesion diagnosis. IEEE J. Biomed. Health Inf. 23 (2), 547–559.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2020. Generative adversarial networks. Commun. ACM 63 (11), 139–144.

Goyal, M., Ng, J., Oakley, A., Yap, M.H., 2019a. Skin lesion boundary segmentation with fully automated deep extreme cut methods. In: Medical Imaging 2019: Biomedical Applications in Molecular, Structural, and Functional Imaging, Vol. 10953. International Society for Optics and Photonics, p. 109530Q.

Goyal, M., Oakley, A., Bansal, P., Dancey, D., Yap, M.H., 2019b. Skin lesion segmentation in dermoscopic images with ensemble deep learning methods. IEEE Access 8, 4171–4181.

Goyal, M., Yap, M.H., Hassanpour, S., 2017. Multi-class semantic segmentation of skin lesions via fully convolutional networks. In: Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies, Comp2Clinic Workshop. pp. 290–295.

Grau, V., Mewes, A.U.J., Alcaniz, M., Kikinis, R., Warfield, S.K., 2004. Improved watershed transform for medical image segmentation using prior information. IEEE Trans. Med. Imaging 23 (4), 447–458.

Green, A., Martin, N., Pfitzner, J., O'Rourke, M., Knight, N., 1994. Computer image analysis in the diagnosis of melanoma. J. Am. Acad. Dermatol. 31 (6), 958–964.

Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A., Badri, O., 2021. Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1820–1828.

Gu, R., Wang, G., Song, T., Huang, R., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., Zhang, S., 2020. CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. IEEE Trans. Med. Imaging 40 (2), 699–711.

Gu, R., Wang, L., Zhang, L., 2022. DE-Net: A deep edge network with boundary information for automatic skin lesion segmentation. Neurocomputing 468, 71–84.

Gu, P., Zheng, H., Zhang, Y., Wang, C., Chen, D.Z., 2021. 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. Springer, pp. 337–347.

Gudhe, N.R., Behravan, H., Sudah, M., Okuma, H., Vanninen, R., Kosma, V.-M., Mannermaa, A., 2021. Multi-level dilated residual network for biomedical image segmentation. Sci. Rep. 11 (1), 1–18.

Guillod, J., Schmid-Saugeon, P., Guggisberg, D., Cerottini, J.P., Braun, R., Krischer, J., Saurat, J.H., Kunt, M., 2002. Validation of segmentation techniques for digital dermoscopy. Skin Res. Technol. 8 (4), 240–249.

Gulzar, Y., Khan, S.A., 2022. Skin lesion segmentation based on vision transformers and convolutional neural networks—A comparative study. Appl. Sci. 12 (12), 5990.

Guo, X., Chen, Z., Yuan, Y., 2020. Complementary network with adaptive receptive fields for melanoma segmentation. In: 2020 IEEE 17th International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 2010–2013.

Gurari, D., Theriault, D., Sameki, M., Isenberg, B., Pham, T.A., Purwada, A., Solski, P., Walker, M., Zhang, C., Wong, J.Y., Betke, M., 2015. 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. pp. 1169–1176.

Gutman, D., Codella, N.C.F., Celebi, M.E., Helba, B., Marchetti, M., Mishra, N., Halpern, A., 2016. 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). http://arxiv.org/abs/1605.01397.

Guy, Jr., G.P., Machlin, S.R., Ekwueme, D.U., Yabroff, K.R., 2015. Prevalence and costs of skin cancer treatment in the US, 2002- 2006 and 2007- 2011. Am. J. Prev. Med. 48 (2), 183–187.

Halpern, A.C., 2003. Total body skin imaging as an aid to melanoma detection. In: Seminars in Cutaneous Medicine and Surgery. pp. 2–8.

Hance, G.A., Umbaugh, S.E., Moss, R.H., Stoecker, W.V., 1996. Unsupervised color image segmentation with application to skin tumor borders. IEEE Eng. Med. Biol. Mag. 15 (1), 104–111.

Hasan, M.K., Dahal, L., Samarakoon, P.N., Tushar, F.I., Martí, R., 2020. DSNet: Automatic dermoscopic skin lesion segmentation. Comput. Biol. Med. 103738.

Hasan, M., Roy, S., Mondal, C., Alam, M., Elahi, M., Toufick, E., Dutta, A., Raju, S., Ahmad, M., et al., 2021. Dermo-DOCTOR: A framework for concurrent skin lesion detection and recognition using a deep convolutional neural network with end-to-end dual encoders. Biomed. Signal Process. Control 68, 102661.

He, K., Gan, C., Li, Z., Rekik, I., Yin, Z., Ji, W., Gao, Y., Wang, Q., Zhang, J., Shen, D., 2022. Transformers in medical image analysis: A review. Intell. Med. http://dx.doi.org/10.1016/j.imed.2022.07.002, URL: https://www.sciencedirect.com/science/article/pii/S2667102622000717.

He, X., Yu, Z., Wang, T., Lei, B., 2017. Skin lesion segmentation via deep RefineNet. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 303–311.

He, X., Yu, Z., Wang, T., Lei, B., Shi, Y., 2018. Dense deconvolution net: Multi path fusion and dense deconvolution for high resolution skin lesion segmentation. Technol. Health Care 26 (S1), 307–316.

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

Henry, H.Y., Feng, X., Wang, Z., Sun, H., 2020. MixModule: Mixed CNN kernel module for medical image segmentation. In: 2020 IEEE 17th International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 1508–1512.

Hornung, A., Steeb, T., Wessely, A., Brinker, T.J., Breakell, T., Erdmann, M., Berking, C., Heppt, M.V., 2021. The value of total body photography for the early detection of melanoma: A systematic review. Int. J. Environ. Res. Public Health 18 (4), 1726.

Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., W., W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q.V., Adam, H., 2019. Searching for MobileNetV3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1314–1324.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7132–7141.

Hu, H., Zhang, Z., Xie, Z., Lin, S., 2019. Local relation networks for image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3464–3473.

Huang, W.-L., Liu, S., Kang, J., Gandjbakhche, A., Armand, M., 2022. DICOM file for total body photography: a work item proposal. In: Photonics in Dermatology and Plastic Surgery 2022, Vol. 11934. SPIE, pp. 64–74.

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

Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J., 1993. Comparing images using the Hausdorff distance. IEEE Trans. Pattern Anal. Mach. Intell. 15 (9), 850–863.

ISIC, 2018. ISIC live leaderboards: 2018.1: Lesion boundary segmentation. https://challenge.isic-archive.com/leaderboards/live/. [Online. Accessed January 17, 2023].

ISIC, 2023. International skin imaging collaboration: Melanoma project. https://www.isic-archive.com/. [Online. Accessed January 17, 2023].

Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1125–1134.

Iyatomi, H., Oka, H., Celebi, M.E., Hashimoto, M., Hagiwara, M., Tanaka, M., Ogawa, K., 2008. An improved internet-based melanoma screening system with dermatologist-like tumor area extraction algorithm. Comput. Med. Imaging Graph. 32 (7), 566–579.

Iyatomi, H., Oka, H., Saito, M., Miyake, A., Kimoto, M., Yamagami, J., Kobayashi, S., Tanikawa, A., Hagiwara, M., Ogawa, K., Argenziano, G., Soyer, H.P., Tanaka, M., 2006. Quantitative assessment of tumor extraction from dermoscopy images and evaluation of computer-based extraction methods for automatic melanoma diagnostic system. Melanoma Res. 16 (2), 183–190.

