Rethinking Data, Annotation, and Targets in Deep Learning for Skin Image Analysis

Kumar Abhishek[0000−0002−7341−9617]

School of Computing Science, Simon Fraser University, Canada
kabhishe@sfu.ca

Abstract. Deep learning for dermatological image analysis now rivals expert accuracy on standard benchmarks. That progress, however, rests on simplifying assumptions that ease benchmarking but diverge from the conditions of clinical practice. My doctoral research revisits one at each stage of the skin-lesion pipeline: that the standard training recipe suits scarce, imbalanced, and imperfectly labeled data; that a lesion has a single correct segmentation; and that the diagnosis is the only output worth predicting. On training, I introduce a Matthews correlation coefficient segmentation loss for severe class imbalance; ζ-mixup, a manifold-preserving generalization of mixup; and a quality-guided semi-supervised framework that grounds pseudolabel reliability in predicted segmentation quality rather than model confidence. On annotation, I discover plausible, diverse segmentation "styles" without annotator correspondence, and show that inter-annotator agreement is associated with malignancy, is predictable from the image alone, and, used as a "soft" clinical feature, improves diagnostic accuracy. On the target, I predict clinical management directly rather than inferring it from diagnosis, and recover lesion elevation, a property clinicians assess by palpation, from 2D images. Alongside these methods I release corrected benchmarks and the largest public multi-annotator skin lesion segmentation dataset. Together, this work treats data quality, annotator disagreement, and the choice of prediction target as design decisions rather than defaults, and provides the public resources needed to study them.

1 Research Problem and Motivation

Skin cancer is among the most commonly diagnosed cancers worldwide, and melanoma, although responsible for only a small fraction of cases, accounts for the majority of skin cancer deaths. However, survival rates depend sharply on the stage at which melanoma is diagnosed, which makes early and accurate diagnosis the single most consequential intervention for skin cancer [12]. Motivated by this, deep-learning (DL)-based computer-aided diagnosis (CAD) systems have made substantial progress and now approach dermatologist-level performance on curated image benchmarks [39].

A CAD system, however, is less a single classifier than a pipeline, and a benchmark score alone does not portray the entire picture. Images and their labels

must be collected and curated; lesions must be delineated from the surrounding skin, both to localize them and to quantify criteria that clinicians assess such as asymmetry and border irregularity [34]; and a prediction must be produced in a form that a clinician can act on. Each stage, however, relies on a working simplification that the clinical setting does not necessarily honor. Dermatological labels are scarce, expensive, and imbalanced [42], and consequently the field has responded with label-efficient training, from semi-/weakly-supervised methods [18] to entirely self-supervised pretraining [14]; yet there is much progress to be made in how the models utilize those labels, namely the training objective, the data augmentation pipeline, and the estimate of which pseudolabels to trust. Additionally, experts consistently exhibit inter- and intra-annotator variabilities in their delineations [32], yet the standard practice is to aggregate these labels into a single consensus target and train against it. Finally, diagnosis, although the near-universal prediction target, is not the only clinically relevant output: the decision acted upon is the management of the lesion, and cues that clinicians often rely on, such as whether a lesion is raised, are not recorded in a 2D RGB image at all.

To this end, I organize my doctoral research around three questions, one per stage in the pipeline: (i) how can models learn reliably from limited and imperfect labels? (ii) is there a single "correct" segmentation, and if not, what does annotator disagreement encode? and (iii) is diagnosis the only prediction target, or do clinically used decisions and features lie beyond it? My contributions (Fig. 1) answer them in turn: training methods robust to imbalance, unrealistic augmentation, and unreliable pseudolabels; a treatment of the inter-annotator variability that consensus masks discard as a signal that is learnable and as I show, diagnostically useful; and a broadening of the prediction target beyond the diagnosis label to include the management decision and physical cues that 2D images do not record.

