Less is More: Sample Selection and Label Conditioning Improve Skin Lesion Segmentation

Vinicius Ribeiro¹ Sandra Avila² Eduardo Valle¹
¹School of Electrical and Computing Engineering (FEEC) ²Institute of Computing (IC)
RECOD Lab., University of Campinas (UNICAMP), Brazil

Abstract

Segmenting skin lesions images is relevant both for itself and for assisting in lesion classification, but suffers from the challenge in obtaining annotated data. In this work, we show that segmentation may improve with less data, by selecting the training samples with best inter-annotator agreement, and conditioning the ground-truth masks to remove excessive detail. We perform an exhaustive experimental design considering several sources of variation, including three different test sets, two different deep-learning architectures, and several replications, for a total of 540 experimental runs. We found that sample selection and detail removal may have impacts corresponding, respectively, to 12% and 16% of the one obtained by picking a better deep-learning model.

1. Introduction

Witholding data to improve machine learning is counter-intuitive, but we will show it brings promising improvements for skin lesion segmentation, by selecting the training samples with best inter-annotator agreement, and conditioning the ground-truth masks to remove excessive detail.

Segmenting skin lesions images, i.e., delimiting the lesion from the surrounding skin, is frequently employed both as an end-result and as an adjutant for lesion classification. Segmentation, however, is a very challenging task, in part due to the difficulty in obtaining properly annotated data.

All supervised machine-learning models need annotated data to be trained, posing unique challenges for medical tasks, where the annotators are specialists whose time is costly and scarce. Segmentation poses additional challenges since the annotations are intricate region borders instead of a single label. Researchers have attempted circumventing the need for data, with techniques like data augmentation [5, 34], the creation of artificial data from generative models [6], or even the use of self-supervision, which allows performing part of the training of supervised models without labels [21].

Quality of training data is another important dimension, in addition to quantity. For skin lesion images, recent works have addressed that issue, exploring the lack of inter-annotator agreement in the ground truth of segmentation images [36], and the possibility of using annotations of different levels of confidence and granularity to learn segmentation masks [28].

In this work, we will follow a novel perspective: instead of finding ways to amplify training data, we will show how less of it can enhance results. The two main contributions of this work are:

• We show how training sample selection, based on inter-annotator agreement, can improve segmentation results, even when such selection is not applied to the test sets;

• We show how removing details from the ground-truth masks — using very simple “conditionings” — can improve segmentation results, even when the same details are still required in the test set, i.e., when the test masks are not “conditioned”.

We evaluate those contributions thoroughly, in an exhaustive experimental design that considers several sources of variation, including three different test sets, two different deep-learning architectures, and several replications, for a total of 540 experimental runs.

The remainder of the text is organized as follows. We survey the related works in section 2. We present the sample selection technique in section 3, and the ground-truth conditioning in section 4. We provide details about our datasets, models, implementation, and experimental design in section 5, with results following in section 6. We conclude the paper in section 7.

2. Literature Review

Segmenting skin-lesion images has attracted scientific interest since the inception of automated skin-lesion anal-
A recent development in skin-lesion segmentation is the use of generative models. Xue et al. [44] proposed SegAN, an end-to-end adversarial network architecture with multiscale loss, and achieved 4th place at the 2018 ISIC Challenge.

Training, and especially, evaluating machine-learning models, require accurate annotations. Ribeiro et al. [36] find, however, that information about inter-annotator agreement in visual datasets is very scarce, and when present, suggest a large variation among different tasks. In particular, for skin-lesion segmentation, they find the degree of agreement is only moderate, with a considerable portion of the samples having very poor inter-annotator agreements.

There are different solutions to that issue. On the one hand, we may ameliorate the quality of the annotations. Because reannotating the data is very expensive, Ribeiro et al. [36] suggest conditioning operations on the ground-truth masks that remove details, improving their agreement.

