Ultrasound-inspired Adaptations for Multi-class Contrastive Segmentation

Rohit Singla RSINGLA@ECE.UBC.CA
Cailin Ringstrom CERINGSTROM@ECE.UBC.CA
Victoria Lessoway
Janice Reid
Robert Rohling ROHLING@ECE.UBC.CA
[1] *Electrical and Computer Engineering, 2332 Main Mall, Vancouver, Canada, V6T 1Z4*
Christopher Nguan CHRIS.NGUAN@UBCUROLOGY.COM
[2] *Urologic Sciences, 2775 Laurel St, Vancouver, Canada, V5Z 1M9*

Editors: Under Review for MIDL 2022

Abstract

Creating ground truth segmentations for medical imaging is a time-consuming and labour-intensive process. Contrastive learning techniques applied to modalities like computed tomography have showed promise in reducing these constraints. This study investigates the possible benefits of employing ultrasound-inspired adaptations to the contrastive learning paradigm. We begin by comparing the label efficiency of fully supervised and contrastive algorithms in a direct comparison. Then we use temporal similarity, which assumes that temporally close frames in an ultrasound video clip share structural similarities, to generate positive and negative pairs and evaluate the effects on accuracy. Finally, we study a loss function based on the Nakagami probability distribution to offer a speckle-based constraint on the learned embeddings. Our preliminary findings in kidney ultrasound suggest that both techniques have mixed results on segmentation accuracy. Future research will investigate optimal approaches to incorporate these contributions.

Keywords: Ultrasound, temporal similarity, speckle, contrastive learning, segmentation

1. Introduction

In medical image analysis, semantic segmentation is a fundamental task.(Litjens et al., 2017) It entails densely assigning a category to each pixel inside the image, providing structured spatial information. Although supervised learning techniques yield great segmentation accuracy, they require a large number of fine-grained annotations to train.(Ronneberger et al., 2015) In medical imaging, these labels are difficult to acquire. They require considerable work to annotate manually, are time demanding to generate, and require clinical expertise from healthcare professionals. Even then, inter-rater heterogeneity in image interpretation, and thus the ground truths, is considerable.(Nir et al., 2018; Ridge et al., 2016; Sahli et al., 2019). Alternatives to manually labeling large data sets have focused on novel learning algorithms, such as self-supervised learning and generative adversarial networks, as well as learned or physics-inspired data augmentations.(Pesteie et al., 2019) As a result, achieving high label efficiency-that is, maintaining high accuracy with a small number of labels-remains a major goal. Contrastive learning, a sort of self-supervised learning, is one strategy that appears to be promising in this regard.

In general, a contrastive learning network aims to learn a lower-embedding space in which embeddings of similar (positive) pairs are close together and dissimilar (negative) pairs are farther apart. To determine pairs, a sampling strategy is required. A common contrastive learning approach has three primary steps. The first is training a large task-agnostic feature extractor with unlabelled pairs. The second is fine-tuning the feature extractor with the available labelled data. Finally, knowledge distillation to a smaller network is performed. To generate pairs, augmentations of the original data may be used to create similar images. Augmentations commonly include colour enhancement, rotation, translation, warping and more. Modality-specific augmentations are an active area of interest within medical imaging such as in ultrasound, computed tomography, and magnetic resonance imaging.(Lee et al., 2021; Frid-Adar et al., 2018; Hao et al., 2020)

Recent years have seen a surge in interest in the application of contrastive learning to semantic segmentation (”contrastive segmentation”). For pre-training (Zhao et al., 2021) used a pixel-wise, label-based contrastive loss, (Alonso et al., 2021) used a memory bank and contrastive learning module, and (Xie et al., 2021) used a pixel-wise contrastive loss with the addition of a propagation consistency scheme. For medical imaging, (Pandey et al., 2021) used a consistency regularisation scheme to aid in contrastive segmentation, and (Chaitanya et al., 2020) demonstrated contrastive learning at both the global and local scales for volumetric medical images.