Izadi, S., Mirikharaji, Z., Kawahara, J., Hamarneh, G., 2018. Generative adversarial networks to segment skin lesions. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE, pp. 881–884.

Jaccard, P., 1901. Distribution de la flore alpine dans le bassin des dranses et dans quelques regions voisines. Bull. Soc. Vaudoise Sci. Nat. 37 (140), 241–272.

Jaccard, P., 1912. The distribution of the flora in the alpine zone. New Phytol. 11 (2), 37–50.

Jafari, M., Auer, D., Francis, S., Garibaldi, J., Chen, X., 2020. DRU-Net: An efficient deep convolutional neural network for medical image segmentation. In: 2020 IEEE 17th International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 1144–1148.

Jafari, M.H., Karimi, N., Nasr-Esfahani, E., Samavi, S., Soroushmehr, S.M.R., Ward, K., Najarian, K., 2016. Skin lesion segmentation in clinical images using deep learning. In: 2016 23rd International Conference on Pattern Recognition. ICPR, IEEE, pp. 337–342.

Jafari, M.H., Nasr-Esfahani, E., Karimi, N., Soroushmehr, S.R., Samavi, S., Najarian, K., 2017. Extraction of skin lesions from non-dermoscopic images for surgical excision of melanoma. Int. J. Comput. Assist. Radiol. Surg. 12 (6), 1021–1030.

Jahanifar, M., Tajeddin, N.Z., Koohbanani, N.A., Gooya, A., Rajpoot, N., 2018. Segmentation of skin lesions and their attributes using multi-scale convolutional neural networks and domain specific augmentations. arXiv preprint arXiv:1809.10243.

Japkowicz, N., Shah, M., 2011. Evaluating Learning Algorithms: A Classification Perspective. Cambridge University Press.

Jaworek-Korjakowska, J., Brodzicki, A., Cassidy, B., Kendrick, C., Yap, M.H., 2021. Interpretability of a deep learning based approach for the classification of skin lesions into main anatomic body sites. Cancers 13 (23), 6048.

Jayapriya, K., Jacob, I.J., 2020. Hybrid fully convolutional networks-based skin lesion segmentation and melanoma detection using deep feature. Int. J. Imaging Syst. Technol. 30 (2), 348–357.

Jensen, J.D., Elewski, B.E., 2015. The ABCDEF rule: combining the “ABCDE rule” and the “ugly duckling sign” in an effort to improve patient self-screening examinations. J. Clin. Aesthet. Dermatol. 8 (2), 15.

Ji, X., Henriques, J.F., Vedaldi, A., 2019. Invariant information clustering for unsupervised image classification and segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9865–9874.

Ji, Y., Zhang, R., Wang, H., Li, Z., Wu, L., Zhang, S., Luo, P., 2021. Multi-compound Transformer for accurate biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 326–336.

Jiang, Y., Cao, S., Tao, S., Zhang, H., 2020. Skin lesion segmentation based on multi-scale attention convolutional neural network. IEEE Access 8, 122811–122825.

Jiang, X., Jiang, J., Wang, B., Yu, J., Wang, J., 2022. SEACU-Net: Attentive ConvLSTM U-Net with squeeze-and-excitation layer for skin lesion segmentation. Comput. Methods Programs Biomed. 225, 107076.

Jiang, F., Zhou, F., Qin, J., Wang, T., Lei, B., 2019. Decision-augmented generative adversarial network for skin lesion segmentation. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). IEEE, pp. 447–450.

Jin, Q., Cui, H., Sun, C., Meng, Z., Su, R., 2021. Cascade knowledge diffusion network for skin lesion diagnosis and segmentation. Appl. Soft Comput. 99, 106881.

Kahn, R.L., 1942. Serology in syphilis control: Principles of sensitivity and specificity with an appendix for health officers and industrial physicians. Am. J. Clin. Path. 12 (8), 446. http://dx.doi.org/10.1093/ajcp/12.8.446d, arXiv:https://academic.oup.com/ajcp/article-pdf/12/8/446/24886161/ajcpath12-0446d.pdf.

Kamalakannan, A., Ganesan, S.S., Rajamanickam, G., 2019. Self-learning AI framework for skin lesion image segmentation and classification. Int. J. Comput. Sci. Inf. Technol. 11 (6), 29–38.

Kapoor, S., Narayanan, A., 2022. Leakage and the reproducibility crisis in ML-based science. arXiv preprint arXiv:2207.07048.

Karimi, D., Dou, H., Warfield, S.K., Gholipour, A., 2020. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. Med. Image Anal. 65, 101759.

Karras, T., Aila, T., Laine, S., Lehtinen, J., 2018. Progressive growing of GANs for improved quality, stability, and variation. In: International Conference on Learning Representations. pp. 1–26, URL: https://openreview.net/forum?id=Hk99zCeAb.

Kats, E., Goldberger, J., Greenspan, H., 2019. A soft STAPLE algorithm combined with anatomical knowledge. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 510–517.

Katsch, F., Rinner, C., Tschandl, P., 2022. Comparison of convolutional neural network architectures for robustness against common artefacts in dermatoscopic images. Dermatol. Pract. Concept. e2022126.

Katz, W.T., Merickel, M.B., 1989. Translation-invariant aorta segmentation from magnetic resonance images. In: Proceedings of the 1989 International Joint Conference on Neural Networks. pp. 327–333.

Kaul, C., Manandhar, S., Pears, N., 2019. Focusnet: an attention-based fully convolutional network for medical image segmentation. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). IEEE, pp. 455–458.

Kaul, C., Pears, N., Dai, H., Murray-Smith, R., Manandhar, S., 2021. Focusnet++: Attentive aggregated transformations for efficient and accurate medical image segmentation. In: 2021 IEEE 18th International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 1042–1046.

Kaur, R., GholamHosseini, H., Sinha, R., 2022a. Skin lesion segmentation using an improved framework of encoder-decoder based convolutional neural network. Int. J. Imaging Syst. Technol..

Kaur, R., GholamHosseini, H., Sinha, R., Lindén, M., 2022b. Automatic lesion segmentation using atrous convolutional deep neural networks in dermoscopic skin cancer images. BMC Med. Imaging 22 (1), 1–13.

Kawahara, J., Daneshvar, S., Argenziano, G., Hamarneh, G., 2019. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE J. Biomed. Health Inf. 23 (2), 538–546.

Kawahara, J., Hamarneh, G., 2018. Fully convolutional neural networks to detect clinical dermoscopic features. IEEE J. Biomed. Health Inf. 23 (2), 578–585.

Kaymak, R., Kaymak, C., Ucar, A., 2020. Skin lesion segmentation using fully convolutional networks: A comparative experimental study. Expert Syst. Appl. 161, 113742.

Kazeminia, S., Baur, C., Kuijper, A., van Ginneken, B., Navab, N., Albarqouni, S., Mukhopadhyay, A., 2020. GANs for medical image analysis. Artif. Intell. Med. 109, 101938.

Kent, A., Berry, M.M., Luehrs Jr., F.U., Perry, J.W., 1955. Machine literature searching: VIII. Operational criteria for designing information retrieval systems. Am. Doc. (Pre-1986) 6 (2), 93–101.

Khan, A.H., Awang Iskandar, D.N., Al-Asad, J.F., Mewada, H., Sherazi, M.A., 2022. Ensemble learning of deep learning and traditional machine learning approaches for skin lesion segmentation and classification. Concurr. Comput.: Pract. Exper. 34 (13), e6907.