On the other hand, we may render our models less sensitive to the noise. Deep learning models are, by nature, fairly insensitive to noisy annotations [37]. An in-depth survey of segmentation techniques for medical images from noisy datasets [40] addresses both the issue of scarce and imperfect annotations, and, for the latter, lists techniques to deal with weak labels (in the technical sense of weakly supervised learning), sparse labels (only part of the image is annotated), and noisy labels (labels with ambiguities and inaccuracies). Specifically for skin-lesion segmentation, Mirikharaji et al. [28] address a continuum of annotations, ranging from fully detailed ground-truths until progressively weaker ones, by using polygons with fewer vertices, and ending with just a bounding box. They proposed a spatial-adaptive reweighting to treat clean and noisy pixel-level annotations in the loss function.

In this work, we propose a third alternative: removing the noisy samples from the dataset and, following Ribeiro et al. [36], removing excessive detail from the ground truths on the remaining samples. While the focus of Ribeiro et al. was improving the inter-annotator agreement on the dataset, here we focus on the machine-learning models and evaluate the impact of removing those details on them.

3. Sample Selection based on Inter-Annotator Agreement

As mentioned, Ribeiro et al. [36] found a broad diversity in the inter-annotator agreement for the ISIC dataset images. In particular, the authors noticed a fairly “heavy tail” of very discordant annotations in their observation.

In this work, we evaluate the actual effect of those observations on segmentation models, by contrasting models learned the usual way, without any data selection, with models learned with fewer samples, eliminating the worst discordant samples in the tail.
To perform a fair comparison, we first selected all samples from the online ISIC Archive dataset with at least two segmentation ground-truth annotations. For each of those samples, we computed the average pairwise Cohen’s Kappa score [26] for all existing ground-truth annotations. All samples with an average score above 0.5 went to the best samples dataset, and all samples, however the score, went to the all samples dataset.

Details about the data and selection procedure are in Sections 5.1, 5.2, 6.1 and 6.2.

4. Detail Elimination with Label Conditioning

In order to enhance inter-annotator agreement, Ribeiro et al. [36] propose applying “conditionings” on the ground-truth segmentation masks, which consist of eliminating details from them. They evaluate (in growing aggressiveness) the morphological operations of opening and closing, the convex hull, the morphological operations combined with the convex hull, and a bounding box.

In this work, we follow up on the idea of conditioning the ground-truth masks, from a different point of view: the machine-learning model. Instead of measuring how much different conditionings affect the mask agreement, we will measure how they affect both the training and the evaluation (when applied to the test set) of segmentation models.

In addition to the original images, we selected the two most promising conditionings proposed by Ribeiro et al. [36] for evaluation (Figure 1):

None  no conditioning: the original images — used as a control;

Opening  this morphological operation removes details like small protrusions in the lesion area. The structuring element was a 5-pixel-wide square;

Convex Hull  opening, just as above, followed by taking the convex hull, i.e., finding the tightest convex polygon that contains the lesion area.

We may interpret conditioning as denoising operations, aiming at preserving the cogent information while discarding details that arise from choosing a particular annotator. Our hypothesis, in this work, is that those annotator-dependent details may prove an expensive distraction for the models to learn.

Details about the procedure are in Sections 5.2 and 6.3.

5. Materials and Methods

5.1. Datasets

All training data used in this work came from the ISIC Archive [1] — curated by the International Skin Imaging Collaboration — the largest publicly available dataset of images of skin lesions. Although a few other datasets also provide segmentation information [2, 27], as far as we know, the ISIC Archive is the only public dataset with more
# of Masks # of Samples
1 11 546
2 2 094
3 100
>4 39
Total 13 779

Table 1. Distribution of samples by number of ground-truth segmentation masks in our ISIC Archive collection.

than one segmentation annotation per lesion, and thus the only one where inter-annotator agreement can be appraised.

The ground truth annotations are highly variable due, in part, to three different methods to create the annotations (Figure 2) and, in part, to differences of opinion and other specificities of human annotators.

At the time we collected our data, the ISIC Archive dataset contained 13 779 images with segmentation ground truth masks, 2 233 of those having multiple masks (Table 1). This latter number limited the training set for our experiments. We derived two training sets: one containing all 2 233 samples (all samples), and other whose average pairwise Cohen’s Kappa score between ground-truth masks was higher than 0.5, (best samples). The latter had 1 808 lesion images, i.e., only 81% of the available samples.