Ultrasonography is a relatively unexplored modality for contrastive learning at the moment. However, because it is non-invasive and does not employ ionising radiation, it is the imaging technique of choice for organs such as the kidney. Contrastive learning is used in ultrasound for a variety of purposes, including measuring the severity of COVID-19 (Xue et al., 2021), labelling in the prenatal setting (He et al., 2021) and classifying views in echocardiography (Chartsias et al., 2021). Ultrasound is unique in its ability to acquire video as a standard ultrasound scan is a two-dimensional (2D) video sequence. Sonographers benefit from differences between consecutive frames when interpreting scans. This is due to the fact that consecutive frames are similar. They share comparable but not identical structural information. Another characteristic of ultrasonography is speckle. Speckle is generated by tissue inhomogeneities as well as the transducer itself. Speckle is a predictable and non-random phenomena induced by the multiple directions in which an ultrasonic pulse scatters.

The purpose of this paper is to demonstrate the utility of ultrasound-inspired adaptations in the context of multi-class contrastive segmentation. We begin by comparing contrastive learning's labelling efficiency to that of a fully supervised network. Then, we focus two significant contributions: temporal similarity for positive-negative pair generation and an ultrasound speckle loss term that acts as an anatomical constraint. We utilize temporal similarity for sampling strategies in a contrastive architecture and systematically evaluate its effects. Finally, we introduce a speckle loss function to impose additional constraints on the learned embeddings. By incorporating these ultrasound-inspired adaptations, we want to create embeddings that are otherwise unaware of the intrinsic features of ultrasound.

2. Methods

2.1. Temporal Similarity

Given an input image i, an augmentation T may be applied.(Chaitanya et al., 2020) This generates the augmented $\widetilde{i}$ where $\widetilde{i} = T(i)$ which could be a considered a positive pair.(Chaitanya et al., 2020). T can be any number of augmentations, such as colour enhancement or geometric transformations. As ultrasound is a video sequence, rather than a single frame, an augmentation in the time domain can be considered. For an input image i_n at frame n, a transformed version of this frame could be at i_{n+d} where $n + d$ is some arbitrary index in the entire video sequence. Here, d is an augmentation parameter that describes the distance from i_n in the video sequence. We treat i_{n+d} as an augmented form of i_n, i.e. $i_{n+d} = T(i_n)$. Using this, new positive pairs and negative pairs can be determined.

The simplest approach is one where any frames within the range (i_{n-d}, i_{n+d}) (a total of $2d$ frames) are considered positive pairs. Any frames outside of this are considered negative pairs. We term this sampling strategy "temporal similarity" where the augmentation is using frames similar in time to one another. For our experiments, we arbitrarily set $d = 5$ frames as to be approximately 5% of the average video sequence length. This sampling approach can be used for several losses, such as the pairwise contrastive loss from (Chopra et al., 2005) or the InfoNCE loss in (Chaitanya et al., 2020).

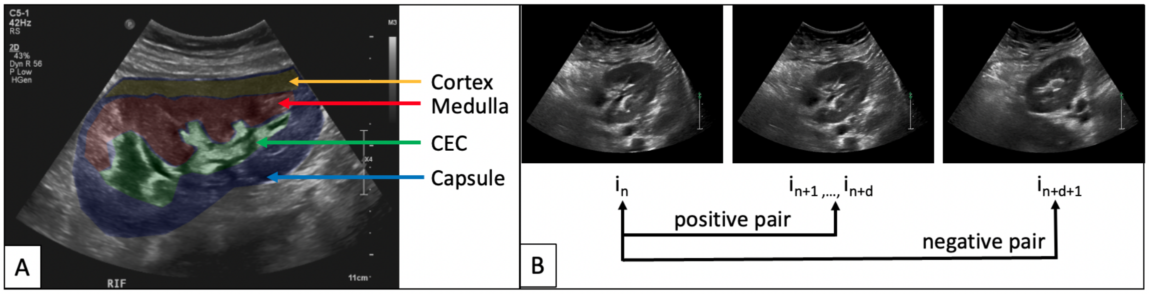

Figure 1: A: Illustrative figure of the different classes. In blue, the kidney capsule outlines the entirety of the organ. Within that class lie the cortex at the top (orange), followed by the medulla (red). In green lies the central echogenic complex (CEC) which is a combination of parts which cannot be delineated in ultrasound. B: An example of the temporal similarity sampling strategy for pair generation. Given an image i_n, any frames within a distance d can be selected as a positive pair. In the example images, the two frames are not the same. Images that are $d + 1$ and beyond can be sampled as negative pairs.