These questions are pertinent to MICCAI in general rather than being limited to dermatology. Severe class imbalance, scarce annotation budgets, disagreement among expert readers, and a mismatch between the label a dataset records and the decision a clinician makes recur across modalities and anatomical regions, and the methods developed herein are formulated so as to be generally applicable. Dermatology is a particularly suitable testbed because public multi-annotator segmentations, recorded management decisions and lesion elevation, and multi-site demographically rich datasets are all available at scale.

2 Relevance and Related Work

This work sits at the intersection of medical image segmentation, label-efficient learning, and clinically grounded prediction. I have surveyed this literature three times as first/joint-first author: a broad review of deep semantic segmentation across natural and medical images [13], a review of DL methods for skin lesion segmentation (SLS) [32], and a review of attribution-based explainability meth-

ods in computer vision [7], the last of which is a complementary requirement for clinical deployment that falls outside the scope of this thesis.

From a model architecture and training perspective, SLS has moved from convolutional encoder-decoders such as U-Net and its self-configuring successors [37, 25] to Transformers and hybrid architectures [17], and more recently to promptable foundation models adapted to medical images [30]. In addition to the model architecture, training rests on two additional, independent choices: the objective, and the data. Overlap-based objectives such as the popular Dice loss [31] measure agreement over the foreground alone and count background pixels only when they are wrongly predicted as lesion, so the correctly classified background, which dominates the image when the lesion is small, never enters the objective. The established remedies reweight, whether over confusion matrix entries (Tversky loss [38]) or individual pixels (focal loss [29]), adjusting how the errors are weighted rather than what the objectives measure. The scale of data is typically constrained by how few annotated images exist, which limits both what can be synthesized from them and what can be drawn from the unlabeled samples. Augmentation attempts to address the former, though *mixup* and its successors [47, 46] combine samples linearly and can place augmented samples off the data manifold with implausible labels. Semi-supervised learning (SSL) attempts to alleviate the latter, often by weighting pseudolabels by model confidence or teacher-student agreement, so that the model both produces and "vets" its own supervision.

The segmentation annotations raise a separate issue. When several experts delineate the same lesion, the prevalent practice is to train against one mask obtained through consensus or label fusion, e.g., STAPLE [44], collapsing their systematic disagreement into a single target. A substantial body of work instead models that disagreement, whether by capturing ambiguity through a latent space [28, 15], modeling per-annotator preferences [26], or arguing that inter-annotator variability from crowd annotations is itself informative [19]. These methods, however, either assume annotator correspondence, i.e., knowledge of which annotator produced which mask, or model the ambiguity without asking whether it carries a diagnostic signal.

For the prediction target, CAD systems overwhelmingly predict the diagnosis and pay little to no attention to the downstream clinical management decision, treating it as a deterministic consequence. To the best of my knowledge, no prior work predicts the management decision directly from images, or recovers lesion elevation from 2D photographs to support diagnosis.

3 Scientific Approach

Question I: Can Models Learn from Limited and Imperfect Labels?

Can models learn reliably when labels are scarce, imbalanced, and imperfect? I aim to (i) establish that the standard training objective and the standard augmentation operator both encode assumptions that skin lesion data violate, and (ii) replace model self-confidence with an external, grounded estimate of