We employed three datasets for testing the models. The first was formed by a random selection of 2 000 images from the 11 546 images of our ISIC Archive collection with only one segmentation mask. Two others are the PH2 dataset [27] collected at the Porto University, with 200 dermoscopic images, and the Edinburgh Dermofit Library [2], with 1 300 focal high-quality clinical images. Since the inter-annotator agreement in those datasets cannot be appraised, all three of them represent “in-the-wild” situations, without sample filtering. The first dataset represents the typical machine-learning evaluation pipeline, with training and evaluation being subsamples of the same dataset. In contrast, the two others represent a cross-dataset scenario, that challenges the generalization abilities of the models.

5.2. Models and conditionings

LinkNet [8] is a traditional “U-Net-like” architecture: encoder–decoder with skip connections between them. DeepLab V3+ [9], in contrast, uses ResNet as the primary feature extractor, introduces new residual blocks for learning multi-scale features, and employs atrous convolutions with different dilation rates in the last residual block to better context understanding and scale invariance.

To train the networks, we split the training samples into 80/20 training and validation sets (Table 2). All lesions have more than one ground-truth mask during training; we randomly select which mask to use every time we pick a sample to compose a batch. Thus, different masks may appear at different times during training. For model selection during validation, for each sample, we evaluate the target metric (Jaccard index) using all available annotations and retain the best (i.e., the highest). The test datasets have a single annotation per lesion, so mask selection and metric computation are straightforward.

We trained each model for 100 epochs, without early stopping, with an Adam optimizer [23] and learning rate of 0.003. The loss function was a weighted sum of the soft Jaccard with the Binary Cross Entropy with Logits [19], with weights, respectively, of 8 and 1.

We applied three data augmentations, aiming to teach our model to be invariant to noise, color, and contrast. We add to each sample a Gaussian noise with zero mean and standard deviation of 2. We also add a color and a contrast enhancement, each parameterized by a Gaussian factor with mean 0.5 and standard deviation of 0.1, implemented using the Pillow image library [25].

We implemented the models, training, and evaluation pipelines using the PyTorch framework for deep learning [33]. We developed all the conditionings in Python, using the morphology package of the scikit-image library [43], and auxiliary code in NumPy [31]. During validation, we always apply to the masks the same conditioning used for training.

All code necessary to reproduce this work is available at our Github repository1.

5.3. Experimental design

We ran a single exhaustive experimental design to validate both the sample selection and the ground-truth conditioning. The design also aimed at capturing sources of variation present in the actual deployment of segmentation models and included the following factors:

| Training Set    | Split | All Samples | Best Samples |
|-----------------|-------|-------------|--------------|
| Training        | 1 786 | 1 449       |
| Validation      | 447   | 359         |
| Total           | 2 233 | 1 808       |

Table 2. Training sets and their splits.

---

1https://github.com/vribeiro1/skin-lesion-segmentation-agreement
The statistical analysis was a full factorial analysis of variance (ANOVA), which we used both to measure significance (p-values) and effect sizes (η²). In addition to the statistical test, we employed interaction plots to elucidate the relationship between the factors.

### 6. Results

#### 6.1. Dataset analysis and sample selection

The exact results varied by test set, with PH2 showing a slight decrease, and ISIC and Edinburgh Dermofit showing a slight increase in performance. Those results showcase distributions are pushed towards higher kappa values. The conditionings are not, however, able to deal with large discordances in the annotations, and all distributions have a fairly heavy tail of very low kappa values. The percentiles of the kappa values in Table 3 also reinforce those findings.

The dotted red line shows the threshold of 0.5 used to select the 1,808 samples of the best samples subset. Notice that this set has a fixed size since the selection is made on the unconditioned kappa values, regardless of the conditionings used in the experiment.

#### 6.2. Impact of sample selection

Figure 3 is an interaction plot highlighting the effect of sample selection, the choice between the all samples vs. best samples in the training set factor of our experimental design. The average effect of that choice can be appreciated on the solid lines in that plot, where selecting the best samples for training appears systematically above picking all samples. Recall that this implies discarding almost 20% of the training samples, and that such selection is not performed on the test sets. Those results are far from trivial, since deep-learning if fairly robust to noise [37] and often presents better results in larger noisier dataset than in smaller cleaner ones.