As an illustrative example, consider the common pairwise contrastive loss in Equation (1) from (Chopra et al., 2005) which receives as input a pair of images (i_1, i_2). Given a feature extractor $f(x)$ (i.e. a neural network), we can measure a similarity metric D between the representations of a given pair such as Euclidean distance or cosine similarity. With temporal similarity, if i_2 is within the m frames from the i_1, it is considered positive and the indicator y is set to 0. Otherwise, y is set to 1. For negative pairs, their representations should be farther than some threshold distance v.

$$L(i_1, i_2) = (1 - Y)\frac{1}{2}D(i_1, i_2)^2 + (Y)\frac{1}{2}max(0, v - D(i_1, i_2))^2 \tag{1}$$

2.2. Speckle Loss

Speckle is tissue-specific. We hypothesize that this property can aid in differentiating similar appearing structures, such as layers of tissue. The kidney for example has separate regions of interest such as the cortex and medulla which are subtly different visually. Incorporating speckle into machine learning may support learning improved delineations. To incorporate speckle, we first characterize it. The Nakagami distribution is a frequently used probability distribution for speckle in ultrasonography. Previously, it was demonstrated that the Nakagami distribution may be used characterize different regions in ultrasound.(Hu et al., 2019). In internal tests (using data separate from these experiments), the Nakagami distribution significantly differed between cortex and medulla regions with excellent goodness of fit. While different distributions can be used to simulate speckle, the Nakagami distribution has demonstrated greater versatility and offers efficient parameter estimators. The Nakagami distribution is defined by two parameters: the shape m and the scale Ω. A Nakagami continuous random variable x follows a probability distribution function as defined in Equation (2) where $\Gamma(*)$ is the Gamma function.

$$N(x) = \frac{2m^m}{\Gamma(m)\Omega^m} x^{2m-1} e^{\frac{-m}{\Omega}x^2} \tag{2}$$

For an image I, we may then estimate the Nakagami parameters for this image ${}_I$ and Ω_I using expectation E and variance Var as described Equation (3) and Equation (4).(Kolar et al., 2004)

$$m = \frac{E[I^2]^2}{Var[I^2]} \tag{3}$$

$$\Omega = E[I^2] \tag{4}$$

Using these parameters, we may obtain the new loss function L_s as in Equation (5), our second contribution. For I, a segmentation network produces predictions $\bar{I}$ for each class $c \in C$. The Nakagami parameters are computed for both I_c and $\bar{I}_c$ using the estimators. Finally, the Euclidean distance between the parameters is summed over all classes. If the predicted region and the ground truth region in a given class are identical, the differences in parameters should be close to zero. This loss term is added to existing loss functions, and each term is normalized for equal weighting.

$$L_s = \sum_{c \in C} (\sqrt{(m_{I_c} - m_{\bar{I}_c}) + (\Omega_{I_c} - \Omega_{\bar{I}_c})} \tag{5}$$

2.3. Data Set and Experiments

Three experiments on multi-class segmentation in kidney ultrasound are presented. Over a five-year period, images are acquired and anonymised from our local institution following our institution's Research Ethics Board approval. The kidney capsule, cortex, medulla and central echogenic complex are annotated with fine-grained polygons in 514 images. Two sonographers, each with over 20 years of expertise, manually generate annotations. For experiments using unlabelled data, we use 7000 ultrasound video sequences with an average of 200 frames. The typical 80/20 ratio between training and testing is adopted. The computations are carried out on a single GPU (NVIDIA Tesla V100 32GB)

As our baseline contrastive model, we use the network from (Chaitanya et al., 2020) for volumetric medical segmentation. This network involves two feature extraction steps at global and local scales using an encoder and a decoder respectively. Both use InfoNCE loss followed by a fine-tuning step with no knowledge distillation. As (Chaitanya et al., 2020) designed their network for volumes, we use time as the third dimension for ultrasound video clips. For our baseline fully supervised model, we use a nnU-net model from (Isensee et al., 2021) to represent the current state-of-the-art.

Experiment 1 - Baselines. We compare the Chaitanya's model with the nnU-net. The average Dice-Sørensen Coefficient (DSC) across all classes is presented. Both networks were trained with labels ranging from 1% to 100% of total labels in increments of 10%.