Khan, A., Kim, H., Chua, L., 2021. PMED-Net: Pyramid based multi-scale encoder-decoder network for medical image segmentation. IEEE Access 9, 55988–55998.

Khouloud, S., Ahlem, M., Fadel, T., Amel, S., 2021. W-net and inception residual network for skin lesion segmentation and classification. Appl. Intell. 1–19.

Kim, M., Lee, B.-D., 2021. A simple generic method for effective boundary extraction in medical image segmentation. IEEE Access 9, 103875–103884.

Kinyanjui, N.M., Odonga, T., Cintas, C., Codella, N.C., Panda, R., Sattigeri, P., Varshney, K.R., 2020. Fairness of classifiers across skin tones in dermatology. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI. Springer, pp. 320–329.

Kittler, H., Pehamberger, H., Wolff, K., Binder, M., 2002. Diagnostic accuracy of dermoscopy. Lancet Oncol. 3 (3), 159–165.

Korotkov, K., Quintana, J., Campos, R., Jesús-Silva, A., Iglesias, P., Puig, S., Malvehy, J., Garcia, R., 2019. An improved skin lesion matching scheme in total body photography. IEEE J. Biomed. Health Inf. 23 (2), 586–598.

Kosgiker, G.M., Deshpande, A., Kauser, A., 2021. SegCaps: An efficient SegCaps network-based skin lesion segmentation in dermoscopic images. Int. J. Imaging Syst. Technol. 31 (2), 874–894.

Kovashka, A., Russakovsky, O., Fei-Fei, L., Grauman, K., 2016. Crowdsourcing in computer vision. Found. Trends Comput. Graph. Vis. 10 (3), 177–243.

Krahenbuhl, P., Koltun, V., 2011. Efficient inference in fully connected CRFs with Gaussian edge potentials. In: Proceedings of the 24th International Conference on Neural Information Processing Systems. pp. 109–117.

Kubat, M., Holte, R.C., Matwin, S., 1998. Machine learning for the detection of oil spills in satellite radar images. Mach. Learn. 30 (2–3), 195–215.

Kwon, Y., Won, J.-H., Kim, B.J., Paik, M.C., 2020. Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation. Comput. Statist. Data Anal. 142, 106816.

Lampert, T.A., Stumpf, A., Gancarski, P., 2016. An empirical study into annotator agreement, ground truth estimation, and algorithm evaluation. IEEE Trans. Image Process. 25 (6), 2557–2572.

Langerak, T.R., van der Heide, U.A., Kotte, A.N.T.J., Viergever, M.A., Van Vulpen, M., Pluim, J.P.W., 2010. Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE). IEEE Trans. Med. Imaging 29 (12), 2000–2008.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86 (11), 2278–2324.

Lee, T.K., McLean, D.I., Atkins, M.S., 2003. Irregularity index: A new border irregularity measure for cutaneous melanocytic lesions. Med. Image Anal. 7 (1), 47–64.

Lei, B., Xia, Z., Jiang, F., Jiang, X., Ge, Z., Xu, Y., Qin, J., Chen, S., Wang, T., Wang, S., 2020. Skin lesion segmentation via generative adversarial networks with dual discriminators. Med. Image Anal. 64, 101716.

Lemay, A., Gros, C., Naga Karthik, E., Cohen-Adad, J., 2022. Label fusion and training methods for reliable representation of inter-rater uncertainty. Mach. Learn. Biomed. Imaging 1, 1–27, URL: https://melba-journal.org/2022:031.

Li, Y., Chen, J., Zheng, Y., 2020b. A multi-task self-supervised learning framework for scopy images. In: 2020 IEEE 17th International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 2005–2009.

Li, Y., Esteva, A., Kuprel, B., Novoa, R., Ko, J., Thrun, S., 2017. Skin cancer detection and tracking using data synthesis and deep learning. In: AAAI Conference on Artificial Intelligence Joint Workshop on Health Intelligence. pp. 1–4.

Li, S., Gao, Z., He, X., 2021a. Superpixel-guided iterative learning from noisy labels for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 525–535.

Li, H., He, X., Zhou, F., Yu, Z., Ni, D., Chen, S., Wang, T., Lei, B., 2018a. Dense deconvolutional network for skin lesion segmentation. IEEE J. Biomed. Health Inf. 23 (2), 527–537.

Li, W., Raj, A.N.J., Tjahjadi, T., Zhuang, Z., 2021b. Digital hair removal by deep learning for skin lesion segmentation. Pattern Recognit. 117, 107994.

Li, Y., Shen, L., 2018. Skin lesion analysis towards melanoma detection using deep learning network. Sensors 18 (2), 556.

Li, R., Wagner, C., Chen, X., Auer, D., 2020a. A generic ensemble based deep convolutional neural network for semi-supervised medical image segmentation. In: 2020 IEEE 17th International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 1168–1172.

Li, Y., Xu, C., Han, J., An, Z., Wang, D., Ma, H., Liu, C., 2022. MHAU-Net: Skin lesion segmentation based on multi-scale hybrid residual attention network. Sensors 22 (22), 8701.

Li, X., Yu, L., Chen, H., Fu, C.-W., Xing, L., Heng, P.-A., 2021c. Transformation-consistent self-ensembling model for semi-supervised medical image segmentation. IEEE Trans. Neural Netw. Learn. Syst. 32 (2), 523–534.

Li, X., Yu, L., Fu, C.-W., Heng, P.-A., 2018b. 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. Springer, pp. 235–243.

Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2980–2988.

Lin, H., Upchurch, P., Bala, K., 2019. Block annotation: Better image annotation with sub-image decomposition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5290–5300.

Lin, A., Xu, J., Li, J., Lu, G., 2022. ConTrans: Improving Transformer with convolutional attention for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 297–307.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M., van Ginneken, B., Sanchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88.

Liu, C., Chen, L.C., Schroff, F., Adam, H., Hua, W., Yuille, A.L., Fei-Fei, L., 2019a. Auto-DeepLab: Hierarchical neural architecture search for semantic image segmentation. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 82–92.

Liu, X., Fan, W., Zhou, D., 2022b. Skin lesion segmentation via intensive atrous spatial Transformer. In: International Conference on Wireless Algorithms, Systems, and Applications. Springer, pp. 15–26.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021b. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022.

Liu, L., Mou, L., Zhu, X.X., Mandal, M., 2019b. Skin lesion segmentation based on improved U-Net. In: 2019 IEEE Canadian Conference of Electrical and Computer Engineering. CCECE, IEEE, pp. 1–4.

Liu, L., Mou, L., Zhu, X.X., Mandal, M., 2020. Automatic skin lesion classification based on mid-level feature learning. Comput. Med. Imaging Graph. 84, 101765.

Liu, L., Tsui, Y.Y., Mandal, M., 2021a. Skin lesion segmentation using deep learning with auxiliary task. J. Imaging 7 (4), 67.

Liu, Q., Wang, J., Zuo, M., Cao, W., Zheng, J., Zhao, H., Xie, J., 2022a. NCRNet: Neighborhood Context Refinement Network for skin lesion segmentation. Comput. Biol. Med. 146, 105545.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440.

Lui, H., et al., 2009. DermWeb. Department of Dermatology and Skin Science, the University of British Columbia, http://www.dermweb.com/. [Online. Accessed January 26, 2022].

Luque, A., Carrasco, A., Martin, A., de las Heras, A., 2020. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognit. 91, 216–231.

Ma, J., Chen, J., Ng, M., Huang, R., Li, Y., Li, C., Yang, X., Martel, A.L., 2021. Loss odyssey in medical image segmentation. Med. Image Anal. 71, 102035.