In addition to those averaged aggregate results, two other results deserve attention. First, there are interactions among sample selection on the training set, conditioning on the training set, and conditioning on the test set, with those facts act synergetically. We will explore those interactions in more detail in subsection 6.4. Second, the results may vary according to the test set. Indeed, for the PH2 dataset there is a slight inversion of the results (although the experiments are quite mixed, as the individual data points show). In contrast, the Edinburgh Dermofit dataset shows the largest positive differences, which is remarkable given that dataset has focal clinical images instead of dermoscopic images and, thus, poses the widest generalization gap for the models to bridge.

#### 6.3. Impact of ground truth conditioning

The most surprising result is that removing details from the ground truths on the training set does not reduce the performance, on average, of models — even when those same details are required on the test sets (leftmost panel). The exact results varied by test set, with PH2 showing a slight decrease, and ISIC and Edinburgh Dermofit showing a slight increase in performance. Those results showcase
Figure 3. Distributions of inter-annotator agreements for the ground-truths pre- (none) and post- the evaluated conditionings (opening and convex hull) on the 2,233 samples of the all samples subset from our ISIC Archive collection. Both plots show the same data. Right: the histograms (shaded areas) and estimated densities (superimposed lines). Left: the original samples (black dots), the estimated densities (violin plots), and the estimated means (red dot). The red dotted line shows the threshold used for the best samples dataset.

Figure 4. Interaction plot highlighting the effect of sample selection, i.e., the training set factor in our experimental design. Points are individual experiments, dashed lines are averages per test set, solid line is the average for all three test sets. Each panel show the results for a choice of conditioning on the test set (leftmost: no conditioning). Sample selection improves results across a large variation of treatments.

That adding excessive detail in the ground-truth masks may be counterproductive. That is particularly true when those details are not needed for the target application. The rightmost panel
shows that when the convex hull conditioning is applied to both training and test ground truth masks — to evaluate, e.g., an application where the rough contour of the lesion is enough — the results sharply increase, for all three test sets.

6.4. Statistical analysis

We can divide, in our statistical analysis, the sources of variation in three groups:

**Design factors** Those are the factors we could actively control in the actual deployment of a machine-learning model. In our experiment, those are model, training set, training conditioning, and test conditioning.

**Nuisance factors** Those are the factors we cannot control in any actual deployment of a model, but we can control in an experiment. In our experiment, the single factor in this category is test set.

**Uncontrolled sources** Those are sources of variation we cannot control in either situation: fluctuations in training (random seeds, numerical errors, etc.), hardware fluctuations, etc.

The statistical analysis was a full factorial ANOVA. All factors were found significant, with tiny p-values ($\sim 10^{-16}$ for test conditioning, $\sim 10^{-16}$ for all others). We considered up to 3rd order interactions, and several of them were significant, notably almost all 2nd order interactions (the exceptions were training set with test conditioning and model with test conditioning).

The main source of variation was test set, which explained 88% of the global variation (i.e., $\eta^2 = 0.88$ effect size). That is perhaps unsurprising considering the three datasets varied widely in difficulty, with the PH2 dataset being much easier to segment than the other two. Uncontrolled sources accounted for less than 2% of the variation. The remainder variation was scattered among the other factors and interactions, model being the largest by far (6% of the variation).

Considering only the variation we can design for, model was the most influential, explaining almost 57% of it. Sample selection was considerable, with training set gathering over 7% of the variation. Training conditioning alone explained just a little over 1% of the variation, while test conditioning alone gathered almost 9%. In addition, both factors interacted to explain almost 4% of the “designable” variation.

7. Conclusions

As previously observed by Ribeiro et al. [36], segmentation ground-truths for skin lesion images present substantial inter-annotator disagreement. Although that, for the moment, can only be measured on the ISIC Archive, there is no reason to believe the results would be different for other datasets, if they had more than one annotation available per sample. In this work, we showcased how a strategy of selecting the samples with largest disagreement may result in significantly improved performance. We also showed how removing details on the segmentation masks (by condition-
ing them with simple operators) may improve the results, especially if those details are not needed on prediction time.