Experiment 2 - Temporal Similarity. We evaluate temporal similarity as a sampling strategy for both InfoNCE and pairwise contrastive loss. In a step-wise manner, we applied this strategy to first the encoder only, then decoder only, and finally both the encoder and decoder. We subsequently evaluated temporal similarity when including the negative pairs for training. All models were evaluated using $10\%, 50\%$, and 100% of labels.

Experiment 3 - Speckle Loss. We investigate whether speckle loss can be aid in the segmentation of three classes in the kidney: the cortex, the medulla, and the central echogenic complex. We present the DSC results for each class, as well as the differences in signal-to-noise ratio (ΔSNR) between the predicted labels and ground truths. ΔSNR is added due to the DSC's sensitivity to alterations in small regions of interest.(Reinke et al., 2021) Additionally, we discuss the differences between different activation functions. The contrastive learning network is trained with 10%, 50% and 100% of labels. A nnU-net using 100% of labels is provided for comparison.

3. Results

Experiment 1's results are presented in Table 1. Across all four classes, the nnU-net achieves an maximum average DSC of 0.57 across all classes at 60% of labels, with no difference when incorporating more. When only 20% of labels are used, Chaitanya's model reaches the same average DSC as the nnU-net with all labels. This supports the hypothesis that a contrastive segmentation network can reduce the label burden.

The findings of introducing temporal similarity are summarised in Table 2. When temporal similarity is used to produce only positive pairs, the average DSC increased when compared to the baseline at 10% of labels. At 50%, no model surpasses the baseline. This indicates a potential benefit for using nearby frames in the case of few labels, such as 50 in this case.

Finally, Table 3 summarises the speckle loss function experiments. Across 10%, 50%, or 100% of labels, the contrastive models see no benefit regardless where the loss function is used. The SNR differences remain similar when only 10% of labels are used, but exceed the baseline when 50% or 100% of labels are used. In comparison, when speckle loss is included, the nnU-net produces larger differences in SNR.

Table 1: Mean DSC of all four classes across differing amounts of labelled data utilized between nnU-net and baseline contrastive learning network. At 20%, the contrastive learning network approaches the maximum mean DSC from the nnU-net.

Network	**1%**	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
nnU-net	0.33	0.44	0.52	0.52	0.55	0.54	0.57	0.55	0.57	0.56	0.57
Chaitanya	0.28	0.50	0.56	0.57	0.56	0.58	0.58	0.58	0.57	0.56	0.57

Table 2: Evaluation of temporal similarity sampling strategies using positive pairs only on mean DSCs. The variations with the highest DSCs are bolded.

Loss	**Variation**	**Stage**	**10%**	**50%**	**100%**
InfoNCE	–	–	0.44	0.55	0.54
InfoNCE	Temporal Positive Pairs	Encoder	0.45	0.51	0.53
InfoNCE	Temporal Positive Pairs	Decoder	**0.47**	0.53	0.54
InfoNCE	Temporal Positive Pairs	Both	0.45	0.50	0.53
Contrastive	Temporal Positive Pairs	Encoder	0.43	0.53	**0.55**
Contrastive	Temporal Positive Pairs	Decoder	0.44	0.53	0.54
Contrastive	Temporal Positive Pairs	Both	0.43	**0.53**	0.53
Contrastive	Temporal Negative Pairs	Encoder	0.45	0.52	0.54
Contrastive	Temporal Negative Pairs	Decoder	0.43	0.52	0.54
Contrastive	Temporal Negative Pairs	Both	**0.46**	0.53	0.54

4. Discussion and Conclusion

We studied how two properties, temporal similarity and speckle, may be used to improve multi-class semantic segmentation. To begin, our preliminary findings support the notion that contrastive learning might be employed in lieu of fully supervised networks, which need massive amounts of labelled data. With only 20% of labelled data available for fine-grained, manually annotated segmentation, the contrastive learning strategy may alleviate the arduous human annotation burden associated with obtaining comparable results.

Second, utilising temporal similarity in the network from (Chaitanya et al., 2020) has mixed results in terms of increasing accuracy at a lower label percentage. Contrary to expectations, we see no improvements when sampling tactics include negative pairs. The

Table 3: The impact of speckle loss as a function of the percentage of labels used. Metrics are (DSC, ΔSNR.)