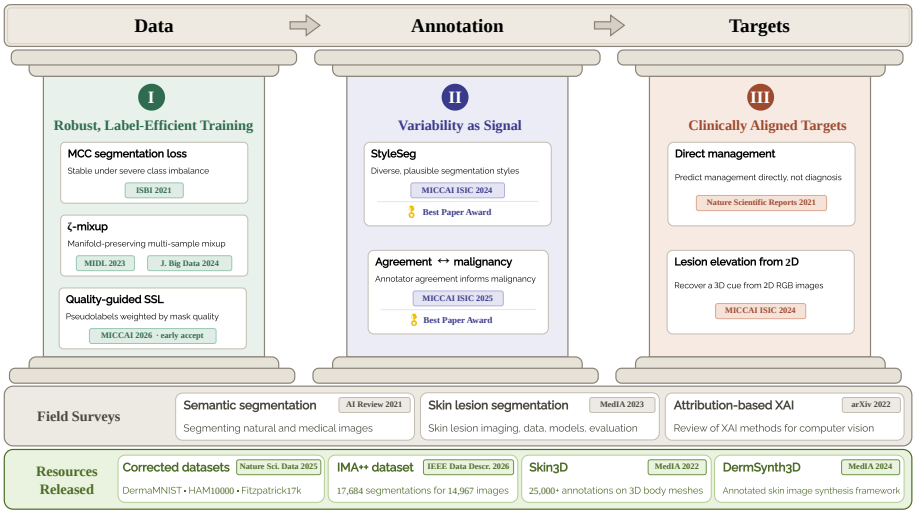

Fig. 1. Dissertation overview organized as three pillars. One question is posed at each stage of the skin image analysis pipeline, and each pillar collects the contributions that answer it (Sec. 3-5). The base gathers the surveys that frame the work and the resources released alongside it. Best viewed online.

pseudolabel reliability in the semi-supervised setting. These works have been published at ISBI 2021 [3], MIDL 2023 and the *Journal of Big Data* 2024 [1, 2], and MICCAI 2026 (early accept) [5].

Question II: Is There A Single Correct Segmentation?

Standard practice reduces multiple expert delineations of a lesion to one consensus mask. I aim to (i) recover the structure in that discarded disagreement without requiring annotator correspondence by learning the latent variation from unpaired masks directly, and (ii) test whether the disagreement is not merely structured but diagnostically informative by treating annotator agreement itself as a measurable quantity. These works have been published at MICCAI ISIC 2024 [9] and MICCAI ISIC 2025 [10], both receiving *Best Paper Awards*, and required curating the ISIC-MultiAnnot and IMA++ datasets [9, 11].

Question III: Beyond the Diagnosis Label

Diagnosis is essential, but is it the only clinically relevant target? I aim to (i) show that the clinical management decision is better predicted directly than inferred from a predicted diagnosis, and (ii) recover a physical cue, lesion elevation, that clinicians use and that 2D imaging does not capture. These works have been published in *Nature Scientific Reports* 2021 [8] and at MICCAI ISIC 2024 [4].

4 Proposed Solution

Question I: Robust, Label-Efficient Training

A metric-aligned segmentation loss: The Dice loss, although standard, is computed from true positives, false positives, and false negatives, and is therefore insensitive to the true negatives that dominate the image when lesions are small. I instead propose a loss based on the Matthews correlation coefficient (MCC), a single measure that accounts for all four entries of the confusion matrix [3]. MCC values range from -1 to 1, with -1 and 1 indicating a completely disjoint and a perfect prediction respectively, and minimizing its differentiable complement as the training loss aligns the objective with balanced segmentation quality.

Manifold-preserving augmentation: Standard mixup convexly combines two samples, which can place augmented points off the data manifold and assign them implausible labels. I propose ζ-mixup, a generalization that convexly combines $T \geq 2$ samples through a p-series interpolant, contains mixup as a special case, and as I show, better preserves the intrinsic dimensionality of the data [1, 2]. The resulting samples stay closer to the data manifold than mixup's, the operation reduces to a single matrix multiplication over a minibatch, and the method is a drop-in replacement requiring no change to the network or the training loop.

Quality-guided semi-supervised learning: Segmentation quality can be estimated without ground truth, but existing estimators assess finished segmentations [43]. I propose a quality-guided SSL framework that instead trains a dedicated network to estimate segmentation quality from image-mask pairs [5]. The predictor learns from masks of controlled, variable quality produced by synthetic corruptions and by partially trained segmentation models, so that it sees the error patterns that arise during training rather than synthetic ones alone. The estimate enters training through a quality-aware regularization loss and a quality-based pseudolabel reweighting scheme, making the framework a drop-in enhancement to existing SSL methods.