Mahbod, A., Tschandl, P., Langs, G., Ecker, R., Ellinger, I., 2020. The effects of skin lesion segmentation on the performance of dermatoscopic image classification. Comput. Methods Programs Biomed. 197, 105725.

Maier-Hein, L., Mersmann, S., Kondermann, D., Bodenstedt, S., Sanchez, A., Stock, C., Kenngott, H.G., Eisenmann, M., Speidel, S., 2014. Can masses of non-experts train highly accurate image classifiers? In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 438–445.

Marchetti, M.A., Codella, N.C.F., Dusza, S.W., Gutman, D.A., Helba, B., Kalloo, A., Mishra, N., Carrera, C., Celebi, M.E., DeFazio, J.L., Jaimes, N., Marghoob, A.A., Quigley, E., Scope, A., Yelamos, O., Halpern, A.C., 2018. Results of the 2016 international skin imaging collaboration international symposium on biomedical imaging challenge: Comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. J. Am. Acad. Dermatol. 78 (2), 270–277.

Maron, R.C., Hekler, A., Krieghoff-Henning, E., Schmitt, M., Schlager, J.G., Utikal, J.S., Brinker, T.J., 2021a. Reducing the impact of confounding factors on skin cancer classification via image segmentation: Technical model study. J. Med. Internet Res. 23 (3), e21695.

Maron, R.C., Schlager, J.G., Haggenmüller, S., von Kalle, C., Utikal, J.S., Meier, F., Gellrich, F.F., Hobelsberger, S., Hauschild, A., French, L., Heinzerling, L., Schlaak, M., Ghoreschi, K., Hilke, F.J., Poch, G., Heppt, M.V., Berking, C., Haferkamp, S., Sondermann, W., Schadendorf, D., Schilling, B., Goebeler, M., Krieghoff-Henning, E., Hekler, A., Fröhling, S., Lipka, D.B., Kather, J.N., Brinker, T.J., 2021b. A benchmark for neural network robustness in skin cancer classification. Eur. J. Cancer 155, 191–199.

Matthews, B.W., 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta 405 (2), 442–451.

Mendonca, T.F., Ferreira, P.M., Marcal, A.R.S., Barata, C., Marques, J.S., Rocha, J., Rozeira, J., 2015. PH^2—A dermoscopic image database for research and benchmarking. In: Celebi, M.E., Mendonca, T., Marques, J.S. (Eds.), Dermoscopy Image Analysis. CRC Press, pp. 419–439.

Mendonca, T., Ferreira, P.M., Marques, J.S., Marcal, A.R.S., Rozeira, J., 2013. 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. pp. 5437–5440.

Menzies, S.W., Crotty, K.A., Ingwar, C., McCarthy, W.H., 2003. An Atlas of Surface Microscopy of Pigmented Skin Lesions: Dermoscopy, second ed. McGraw-Hill.

Miller, G.A., Nicely, P.E., 1955. An analysis of perceptual confusions among some English consonants. J. Acoust. Soc. Am. 27 (2), 338–352.

Mirikharaji, Z., Abhishek, K., Izadi, S., Hamarneh, G., 2021. D-LEMA: Deep learning ensembles from multiple annotations-application to skin lesion segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1837–1846.

Mirikharaji, Z., Hamarneh, G., 2018. Star shape prior in fully convolutional networks for skin lesion segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 737–745.

Mirikharaji, Z., Izadi, S., Kawahara, J., Hamarneh, G., 2018. Deep auto-context fully convolutional neural network for skin lesion segmentation. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE, pp. 877–880.

Mirikharaji, Z., Yan, Y., Hamarneh, G., 2019. Learning to segment skin lesions from noisy annotations. In: Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data. Springer, pp. 207–215.

Mirzaalian, H., Lee, T.K., Hamarneh, G., 2016. Skin lesion tracking using structured graphical models. Med. Image Anal. 27, 84–92.

Mishra, R., Daescu, O., 2017. Deep learning for skin lesion segmentation. In: 2017 IEEE International Conference on Bioinformatics and Biomedicine. BIBM, IEEE, pp. 1189–1194.

Nachbar, F., Stolz, W., Merkle, T., Cognetta, A.B., Vogt, T., Landthaler, M., Bilek, P., Braun-Falco, O., Plewig, G., 1994. The ABCD rule of dermatoscopy: High prospective value in the diagnosis of doubtful melanocytic skin lesions. J. Am. Acad. Dermatol. 30 (4), 551–559.

Nasr-Esfahani, E., Rafiei, S., Jafari, M.H., Karimi, N., Wrobel, J.S., Samavi, S., Soroushmehr, S.R., 2019. Dense pooling layers in fully convolutional network for skin lesion segmentation. Comput. Med. Imaging Graph. 78, 101658.

Nathan, S., Kansal, P., 2020. Lesion net–skin lesion segmentation using coordinate convolution and deep residual units. arXiv preprint arXiv:2012.14249.

Navarro, F., Escudero-Vinolo, M., Bescós, J., 2018. Accurate segmentation and registration of skin lesion images to evaluate lesion change. IEEE J. Biomed. Health Inf. 23 (2), 501–508.

Ning, F., Delhomme, D., LeCun, Y., Piano, F., Bottou, L., Barbano, P.E., 2005. Toward automatic phenotyping of developing embryos from videos. IEEE Trans. Image Process. 14 (9), 1360–1371.

Norton, K.A., Iyatomi, H., Celebi, M.E., Ishizaki, S., Sawada, M., Suzaki, R., Kobayashi, K., Tanaka, M., Ogawa, K., 2012. Three-phase general border detection method for dermoscopy images using non-uniform illumination correction. Skin Res. Technol. 18 (3), 290–300.

Nosrati, M.S., Hamarneh, G., 2016. Incorporating prior knowledge in medical image segmentation: a survey. arXiv preprint arXiv:1607.01092.

Oakley, A., et al., 1995. DermNet New Zealand trust. https://dermnetnz.org/. [Online. Accessed January 26, 2022].

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al., 2018. Attention U-Net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999.

Öztürk, Ş., Özkaya, U., 2020. Skin lesion segmentation with improved convolutional neural network. J. Digit. Imaging 33, 958–970.

Pacheco, A.G., Lima, G.R., Salomão, A.S., Krohling, B., Biral, I.P., de Angelo, G.G., Alves Jr., F.C., Esgario, J.G., Simora, A.C., Castro, P.B., et al., 2020. PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. Data Brief 32, 106221.

Pakzad, A., Abhishek, K., Hamarneh, G., 2023. CIRCLe: Color invariant representation learning for unbiased classification of skin lesions. In: Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV. Springer, pp. 203–219.

Papadopoulos, D.P., Uijlings, J.R., Keller, F., Ferrari, V., 2017. Extreme clicking for efficient object annotation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4930–4939.

Papandreou, G., Chen, L.C., Murphy, K.P., Yuille, A.L., 2015. Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1742–1750.

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., Tran, D., 2018. Image transformer. In: International Conference on Machine Learning. PMLR, pp. 4055–4064.

Pearson, K., 1904. On the Theory of Contingency and its Relation to Association and Normal Correlation, Vol. 1. Dulau and Company, London, UK.

Peng, B., Li, T., 2013. A probabilistic measure for quantitative evaluation of image segmentation. IEEE Signal Process. Lett. 20 (7), 689–692.