Loss	**Percent**	**Cortex**	**Medulla**	**CEC**	**All Classes**
Chaitanya *et al.*	10%	0.21, 0.34	0.29, 0.28	0.54, 0.13	0.42, 0.25
+Speckle	10%	0.04, 0.29	0.13, 0.22	0.44, 0.15	0.34, 0.22
Chaitanya *et al.*	50%	0.31, 0.30	0.38, 0.22	0.68, 0.10	0.55, 0.20
+Speckle	50%	0.20, **0.38**	**0.33**, **0.24**	0.57, **0.12**	0.48, **0.25**
Chaitanya *et al.*	100%	0.30, 0.31	0.38, 0.22	0.67, 0.31	0.54, 0.20
+Speckle	100%	0.19, **0.36**	0.23, **0.23**	0.55, **0.36**	0.45, **0.24**
nnU-net	100%	0.43, 0.24	0.46, 0.19	0.75, 0.08	0.57, 0.17
+Speckle	100%	0.37, **0.32**	0.28, **0.23**	0.29, **0.24**	0.39, **0.26**

introduction of "hard negative pairs", that is, pairs that are distinct but difficult to discriminate, has been shown to boost contrastive learning performance in other works; however, this is not the case here. Our sampling strategy however was simple; more principled strategies using nearby/far frames in a video sequence and across different video sequences may yield different results for single image segmentation.

Third, it appears as though using speckle loss improves the SNR of predicted annotations, but not the DSC measure. We acknowledge that this is a difficult data set to evaluate simply via DSC. Given the small size of the regions of interest, we performed a sensitivity analysis by adding a 1-pixel and 10-pixel erosion and dilatation of the ground truth mask and comparing it against the original. We found changes in DSC of around 1-2% at 1-pixel, and 20-30% at 10-pixels. When we evaluate these networks using the SNR, we see a constant improvement with the exception of networks examined with 10% of labels. This suggests that the network is producing regions that, while not accurate in shape, are improving in signal and less noisy.

Fourth, however, is that the speckle loss does not improve DSC results. This is counter-intuitive, given that the speckle is property of tissue, and penalizing the network for deviating from this property should provide better results. Recall that the Nakagami distribution is fitted onto the ground truth region of interest, and summarized as two parameters for the entire region. It could be possible that this loses relevant spatial information, and in doing so does not aid the network. Creating a Nakagami parameter map may be an alternative. It could also be that the loss term itself needs a different formulation to capture deviations from the ground truth Nakagami distribution, or that speckle information should be included in other form such as augmentation. As presented, the Euclidean distance between Nakagami parameters does not provide a meaningful benefit in DSC scores. Further work is needed to investigate how tissue properties like speckle can be effectively incorporated.

The Nakagami model employed in this study is one of several possible probability distributions for characterising speckle in ultrasound pictures. It is worthwhile to investigate alternative speckle models and how their performance varies. Other limitations of this experiment include the fact that only one organ, the kidney, was examined. Humans have difficulty delineating the regions of interest within the kidney, notably the cortical and medullary sections. It remains to be seen if classification, regression, and other segmentation tasks such as inter-organ segmentation may benefit from temporal similarity and speckle. Different organs have more different tissue structures, which can be defined by their change through time (ex: the heart in different stages of the cardiac cycle) or by their speckle content (ex: the liver). It remains to be observed how the addition of speckle loss affects learned representations, intra-class compactness, and inter-class separability.

In summary, we provide a detailed evaluation of temporal similarity and pioneer the use of speckle removal. We conduct three experiments systemically, totaling 60 models. Other machine learning techniques may benefit from similar ultrasound-inspired improvements. Incorporating temporal and speckle characteristics into machine learning techniques applied to ultrasound may be beneficial in a variety of ways, including data augmentation, data sampling, and the addition of novel components to networks.

Acknowledgments

The authors acknowledge funding from the Natural Sciences and Engineering Research Council of Canada (NSERC), the Kidney Foundation of Canada (KFOC), and the Vanier Canada Graduate Scholarship. The authors thank Tim Salcudean for infrastructure and support.

References

Inigo Alonso, Alberto Sabater, David Ferstl, Luis Montesano, and Ana C Murillo. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. *arXiv preprint arXiv:2104.13415*, 2021.

Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. *arXiv preprint arXiv:2006.10511*, 2020.