Question II: Variability as Signal

Discovering segmentation styles without correspondence: Multiple annotators produce systematically different yet plausible segmentations of the same lesion. I introduce the problem of segmentation style discovery and propose StyleSeg, which learns plausible, diverse, and semantically consistent segmentation "styles" from a corpus of image-mask pairs with no knowledge of annotator correspondence, together with a measure of annotation style, AS^2, that quantifies the alignment between discovered styles and annotator preferences [9].

Inter-annotator agreement as a clinical feature: Building on this, I then treat inter-annotator agreement (IAA) as a measurable quantity and ask three things of it in sequence [10]: whether it is associated with malignancy, whether it can be predicted from the image alone, and whether that prediction is useful. As I show, the last is particularly consequential: predicted IAA can be incorporated as a "soft" clinical feature in a multi-task objective, so that a model trained only

on images can condition its malignancy prediction on how much experts would be expected to disagree.

Question III: Clinically Aligned Targets

Predicting management directly: What a clinician acts on is the management decision (e.g., reassure, monitor, or excise), and it is not always a deterministic function of the diagnosis: the same diagnosis maps to different management depending on lesion site and patient metadata, and in some cases the management step, such as an excisional biopsy, is what establishes the diagnosis. I therefore propose to predict management directly from lesion images together with patient metadata, rather than inferring it from a predicted diagnosis [8]. Management is also a coarser target than the fine-grained diagnosis label, so predicting it directly prevents propagating diagnostic uncertainty into the clinical decision.

Recovering lesion elevation from 2D images: Dermatologists assess elevation by palpation, and it informs the "evolving/elevation" component of the ABCDE rule [41], yet teledermatology and image-only CAD typically have no access to it. The clinical literature notes that raised lesions are difficult to distinguish from flat ones in two-dimensional photographs, however good their quality [20]. I propose to predict image-level elevation labels from a single 2D image, use the trained predictor to label datasets that lack them, and feed the estimated elevation to the diagnostic model as an auxiliary input [4].

5 Results and Contributions

Question I

The MCC loss improves, statistically significantly, the mean Jaccard index over Dice-loss training across three public datasets (by up to 11.25%), with consistent gains in with a better sensitivity-specificity trade-off across all three datasets. ζ-mixup better preserves intrinsic dimensionality of the data, and across 26 diverse natural- and medical-image classification datasets, outperforms mixup, CutMix, and standard augmentation [2]. Quality-guided SSL outperforms confidence- and uncertainty-based SSL baselines across five datasets spanning dermoscopy and colonoscopy and three segmentation architectures, with the labeled and unlabeled sets drawn from different sources, a setting closer to clinical practice [5]. None of the three requires architectural modification, which suggests that the objective, the augmentation, and the pseudolabel reliability estimate remain productive places to look for label efficiency alongside advances in architecture.

Question II

StyleSeg consistently outperforms competing methods on four public SLS datasets while producing diverse, plausible segmentations. Even when ten styles are modeled, its least plausible segmentation agrees with the ground truth far more closely than that of a multiple-hypothesis baseline, while its discovered styles align with the recorded annotator preferences [9] On the IMA++ dataset [11], benign lesions show higher inter-annotator agreement than malignant ones; because a comparison of means alone can mislead, I compare the full distributions

of benign-versus-malignant IAA scores, which confirms a statistically significant association between agreement and malignancy ($p < 0.001$) [10]. Agreement is predictable directly from the dermoscopic image and, incorporated as a "soft" clinical feature in a multi-task model, improves balanced accuracy of malignancy classification by 4.2% across IMA++ and four public datasets, thus indicating that the disagreement that consensus masks discard carries a diagnostic signal.