Peng, Y., Wang, N., Wang, Y., Wang, M., 2019. Segmentation of dermoscopy image using adversarial networks. Multimedia Tools Appl. 78 (8), 10965–10981.

Peng, B., Wang, X., Yang, Y., 2016. Region based exemplar references for image segmentation evaluation. IEEE Signal Process. Lett. 23 (4), 459–462.

Peng, B., Zhang, L., Mou, X., Yang, M.H., 2017a. Evaluation of segmentation quality via adaptive composition of reference segmentations. IEEE Trans. Pattern Anal. Mach. Intell. 39 (10), 1929–1941.

Peng, C., Zhang, X., Yu, G., Luo, G., Sun, J., 2017b. Large kernel matters–improve semantic segmentation by global convolutional network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4353–4361.

Perez, F., Vasconcelos, C., Avila, S., Valle, E., 2018. Data augmentation for skin lesion analysis. In: Proceedings of the Third ISIC Workshop on Skin Image Analysis. pp. 303–311.

Peserico, E., Silletti, A., 2010. Is (N)PRI suitable for evaluating automated segmentation of cutaneous lesions? Pattern Recognit. Lett. 31 (16), 2464–2467.

Pinheiro, P.H., Collobert, R., 2014. Recurrent convolutional neural networks for scene labeling. In: 31st International Conference on Machine Learning. ICML, PMLR, pp. 82–90.

Pollastri, F., Bolelli, F., Palacios, R.P., Grana, C., 2018. Improving skin lesion segmentation with generative adversarial networks. In: 2018 IEEE 31st International Symposium on Computer-Based Medical Systems. CBMS, IEEE, pp. 442–443.

Pollastri, F., Bolelli, F., Paredes, R., Grana, C., 2020. Augmenting data with GANs to segment melanoma skin lesions. Multimedia Tools Appl. 79 (21), 15575–15592.

Poudel, S., Lee, S.-W., 2021. Deep multi-scale attentional features for medical image segmentation. Appl. Soft Comput. 109, 107445.

Pour, M.P., Seker, H., 2020. Transform domain representation-driven convolutional neural networks for skin lesion segmentation. Expert Syst. Appl. 144, 113129.

Qiu, Y., Cai, J., Qin, X., Zhang, J., 2020. Inferring skin lesion deep convolutional neural networks. IEEE Access 8, 144246–144258.

Rajchl, M., Lee, M.C., Schrans, F., Davidson, A., Passerat-Palmbach, J., Tarroni, G., Alansary, A., Oktay, O., Kainz, B., Rueckert, D., 2016. Learning under distributed weak supervision. arXiv preprint arXiv:1606.01100.

Ramachandram, D., DeVries, T., 2017. LesionSeg: semantic segmentation of skin lesions using deep convolutional neural network. arXiv preprint arXiv:1703.03372.

Ramachandram, D., Taylor, G.W., 2017. Skin lesion segmentation using deep hypercolumn descriptors. J. Comput. Vis. Imaging Syst. 3 (1).

Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., Shlens, J., 2019. Stand-alone self-attention in vision models. Adv. Neural Inf. Process. Syst. 32.

Ramadan, R., Aly, S., Abdel-Atty, M., 2022. Color-invariant skin lesion semantic segmentation based on modified U-Net deep convolutional neural network. Health Inf. Sci. Syst. 10 (1), 1–12.

Ramani, D.R., Ranjani, S.S., 2019. 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. pp. 81–101.

Rand, W.M., 1971. Objective criteria for the evaluation of clustering methods. J. Amer. Statist. Assoc. 66 (336), 846–850.

Ranftl, R., Bochkovskiy, A., Koltun, V., 2021. Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12179–12188.

Redekop, E., Chernyavskiy, A., 2021. Uncertainty-based method for improving poorly labeled segmentation datasets. In: 2021 IEEE 18th International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 1831–1835.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 779–788.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. Adv. Neural Inf. Process. Syst. 28.

Ren, Y., Yu, L., Tian, S., Cheng, J., Guo, Z., Zhang, Y., 2021. Serial attention network for skin lesion segmentation. J. Ambient Intell. Humaniz. Comput. 1–12.

Renard, F., Guedria, S., Palma, N.D., Vuillerme, N., 2020. Variability and reproducibility in deep learning for medical image segmentation. Sci. Rep. 10 (1), 1–16.

Ribeiro, V., Avila, S., Valle, E., 2020. 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. pp. 738–739.

Rohlfing, T., Maurer, C.R., 2006. Shape-based averaging. IEEE Trans. Image Process. 16 (1), 153–161.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241.

Ross-Howe, S., Tizhoosh, H.R., 2018. 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, pp. 1–8.

Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., Combalia, M., Dusza, S., Guitera, P., Gutman, D., Halpern, A., Helba, B., Kittler, H., Kose, K., Langer, S., Lioprys, K., Malvehy, J., Musthaq, S., Nanda, J., Reiter, O., Shih, G., Stratigos, A., Tschandl, P., Weber, J., Soyer, H.P., 2021. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. Sci. Data 8, 34.

Roth, H.R., Yang, D., Xu, Z., Wang, X., Xu, D., 2021. Going to extremes: Weakly supervised medical image segmentation. Mach. Learn. Knowl. Extr. 3 (2), 507–524.

Rother, C., Kolmogorov, V., Blake, A., 2004. "GrabCut" interactive foreground extraction using iterated graph cuts. ACM Trans. Graph. 23 (3), 309–314.

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

Saba, T., Khan, M.A., Rehman, A., Marie-Sainte, S.L., 2019. Region extraction and classification of skin cancer: A heterogeneous framework of deep CNN features fusion and reduction. J. Med. Syst. 43 (9), 289.

Sachin, T.S., Sowmya, V., Soman, K., 2021. Performance analysis of deep learning models for biomedical image segmentation. In: Deep Learning for Biomedical Applications. CRC Press, pp. 83–100.

Sagi, O., Rokach, L., 2018. Ensemble learning: A survey. Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. 8 (4), e1249.

Saha, A., Prasad, P., Thabit, A., 2020. Leveraging adaptive color augmentation in convolutional neural networks for deep skin lesion segmentation. In: 2020 IEEE 17th International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 2014–2017.

Şahin, N., Alpaslan, N., Hanbay, D., 2021. Robust optimization of SegNet hyperparameters for skin lesion segmentation. Multimedia Tools Appl. 1–21.

Saini, S., Gupta, D., Tiwari, A.K., 2019. Detector-SegMentor network for skin lesion localization and segmentation. In: National Conference on Computer Vision, Pattern Recognition, Image Processing, and Graphics. Springer, pp. 589–599.

Saini, S., Jeon, Y.S., Feng, M., 2021. B-SegNet: branched-SegMentor network for skin lesion segmentation. In: Proceedings of the Conference on Health, Inference, and Learning. pp. 214–221.

Sarker, M., Kamal, M., Rashwan, H.A., Abdel-Nasser, M., Singh, V.K., Banu, S.F., Akram, F., Chowdhury, F.U., Choudhury, K.A., Chambon, S., et al., 2019. Mobile-GAN: Skin lesion segmentation using a lightweight generative adversarial network. arXiv preprint arXiv:1907.00856.

Sarker, M.M.K., Rashwan, H.A., Akram, F., Banu, S.F., Saleh, A., Singh, V.K., Chowdhury, F.U., Abdulwahab, S., Romani, S., Radeva, P., et al., 2018. SLSDeep: Skin lesion segmentation based on dilated residual and pyramid pooling networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 21–29.