Agisilaos Chartsias, Shan Gao, Angela Mumith, Jorge Oliveira, Kanwal Bhatia, Bernhard Kainz, and Arian Beqiri. Contrastive learning for view classification of echocardiograms. In *International Workshop on Advances in Simplifying Medical Ultrasound*, pages 149–158. Springer, 2021.

Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005.

Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018.

Ruqian Hao, Khashayar Namdar, Lin Liu, Masoom A Haider, and Farzad Khalvati. A comprehensive study of data augmentation strategies for prostate cancer detection in diffusion-weighted mri using convolutional neural networks. *arXiv preprint arXiv:2006.01693*, 2020.

Shuangchi He, Zehui Lin, Xin Yang, Chaoyu Chen, Jian Wang, Xue Shuang, Ziwei Deng, Qin Liu, Yan Cao, Xiduo Lu, et al. Statistical dependency guided contrastive learning for multiple labeling in prenatal ultrasound. In *International Workshop on Machine Learning in Medical Imaging*, pages 190–198. Springer, 2021.

Ricky Hu, Rohit Singla, Farah Deeba, and Robert N Rohling. Acoustic shadow detection: study and statistics of b-mode and radiofrequency data. *Ultrasound in medicine & biology*, 45(8):2248–2257, 2019.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Radim Kolar, Radovan Jirik, and Jiri Jan. Estimator comparison of the nakagami-m parameter and its application in echocardiography. *Radioengineering*, 13(1):8–12, 2004.

Lok Hin Lee, Yuan Gao, and J Alison Noble. Principled ultrasound data augmentation for classification of standard planes. In *International Conference on Information Processing in Medical Imaging*, pages 729–741. Springer, 2021.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

Guy Nir, Soheil Hor, Davood Karimi, Ladan Fazli, Brian F Skinnider, Peyman Tavassoli, Dmitry Turbin, Carlos F Villamil, Gang Wang, R Storey Wilson, et al. Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts. *Medical image analysis*, 50:167–180, 2018.

Prashant Pandey, Ajey Pai, Nisarg Bhatt, Prasenjit Das, Govind Makharia, Prathosh AP, et al. Contrastive semi-supervised learning for 2d medical image segmentation. *arXiv preprint arXiv:2106.06801*, 2021.

Mehran Pesteie, Purang Abolmaesumi, and Robert N Rohling. Adaptive augmentation of medical data using independently conditional variational auto-encoders. *IEEE transactions on medical imaging*, 38(12):2807–2820, 2019.

Annika Reinke, Matthias Eisenmann, Minu D Tizabi, Carole H Sudre, Tim Rädsch, Michela Antonelli, Tal Arbel, Spyridon Bakas, M Jorge Cardoso, Veronika Cheplygina, et al. Common limitations of image processing metrics: A picture story. *arXiv preprint arXiv:2104.05642*, 2021.

Carole A Ridge, Afra Yildirim, Phillip M Boiselle, Tomas Franquet, Cornelia M Schaefer-Prokop, Denis Tack, Pierre Alain Gevenois, and Alexander A Bankier. Differentiating between subsolid and solid pulmonary nodules at ct: inter-and intraobserver agreement between experienced thoracic radiologists. *Radiology*, 278(3):888–896, 2016.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Zeyad T Sahli, Ashwyn K Sharma, Joseph K Canner, Farah Karipineni, Osama Ali, Satomi Kawamoto, Jen-Fan Hang, Aarti Mathur, Syed Z Ali, Martha A Zeiger, et al. Tirads interobserver variability among indeterminate thyroid nodules: A single-institution study. *Journal of Ultrasound in Medicine*, 38(7):1807–1813, 2019.

Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16684–16693, 2021.

Wufeng Xue, Chunyan Cao, Jie Liu, Yilian Duan, Haiyan Cao, Jian Wang, Xumin Tao, Zejian Chen, Meng Wu, Jinxiang Zhang, et al. Modality alignment contrastive learning for severity assessment of covid-19 from lung ultrasound and clinical information. *Medical Image Analysis*, 69:101975, 2021.

Xiangyun Zhao, Raviteja Vemulapalli, Philip Andrew Mansfield, Boqing Gong, Bradley Green, Lior Shapira, and Ying Wu. Contrastive learning for label efficient semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10623–10633, 2021.