Question III

On the Interactive Atlas of Dermoscopy (1,011 cases, 20 diagnoses, three management decisions), predicting management directly is more accurate than predicting the diagnosis and then inferring management (by 13.73% in overall accuracy), and it reduces the over-excision rate substantially, with 24.56% fewer cases wrongly predicted for excision [8]. Because both models see the same images and share a backbone, the gain is attributable to the change of the prediction target rather than a relabeling of the diagnosis task. On the public MClass-D benchmark [16], the model's management recommendations agree with those of 157 dermatologists as much as the dermatologists agree among themselves. For lesion elevation, predicted elevation labels used as auxiliary inputs improve diagnostic classification (AUROC gains of up to 6.29%), indicating that a physical cue clinicians rely on can be partially recovered from appearance alone.

Data and Resources Released

A recurring lesson across these projects is that datasets' flaws silently propagate into the models and thus the reported results, and so a substantial part of my dissertation consists of resources released to the community. Benchmark quality is not a solved problem even in mainstream computer vision, where pervasive test-set label errors change model rankings [35], and dermatology benchmarks add further complexity in the form of patient- and lesion-level data hierarchy that naive partitioning ignores. I audited three widely used benchmarks: DermaMNIST [45], its source HAM10000 [42], and Fitzpatrick17k [24], and found data quality issues that inflate reported performance [6]. HAM10000's 10,015 images derive only from 7,470 unique lesions, an overlap not accounted for when partitioning DermaMNIST, which leaves 13.47% of lesions in more than one split. I measured the effect on published results and released corrected datasets: DermaMNIST-C, DermaMNIST-E, and Fitzpatrick-C, with updated benchmarks; this is crucial for fairness audits that these datasets increasingly support [21]. To study annotation variability, I curated ISIC-MultiAnnot [9] and then IMA++ [10, 11], the largest public multi-annotator SLS dataset, containing 17,684 segmentation masks of 14,967 dermoscopic images, 2,394 of which carry two to five masks each, with annotator skill level and segmentation tool metadata that enable fine-grained analysis.

Collaborative Contributions

Several collaborative efforts extend these themes beyond my first-authored work. As lead PhD student, I contributed to Skin3D (MedIA 2022), which tracks skin lesions on 3D total-body meshes and releases over 25,000 annotations, the first

large-scale resource of its kind [48]. The variability thread explored in Question II traces its roots to D-LEMA (CVPR ISIC 2021, *Best Paper Award*), which proposed DL-based ensembles from multiple annotations [33]. CIRCLe (ECCV ISIC 2022) learns color-invariant representations for fairer classification across skin tones [36]. Two contributions focus on synthetic data: DermSynth3D (MedIA 2024) synthesizes in-the-wild annotated dermatology images [40], while a companion analysis (under review) presents the intrinsic limits of synthetic data along with its associated ethical risks and proposes actionable stakeholder recommendations [27]. PET-Disentangler (SNMMI 2024; ISBI 2025, *Best Paper Award*) carries the segmentation work into another modality, positron emission tomography (PET), by disentangling disease from healthy anatomy [22, 23].

6 Open Challenges and Future Work

While this thesis makes substantial progress on each of the three questions posed above, several open challenges remain. First, the three threads have so far been developed largely independently, and the next step would be to carry annotation variability and predicted segmentation quality through to the management decision, building an uncertainty-aware decision support model that reports not only a management prediction but a calibrated estimate of how much experts would be expected to disagree about the case. Second, inter-annotator agreement is currently, both in my work and across the literature, summarized by averaging pairwise metrics (Dice, boundary distances) over annotator pairs, which cannot distinguish a case where one annotator disagrees with the rest from one where all annotators disagree "mildly". I believe that principled groupwise measures would distinguish these cases and give a more faithful account of consensus and its uncertainty. Finally, (dis)agreement is at present predicted as one number per image, whereas in reality, disagreement is spatially localized, concentrated at ambiguous boundaries. Predicting where annotators are likely to agree, as a dense spatial map rather than a scalar, could guide annotation effort, and feed uncertainty back into segmentation.