To put our findings in perspective, consider the improvement brought from moving from LinkNet to DeepLab: this was the most important "designable" factor we found but, of course, creating new deep learning architectures is a laborious and haphazardous enterprise. One can obtain 12% of that improvement simply by throwing away 1/5th of the training data. By giving up detail on the segmentation masks, one can obtain 16% of that improvement. For many applications of lesions segmentation (e.g., finding a rough contour, or determining a lesion diameter) less is more, in a very concrete sense.

The sample selection technique proposed in this paper requires multiple annotations per sample, a condition that makes it applicable to very few of the available training data. In the future, we would like to extend it to samples with a single ground-truth mask, greatly increasing its applicability.

8. Acknowledgements

E. Valle is partially funded by a CNPq PQ-2 grant (311905/2017-0), and by a FAPESP grant (2019/05018-1). S. Avila is partially funded by Google Research Awards for Latin America 2018 & 2019, FAPESP (2017/16246-0) and FAEPPEX (3125/17). This project is partially funded by CNPq Universal grant (424958/2016-3). RECOD Lab. is partially supported by diverse projects and grants from FAPESP, CNPq, and CAPES. The funding sources had no involvement in the data acquisition, study design, result analysis, or in the manuscript writing. We gratefully acknowledge the donation of GPU Hardware by NVIDIA Corporation, used in this work.

References

[1] International Skin Imaging Collaboration: Melanoma Project. https://isic-archive.com. 3
[2] L. Ballerini, R. B Fisher, B. Aldridge, and J. Rees. A color and texture based hierarchical k-nn approach to the classification of non-melanoma skin lesions. In Color Medical Image Analysis, pages 63–86. Springer, 2013. 3, 4
[3] M. Berseth. Isic 2017-skin lesion analysis towards melanoma detection. arXiv preprint arXiv:1703.00523, 2017. 2
[4] L. Bi, J. Kim, E. Ahn, and D. Feng. Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks. arXiv preprint arXiv:1703.04197, 2017. 2
[5] D. Bisla, A. Choromanska, R. S. Berman, J. A. Stein, and D. Polsky. Towards automated melanoma detection with deep learning: Data purification and augmentation. In IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019. 1
[6] A. Bissoto, F. Perez, E. Valle, and S. Avila. Skin lesion synthesis with generative adversarial networks. In OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis, pages 294–302. 2018. 1
[7] M. E. Celebi, Q. Wen, H. Iyatomi, K. Shimizu, H. Zhou, and G. Schaefer. A state-of-the-art survey on lesion border detection in dermoscopy images. Dermoscopy Image Analysis, pages 97–129, 2015. 2
[8] A. Chaurasia and E. Culurciello. Linknet: Exploiting encoder representations for efficient semantic segmentation. In 2017 IEEE Visual Communications and Image Processing (VCIP), pages 1–4. IEEE, 2017. 4
[9] L.C. Chen, G. Papandreou, F. Schroff, and H. Adam. Re-thinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587, 2017. 2, 4
[10] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC). In IEEE International Symposium on Biomedical Imaging, pages 168–172, 2018. 2
[11] N. C. F. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, et al. Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC). arXiv preprint arXiv:1902.03368, 2019. 2
[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In IEEE Conference on Computer Vision and Pattern Recognition, pages 248–255, 2009. 2
[13] A. P. Dhawan and A. Sim. Segmentation of images of skin lesions using color and texture information of surface pigmentation. Computerized Medical Imaging and Graphics, 16(3):163–177, 1992. 2
[14] H. Du, J. Young Seok, D. Ng, N. K. Yuan, and M. Feng. Team holidayburned at isic challenge 2018. Technical report, 2018. 2
[15] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. International Journal of Computer Vision, 88(2):303–338, 2010. 2
[16] M. Fornaciari, M. Carvalho, F. V. Bittencourt, S. Avila, and E. Valle. Towards automated melanoma screening: Proper computer vision & reliable results. arXiv preprint arXiv:1604.04024, 2016. 2