Sarker, M.M.K., Rashwan, H.A., Akram, F., Singh, V.K., Banu, S.F., Chowdhury, F.U., Choudhury, K.A., Chambon, S., Radeva, P., Puig, D., et al., 2021. SLSNet: Skin lesion segmentation using a lightweight generative adversarial network. Expert Syst. Appl. 183, 115433.

Schaefer, G., Rajab, M.I., Celebi, M.E., Iyatomi, H., 2011. Colour and contrast enhancement for improved skin lesion segmentation. Comput. Med. Imaging Graph. 35 (2), 99–104.

Shahin, A.H., Amer, K., Elattar, M.A., 2019. 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). IEEE, pp. 451–454.

Shamshad, F., Khan, S., Zamir, S.W., Khan, M.H., Hayat, M., Khan, F.S., Fu, H., 2022. Transformers in medical imaging: A survey. arXiv preprint arXiv:2201.09873.

Shamsolmoali, P., Zareapoor, M., Granger, E., Zhou, H., Wang, R., Celebi, M.E., Yang, J., 2021. Image synthesis with adversarial networks: A comprehensive survey and case studies. Inf. Fusion 72, 126–146.

Sharma, M., Saha, O., Sriraman, A., Hebbalaguppe, R., Vig, L., Karande, S., 2017. Crowdsourcing for chromosome segmentation and deep classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 786–793.

Shimizu, K., Iyatomi, H., Celebi, M.E., Norton, K.A., Tanaka, M., 2015. Four-class classification of skin lesions with task decomposition strategy. IEEE Trans. Biomed. Eng. 62 (1), 274–283.

Shorten, C., Khoshgoftaar, T.M., 2019. A survey on image data augmentation for deep learning. J. Big Data 6 (1), 60.

Siegel, R.L., Miller, K.D., Wagle, N.S., Jemal, A., 2023. Cancer statistics, 2023. CA: Cancer J. Clin. 73 (1), 17–48. http://dx.doi.org/10.3322/caac.21763.

Silveira, M., Nascimento, J.C., Marques, J.S., Marcal, A.R.S., Mendonca, T., Yamauchi, S., Maeda, J., Rozeira, J., 2009. Comparison of segmentation methods for melanoma diagnosis in dermoscopy images. IEEE J. Sel. Top. Sign. Proces. 3 (1), 35–45.

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

Singh, V.K., Abdel-Nasser, M., Rashwan, H.A., Akram, F., Pandey, N., Lalande, A., Presles, B., Romani, S., Puig, D., 2019. FCA-Net: Adversarial learning for skin lesion segmentation based on multi-scale features and factorized channel attention. IEEE Access 7, 130552–130565.

Singh, L., Janghel, R.R., Sahu, S.P., 2023. An empirical review on evaluating the impact of image segmentation on the classification performance for skin lesion detection. IETE Tech. Rev. 40 (2), 190–201.

Sinha, A., Kawahara, J., Pakzad, A., Abhishek, K., Ruthven, M., Ghorbel, E., Kacem, A., Aouada, D., Hamarneh, G., 2023. DermSynth3D: Synthesis of in-the-wild annotated dermatology images. arXiv preprint arXiv:2305.12621.

Smyth, P., Fayyad, U.M., Burl, M.C., Perona, P., Baldi, P., 1995. Inferring ground truth from subjective labelling of venus images. In: Advances in Neural Information Processing Systems. pp. 1085–1092.

Soenksen, L.R., Kassis, T., Conover, S.T., Marti-Fuster, B., Birkenfeld, J.S., Tucker-Schwartz, J., Naseem, A., Stavert, R.R., Kim, C.C., Senna, M.M., Avilés-Izquierdo, J., Collins, J.J., Barzilay, R., Gray, M.L., 2021. Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images. Sci. Transl. Med. 13 (581), eabb3652.

Song, L., Lin, J., Wang, Z.J., Wang, H., 2019. Dense-residual attention network for skin lesion segmentation. In: International Workshop on Machine Learning in Medical Imaging. Springer, pp. 319–327.

Sørensen, T.A., 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. Biol. Skar. 5 (4), 1–34.

Soudani, A., Barhoumi, W., 2019. An image-based segmentation recommender using crowdsourcing and transfer learning for skin lesion extraction. Expert Syst. Appl. 118, 400–410.

Strudel, R., Garcia, R., Laptev, I., Schmid, C., 2021. Segmenter: Transformer for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7262–7272.

Sun, C., Shrivastava, A., Singh, S., Gupta, A., 2017. Revisiting unreasonable effectiveness of data in deep learning era. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 843–852.

Sun, X., Yang, J., Sun, M., Wang, K., 2016. A benchmark for automatic visual classification of clinical skin disease images. In: European Conference on Computer Vision. Springer, pp. 206–222.

Taghanaki, S.A., Abhishek, K., Hamarneh, G., 2019. Improved inference via deep input transfer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 819–827.

Taha, A.A., Hanbury, A., 2015. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. BMC Med. Imaging 15 (1), 29.

Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J.N., Wu, Z., Ding, X., 2020. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. Med. Image Anal. 63, 101693.

Tan, M., Chen, B., Pang, R., Vasudevan, V., Sandler, M., Howard, A., Le, Q.V., 2019a. MnasNet: Platform-aware neural architecture search for mobile. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2820–2828.

Tan, M., Le, Q., 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the International Conference on Machine Learning. pp. 6105–6114.

Tan, T.Y., Zhang, L., Lim, C.P., Fielding, B., Yu, Y., Anderson, E., 2019b. Evolving ensemble models for image segmentation using enhanced particle swarm optimization. IEEE Access 7, 34004–34019.

Tang, P., Liang, Q., Yan, X., Xiang, S., Sun, W., Zhang, D., Coppola, G., 2019a. Efficient skin lesion segmentation using separable-Unet with stochastic weight averaging. Comput. Methods Programs Biomed. 178, 289–301.

Tang, X., Peng, J., Zhong, B., Li, J., Yan, Z., 2021b. Introducing frequency representation into convolution neural networks for medical image segmentation via twin-Kernel Fourier convolution. Comput. Methods Programs Biomed. 205, 106110.

Tang, P., Yan, X., Liang, Q., Zhang, D., 2021a. AFLN-DGCL: Adaptive feature learning network with difficulty-guided curriculum learning for skin lesion segmentation. Appl. Soft Comput. 110, 107656.

Tang, Y., Yang, F., Yuan, S., et al., 2019b. A multi-stage framework with context information fusion structure for skin lesion segmentation. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). IEEE, pp. 1407–1410.

Tao, S., Jiang, Y., Cao, S., Wu, C., Ma, Z., 2021. Attention-guided network with densely connected convolution for skin lesion segmentation. Sensors 21 (10), 3462.

Tong, X., Wei, J., Sun, B., Su, S., Zuo, Z., Wu, P., 2021. ASCU-Net: Attention gate, spatial and channel attention U-Net for skin lesion segmentation. Diagnostics 11 (3), 501.

Torralba, A., Efros, A.A., 2011. Unbiased look at dataset bias. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1521–1528.

Tran, H., Chen, K., Lim, A.C., Jabbour, J., Shumack, S., 2005. Assessing diagnostic skill in dermatology: a comparison between general practitioners and dermatologists. Australas. J. Dermatol. 46 (4), 230–234.

Tran, T.-T., Pham, V.-T., 2022. Fully convolutional neural network with attention gate and fuzzy active contour model for skin lesion segmentation. Multimedia Tools Appl. 81 (10), 13979–13999.

Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., Janda, M., Lallas, A., Longo, C., Malvehy, J., Paoli, J., Puig, S., Rosendahl, C., Soyer, H.P., Zalaudek, I., Kittler, H., 2020. Human–computer collaboration for skin cancer recognition. Nat. Med. 26 (8), 1229–1234. http://dx.doi.org/10.1038/s41591-020-0942-0.

Tschandl, P., Rosendahl, C., Kittler, H., 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data 180161.

Tschandl, P., Sinz, C., Kittler, H., 2019. Domain-specific classification-pretrained fully convolutional network encoders for skin lesion segmentation. Comput. Biol. Med. 104, 111–116.

Tu, W., Liu, X., Hu, W., Pan, Z., 2019. Dense-residual network with adversarial learning for skin lesion segmentation. IEEE Access 7, 77037–77051.

Unnikrishnan, R., Pantofaru, C., Hebert, M., 2007. Toward objective evaluation of image segmentation algorithms. IEEE Trans. Pattern Anal. Mach. Intell. 29 (6), 929–944.

Ünver, H.M., Ayan, E., 2019. Skin lesion segmentation in dermoscopic images with combination of YOLO and GrabCut algorithm. Diagnostics 9 (3), 72.

Usatine, R.P., Madden, B.D., 2013. Interactive Dermatology Atlas. Department of Dermatology and Cutaneous Surgery, University of Texas, https://www.dermatlas.net/ [Accessed January 26, 2022].

Valanarasu, J.M.J., Patel, V.M., 2022. UNeXt: MLP-based rapid medical image segmentation network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 23–33.

Valle, E., Fornaciali, M., Menegola, A., Tavares, J., Bittencourt, F.V., Li, L.T., Avila, S., 2020. Data, depth, and design: Learning reliable models for skin lesion analysis. Neurocomputing 383, 303–313.

van Rijsbergen, C.J., 1979. Information Retrieval, second ed. Butterworth–Heinemann.

Vandewalle, P., 2012. Code sharing is associated with research impact in image processing. Comput. Sci. Eng. 14 (4), 42–47.

Vanker, A.D., Van Stoecker, W., 1984. An expert diagnostic program for dermatology. Comput. Biomed. Res. 17 (3), 241–247.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Venkatesh, G., Naresh, Y., Little, S., O'Connor, N.E., 2018. 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. Springer, pp. 277–284.

Vesal, S., Patil, S.M., Ravikumar, N., Maier, A.K., 2018a. 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. Springer, pp. 285–293.

Vesal, S., Ravikumar, N., Maier, A., 2018b. SkinNet: A deep learning framework for skin lesion segmentation. In: 2018 IEEE Nuclear Science Symposium and Medical Imaging Conference Proceedings (NSS/MIC). IEEE, pp. 1–3.

ViDIR Dataverse, 2020. HAM10000 binary lesion segmentations. http://dx.doi.org/10.7910/DVN/DBW86T, [Online. Accessed January 9, 2023].

Wang, R., Chen, S., Fan, J., Li, Y., 2020a. Cascaded context enhancement for automated skin lesion segmentation. arXiv preprint arXiv:2004.08107.

Wang, X., Ding, H., Jiang, X., 2019b. Dermoscopic image segmentation through the enhanced high-level parsing and class weighted loss. In: 2019 IEEE International Conference on Image Processing. ICIP, IEEE, pp. 245–249.

Wang, X., Jiang, X., Ding, H., Liu, J., 2019c. Bi-directional dermoscopic feature learning and multi-scale consistent decision fusion for skin lesion segmentation. IEEE Trans. Image Process. 29, 3039–3051.

Wang, X., Jiang, X., Ding, H., Zhao, Y., Liu, J., 2021b. Knowledge-aware deep framework for collaborative skin lesion segmentation and melanoma recognition. Pattern Recognit. 120, 108075.

Wang, T., Lan, J., Han, Z., Hu, Z., Huang, Y., Deng, Y., Zhang, H., Wang, J., Chen, M., Jiang, H., et al., 2022b. O-Net: a novel framework with deep fusion of CNN and Transformer for simultaneous segmentation and classification. Front. Neurosci. 16.

Wang, J., Li, B., Guo, X., Huang, J., Song, M., Wei, M., 2022a. CTCNet: A bi-directional cascaded segmentation network combining Transformers with CNNs for skin lesions. In: Chinese Conference on Pattern Recognition and Computer Vision. PRCV, Springer, pp. 215–226.

Wang, M., Liu, B., Foroosh, H., 2017. Factorized convolutional neural networks. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 545–553.

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., Catanzaro, B., 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8798–8807.

Wang, Z., Lyu, J., Luo, W., Tang, X., 2022d. Superpixel inpainting for self-supervised skin lesion segmentation from dermoscopic images. In: 2022 IEEE 19th International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 1–4.

Wang, Y., Wang, S., 2022. Skin lesion segmentation with attention-based SC-Conv U-Net and feature map distortion. Signal Image Video Process. 1–9.

Wang, H., Wang, G., Sheng, Z., Zhang, S., 2019a. Automated segmentation of skin lesion based on pyramid attention network. In: International Workshop on Machine Learning in Medical Imaging. Springer, pp. 435–443.

Wang, Y., Wei, Y., Qian, X., Zhu, L., Yang, Y., 2020b. DONet: Dual objective networks for skin lesion segmentation. arXiv preprint arXiv:2008.08278.

Wang, J., Wei, L., Wang, L., Zhou, Q., Zhu, L., Qin, J., 2021a. Boundary-aware Transformers for skin lesion segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 206–216.

Wang, Y., Xu, Z., Tian, J., Luo, J., Shi, Z., Zhang, Y., Fan, J., He, Z., 2022c. Cross-domain few-shot learning for rare-disease skin lesion segmentation. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 1086–1090.

Warfield, S.K., Zou, K.H., Wells, W.M., 2004. Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. IEEE Trans. Med. Imaging 23 (7), 903–921.

Wei, Z., Song, H., Chen, L., Li, Q., Han, G., 2019. Attention-based DenseUnet network with adversarial training for skin lesion segmentation. IEEE Access 7, 136616–136629.

Weng, Y., Zhou, T., Li, Y., Qiu, X., 2019. NAS-Unet: Neural architecture search for medical image segmentation. IEEE Access 7, 44247–44257.

Wibowo, A., Purnama, S.R., Wirawan, P.W., Rasyidi, H., 2021. Lightweight encoder-decoder model for automatic skin lesion segmentation. Inform. Med. Unlocked 100640.

Wu, H., Chen, S., Chen, G., Wang, W., Lei, B., Wen, Z., 2022a. FAT-Net: Feature adaptive Transformers for automated skin lesion segmentation. Med. Image Anal. 76, 102327.

Wu, J., Fang, H., Shang, F., Yang, D., Wang, Z., Gao, J., Yang, Y., Xu, Y., 2022b. SeATrans: Learning segmentation-assisted diagnosis model via Transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 677–687.

Wu, H., Pan, J., Li, Z., Wen, Z., Qin, J., 2020. Automated skin lesion segmentation via an adaptive dual attention module. IEEE Trans. Med. Imaging 40 (1), 357–370.

Wu, Y., Zeng, D., Xu, X., Shi, Y., Hu, J., 2022c. FairPrune: Achieving fairness through pruning for dermatological disease diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 743–753.

Xie, Z., Tu, E., Zheng, H., Gu, Y., Yang, J., 2021. Semi-supervised skin lesion segmentation with learning model confidence. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 1135–1139.