7 Dissertation Status and Long-Term Goals

The three threads of this dissertation, namely robust training, annotator disagreement modeled as structured signal, and clinically relevant prediction targets, are steps toward a single longer-term goal: dermatological image analysis that is dependable by construction, trained on verified data, explicit about disagreement and uncertainty rather than reduced to a single confident number, and evaluated on clinically meaningful endpoints. Realizing that goal requires carrying the work beyond the retrospective, benchmark setting in which it was developed, and into prospective evaluation with clinical collaborators, which is what I intend to pursue beyond the dissertation.

References

1. Abhishek, K., Brown, C.J., Hamarneh, G.: ζ-mixup: Richer, More Realistic Mixing of Multiple Images. In: MIDL (2023)
2. Abhishek, K., Brown, C.J., Hamarneh, G.: Multi-sample ζ-mixup: Richer, More Realistic Synthetic Samples from a p-series Interpolant. Journal of Big Data (2024)
3. Abhishek, K., Hamarneh, G.: Matthews Correlation Coefficient Loss For Deep Convolutional Networks: Application To Skin Lesion Segmentation. In: ISBI (2021)
4. Abhishek, K., Hamarneh, G.: Lesion Elevation Prediction from Skin Images Improves Diagnosis. In: MICCAI ISIC (2024)
5. Abhishek, K., Hamarneh, G.: Quality-guided semi-supervised learning for medical image segmentation. In: MICCAI (2026)
6. Abhishek, K., Jain, A., Hamarneh, G.: Investigating the Quality of DermaMNIST and Fitzpatrick17k Dermatological Image Datasets. Nature Scientific Data (2025)
7. Abhishek, K., Kamath, D.: Attribution-based XAI Methods in Computer Vision: A Review. arXiv preprint arXiv:2211.14736 (2022)
8. Abhishek, K., Kawahara, J., Hamarneh, G.: Predicting the Clinical Management of Skin Lesions using Deep Learning. Nature Scientific Reports (2021)
9. Abhishek, K., Kawahara, J., Hamarneh, G.: Segmentation Style Discovery: Application to Skin Lesion Images. In: MICCAI ISIC (2024)
10. Abhishek, K., Kawahara, J., Hamarneh, G.: What Can We Learn from Inter-Annotator Variability in Skin Lesion Segmentation? In: MICCAI ISIC (2025)
11. Abhishek, K., Kawahara, J., Hamarneh, G.: Descriptor: ISIC Archive Multi-Annotator Dermoscopic Skin Lesion Segmentation Dataset (IMA++). IEEE Data Descriptions (2026)
12. American Cancer Society: Cancer Facts & Figures 2026 (2026)
13. Asgari Taghanaki*, S., Abhishek*, K., Cohen, J.P., Cohen-Adad, J., Hamarneh, G.: Deep Semantic Segmentation of Natural and Medical Images: A Review. Artificial Intelligence Review (2021)
14. Azizi, S., et al.: Big self-supervised models advance medical image classification. In: ICCV (2021)
15. Baumgartner, C.F., et al.: PHiSeg: Capturing uncertainty in medical image segmentation. In: MICCAI 2019 (2019)
16. Brinker, T.J., et al.: Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. European Journal of Cancer (2019)
17. Cao, H., et al.: Swin-Unet: Unet-like pure transformer for medical image segmentation. In: ECCV MCV (2022)
18. Chaitanya, K., et al.: Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation. Med Image Anal (2023)
19. Cheplygina, V., et al.: Crowd disagreement about medical images is informative. In: MICCAI LABELS (2018)
20. Cox, N.H.: Palpation of the skin – an important issue. J R Soc Med (2006)
21. Daneshjou, R., et al.: Disparities in dermatology AI performance on a diverse, curated clinical image set. Science Advances (2022)
22. Gatsak, T., Abhishek, K., Ben Yedder, H., Asgari Taghanaki, S., Hamarneh, G.: PET-Disentangler: PET Lesion Segmentation via Disentangled Healthy and Disease Feature Representations. Journal of Nuclear Medicine (2024)
23. Gatsak, T., Abhishek, K., Yedder, H.B., Taghanaki, S.A., Hamarneh, G.: Disentangled PET Lesion Segmentation. In: ISBI (2025)