3189
[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In IEEE conference on computer vision and pattern recognition, pages 770–778, 2016. 
[18] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In IEEE international conference on computer vision, pages 2961–2969, 2017. 
[19] V. Iglovikov, S. Mushinskiiy, and V. Osin. Satellite imagery feature detection using deep convolutional neural network: A kaggle competition. arXiv preprint arXiv:1706.06169, 2017. 
[20] Y. Ji, X. Li, G. Zhang, D. Lin, and H. Chen. Automatic skin lesion segmentation by feature aggregation convolutional neural network. Technical report, 2018. 
[21] L. Jing and Y. Tian. Self-supervised visual feature learning with deep neural networks: A survey. arXiv preprint arXiv:1902.06162, 2019. 
[22] Alexandr A Kalinin, Vladimir I Iglovikov, Alexander Rakhlin, and Alexey A Shvets. Medical image segmentation using deep neural networks with pre-trained encoders. In Deep Learning Applications, pages 39–52. Springer, 2020. 
[23] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014. 
[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems, pages 1097–1105, 2012. 
[25] F. Lundh, M. Ellis, et al. Python imaging library (pil), 2012. 
[26] M. L. McHugh. Interrater reliability: the kappa statistic. Biochemia medica: Biochemia medica, 22(3):276–282, 2012. 
[27] T. F. Mendonca, M. E. Celebi, T. Mendonca, and J. S. Marques. Ph2: A public database for the analysis of dermoscopic images. Dermoscopy image analysis, 2015. 
[28] Z. Mirikharaji, Y. Yan, and G. Hamarneh. Learning to segment skin lesions from noisy annotations. In Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data, pages 207–215. Springer, 2019. 
[29] R. H. Moss, G. A. Hance, S. E. Umbaugh, and W. V. Stoecker. Unsupervised color image segmentation: with application to skin tumor borders. 1996. 
[30] F. Nachbar, W. Stolz, T. Merkle, A. B. Cognetta, T. Vogt, M. Landthaler, P. Bilek, O. Braun-Falco, and G. Plewig. The abcd rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions. Journal of the American Academy of Dermatology, 30(4):551–559, 1994. 
[31] T. E. Oliphant. A guide to NumPy, volume 1. Trelgol Publishing USA, 2006. 
[32] R. B. Oliveira, J. P. Papa, A. S. Pereira, and J. M. R. S. Tavares. Computational methods for pigmented skin lesion classification in images: review and future trends. Neural Computing and Applications, 29(3):613–636, 2018. 
[33] A. Paszke, S. Gross, S. Chintal, G. Chanan, E. Yang, Z. De Vito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017. 
[34] F. Perez, C. Vasconcelos, S. Avila, and E. Valle. Data augmentation for skin lesion analysis. In OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis, pages 303–311. Springer, 2018. 
[35] C. Qian, T. Liu, H. Jiang, Z. Wang, P. Wang, M. Guan, and B. Sun. A two-stage method for skin lesion analysis. arXiv preprint arXiv:1809.03917, 2018. 
[36] V. Ribeiro, S. Avila, and E. Valle. Handling inter-annotator agreement for automated skin lesion segmentation. arXiv preprint arXiv:1906.02415, 2019. 
[37] D. Rolnick, A. Veit, S. Belongie, and N. Shavit. Deep learning is robust to massive label noise. arXiv preprint arXiv:1705.10694, 2017. 
[38] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. pages 234–241, 2015. 
[39] J. Sivic and A. Zisserman. Video google: Efficient visual search of videos. In Toward category-level object recognition, pages 127–144. Springer, 2006. 
[40] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. Chiang, Z. Wu, and X. Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. arXiv preprint arXiv:1908.10454, 2019. 
[41] Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffery N Chiang, Zhihao Wu, and Xiaowei Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. Medical Image Analysis, page 101693, 2020. 
[42] E. Valle, M. Fornaciari, A. Menegola, J. Tavares, F. Vittencourt, L. T. Li, and S. Avila. Data, depth, and design: Learning reliable models for skin lesion analysis. Neurocomputing, 383:303–313, 2020. 
[43] D. Van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu. Scikit-image: image processing in python. PeerJ, 2:e453, 2014. 
[44] Y. Xue, T. Xu, and X. Huang. Adversarial learning with multi-scale loss for skin lesion segmentation. In International Symposium on Biomedical Imaging, pages 859–863, 2018.
[45] Y. Yuan. Automatic skin lesion segmentation with fully convolutional-deconvolutional networks. *arXiv preprint arXiv:1703.05165*, 2017.

[46] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2881–2890, 2017.