Xie, F., Yang, J., Liu, J., Jiang, Z., Zheng, Y., Wang, Y., 2020a. Skin lesion segmentation using high-resolution convolutional neural network. Comput. Methods Programs Biomed. 186, 105241.

Xie, Y., Zhang, J., Xia, Y., Shen, C., 2020b. A mutual bootstrapping model for automated skin lesion segmentation and classification. IEEE Trans. Med. Imaging 39 (7), 2482–2493.

Xu, R., Wang, C., Xu, S., Meng, W., Zhang, X., 2021. DC-Net: Dual context network for 2D medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 503–513.

Xue, Y., Xu, T., Huang, X., 2018. Adversarial learning with multi-scale loss for skin lesion segmentation. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE, pp. 859–863.

Yan, Y., Kawahara, J., Hamarneh, G., 2019. Melanoma recognition via visual attention. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 793–804.

Yang, X., Li, H., Wang, L., Yeo, S.Y., Su, Y., Zeng, Z., 2018. 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, IEEE, pp. 1263–1266.

Yang, C.-H., Ren, J.-H., Huang, H.-C., Chuang, L.-Y., Chang, P.-Y., 2021. Deep hybrid convolutional neural network for segmentation of melanoma skin lesion. Comput. Intell. Neurosci. 2021.

Yerushalmy, J., 1947. Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. Public Health Rep. (1896-1970) 62 (40), 1432–1449.

Yi, X., Walia, E., Babyn, P., 2019. Generative adversarial network in medical imaging: A review. Med. Image Anal. 58, 101552.

Yu, L., Chen, H., Dou, Q., Qin, J., Heng, P.A., 2017a. Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE Trans. Med. Imaging 36 (4), 994–1004.

Yu, Y., Gong, Z., Zhong, P., Shan, J., 2017b. Unsupervised representation learning with deep convolutional neural network for remote sensing images. In: International Conference on Image and Graphics. pp. 97–108.

Yu, F., Koltun, V., 2016. Multi-scale context aggregation by dilated convolutions. In: International Conference on Learning Representations.

Yu, B., Yu, L., Tian, S., Wu, W., Zhang, D., Kang, X., 2022. mCA-Net: modified comprehensive attention convolutional neural network for skin lesion segmentation. Comput. Methods Biomech. Biomed. Eng.: Imaging Vis. 10 (1), 85–95.

Yuan, Y., Chao, M., Lo, Y.-C., 2017. Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance. IEEE Trans. Med. Imaging 36 (9), 1876–1886.

Yuan, Y., Lo, Y.C., 2019. Improving dermoscopic image segmentation with enhanced convolutional-deconvolutional networks. IEEE J. Biomed. Health Inf. 23 (2), 519–526.

Zafar, K., Gilani, S.O., Waris, A., Ahmed, A., Jamil, M., Khan, M.N., Sohail Kashif, A., 2020. Skin lesion segmentation from dermoscopic images using convolutional neural network. Sensors 20 (6), 1601.

Zeng, G., Zheng, G., 2018. Multi-scale fully convolutional DenseNets for automated skin lesion segmentation in dermoscopy images. In: International Conference Image Analysis and Recognition. Springer, pp. 513–521.

Zhang, Y., Chen, Z., Yu, H., Yao, X., Li, H., 2022a. Feature fusion for segmentation and classification of skin lesions. In: 2022 IEEE 19th International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 1–5.

Zhang, H., Fritts, J.E., Goldman, S.A., 2008. Image segmentation evaluation: A survey of unsupervised methods. Comput. Vis. Image Underst. 110 (2), 260–280.

Zhang, Y., Liu, H., Hu, Q., 2021b. TransFuse: Fusing Transformers and CNNs for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 14–24.

Zhang, R., Liu, S., Yu, Y., Li, G., 2021a. Self-supervised correction learning for semi-supervised biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 134–144.

Zhang, J., Petitjean, C., Ainouz, S., 2020a. Kappa loss for skin lesion segmentation in fully convolutional network. In: 2020 IEEE 17th International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 2001–2004.

Zhang, G., Shen, X., Chen, S., Liang, L., Luo, Y., Yu, J., Lu, J., 2019a. DSM: A deep supervised multi-scale network learning for skin cancer segmentation. IEEE Access 7, 140936–140945.

Zhang, L., Tanno, R., Bronik, K., Jin, C., Nachev, P., Barkhof, F., Ciccarelli, O., Alexander, D.C., 2020b. Learning to segment when experts disagree. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 179–190.

Zhang, Z., Tian, C., Gao, X., Wang, C., Feng, X., Bai, H.X., Jiao, Z., 2022b. Dynamic prototypical feature representation learning framework for semi-supervised skin lesion segmentation. Neurocomputing 507, 369–382.

Zhang, Y., Yang, Q., 2022. A survey on multi-task learning. IEEE Trans. Knowl. Data Eng..

Zhang, L., Yang, G., Ye, X., 2019b. Automatic skin lesion segmentation by coupling deep fully convolutional networks and shallow network with textons. J. Med. Imaging 6 (2), 024001.

Zhang, X., Zhou, X., Lin, M., Sun, J., 2018. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6848–6856.

Zhao, H., Jia, J., Koltun, V., 2020. Exploring self-attention for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10076–10085.

Zhao, M., Kawahara, J., Abhishek, K., Shamanian, S., Hamarneh, G., 2022a. Skin3D: Detection and longitudinal tracking of pigmented skin lesions in 3D total-body textured meshes. Med. Image Anal. 77, 102329.

Zhao, Z., Lu, W., Zeng, Z., Xu, K., Veeravalli, B., Guan, C., 2022b. Self-supervised assisted active learning for skin lesion segmentation. In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, pp. 5043–5046. http://dx.doi.org/10.1109/embc48229.2022.9871734.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2881–2890.

Zhao, C., Shuai, R., Ma, L., Liu, W., Wu, M., 2021. Segmentation of dermoscopy images based on deformable 3D convolution and ResU-NeXt++. Med. Biol. Eng. Comput. 59 (9), 1815–1832.

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al., 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6881–6890.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2921–2929.

Zhu, Q., 2020. On the performance of matthews correlation coefficient (MCC) for imbalanced dataset. Pattern Recognit. Lett. 136, 71–80.

Zhu, L., Feng, S., Zhu, W., Chen, X., 2020. ASNet: An adaptive scale network for skin lesion segmentation in dermoscopy images. In: Medical Imaging 2020: Biomedical Applications in Molecular, Structural, and Functional Imaging, Vol. 11317. International Society for Optics and Photonics. SPIE, pp. 226–231.

Zijdenbos, A.P., Dawant, B.M., Margolin, R.A., Palmer, A.C., 1994. Morphometric analysis of white matter lesions in MR images: Method and validation. IEEE Trans. Med. Imaging 13 (4), 716–724.

Zortea, M., Skrøvseth, S.O., Schopf, T.R., Kirchesch, H.M., Godtliebsen, F., 2011. Automatic segmentation of dermoscopic images by iterative classification. Int. J. Biomed. Imaging 2011.

Zou, K.H., Warfield, S.K., Bharatha, A., Tempany, C.M., Kaus, M.R., Haker, S.J., Wells III, W.M., Jolesz, F.A., Kikinis, R., 2004. Statistical validation of image segmentation quality based on a spatial overlap index. Academic Radiol. 11 (2), 178–189.

Zunair, H., Hamza, A.B., 2021. Sharp U-Net: Depthwise convolutional network for biomedical image segmentation. Comput. Biol. Med. 136, 104699.