24. Groh, M., et al.: Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset. In: CVPR ISIC (2021)
25. Isensee, F., et al.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods (2021)
26. Ji, W., et al.: Learning calibrated medical image segmentation via multi-rater agreement modeling. In: CVPR (2021)
27. Jin, W., Sinha, A., Abhishek, K., Hamarneh, G.: Ethical Medical Image Synthesis. arXiv preprint arXiv:2508.09293 (2025)
28. Kohl, S., et al.: A probabilistic U-Net for segmentation of ambiguous images. In: NeurIPS (2018)
29. Lin, T.Y., et al.: Focal loss for dense object detection. In: ICCV (2017)
30. Ma, J., et al.: Segment anything in medical images. Nat Commun (2024)
31. Milletari, F., et al.: V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: 3DV (2016)
32. Mirikharaji*, Z., Abhishek*, K., Bissoto, A., Barata, C., Avila, S., Valle, E., Celebi, M.E., Hamarneh, G.: A Survey on Deep Learning for Skin Lesion Segmentation. Medical Image Analysis (2023)
33. Mirikharaji, Z., Abhishek, K., Izadi, S., Hamarneh, G.: D-LEMA: Deep Learning Ensembles from Multiple Annotations - Application to Skin Lesion Segmentation. In: CVPR ISIC (2021)
34. Nachbar, F., et al.: The ABCD rule of dermatoscopy. J Am Acad Dermatol (1994)
35. Northcutt, C.G., et al.: Pervasive label errors in test sets destabilize machine learning benchmarks. In: NeurIPS (2021)
36. Pakzad, A., Abhishek, K., Hamarneh, G.: CIRCLe: Color Invariant Representation Learning for Unbiased Classification of Skin Lesions. In: ECCV ISIC (2022)
37. Ronneberger, O., et al.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
38. Salehi, S.S.M., et al.: Tversky loss function for image segmentation using 3D fully convolutional deep networks. In: MICCAI MLMI (2017)
39. Salinas, M.P., et al.: A systematic review and meta-analysis of artificial intelligence versus clinicians for skin cancer diagnosis. NPJ Digit Med (2024)
40. Sinha, A., Kawahara, J., Pakzad, A., Abhishek, K., Ruthven, M., Ghorbel, E., Kacem, A., Aouada, D., Hamarneh, G.: DermSynth3D: Synthesis of in-the-wild Annotated Dermatology Images. Medical Image Analysis (2024)
41. Strayer, S.M., et al.: Diagnosing skin malignancy: Assessment of predictive clinical criteria and risk factors. J Fam Pract (2003)
42. Tschandl, P., et al.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Nat Sci Data (2018)
43. Valindria, V.V., et al.: Reverse Classification Accuracy: Predicting Segmentation Performance in the Absence of Ground Truth. IEEE Trans Med Imaging (2017)
44. Warfield, S., et al.: Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. IEEE Trans Med Imaging (2004)
45. Yang, J., et al.: MedMNIST v2 - A large-scale lightweight benchmark for 2D and 3D biomedical image classification. Nat Sci Data (2023)
46. Yun, S., et al.: CutMix: Regularization strategy to train strong classifiers with localizable features. In: ICCV (2019)
47. Zhang, H., et al.: mixup: Beyond empirical risk minimization. In: ICLR (2018)
48. Zhao, M., Kawahara, J., Abhishek, K., Shamanian, S., Hamarneh, G.: Skin3D: Detection and Longitudinal Tracking of Pigmented Skin Lesions in 3D Total-Body Textured Meshes. Medical Image Analysis (2022)