Reducing Contextual Bias in Cardiac Magnetic Resonance Imaging Deep Learning Using Contrastive Self-Supervision

Makiya Nakashima NAKASHM2@CCF.ORG
Cardiovascular Innovation Research Center
Cleveland Clinic
Cleveland, OH, USA

Donna Salem SALAMD@CCF.ORG
Heart Vascular and Thoracic Institute
Cleveland Clinic
Cleveland, OH, USA

HW Wilson Tang TANGW@CCF.ORG
Heart Vascular and Thoracic Institute
Cleveland Clinic
Cleveland, OH, USA

Christopher Nguyen NGUYENC6@CCF.ORG
Cardiovascular Innovation Research Center
Cleveland Clinic
Cleveland, OH, USA

Tae Hyun Hwang HWANG.TAEHYUN@MAYO.EDU
Department of Artificial Intelligence and Informatics
Mayo Clinic
Jacksonville, FL, USA

Ding Zhao DINGZHAO@CMU.EDU
Department of Mechanical Engineering
Carnegie Mellon University
Pittsburgh, PA, USA

Byung-Hak Kim BHAK.KIM@CJ.NET
AI Center
CJ Corporation
Seoul, South Korea

Deborah Kwon KWOND@CCF.ORG
Heart Vascular and Thoracic Institute
Cleveland Clinic
Cleveland, OH, USA

David Chen CHEND3@CCF.ORG
Cardiovascular Innovation Research Center
Cleveland Clinic
Cleveland, OH, USA

Abstract

Applying deep learning to medical imaging tasks is not straightforward due to the variable quality and relatively low volume of healthcare data. There is often considerable risk that deep learning models may use contextual cues instead of physiologically relevant features to achieve the clinical task. Although these cues can provide shortcuts to high performance within a carefully crafted training set, they often lead to poor performance in real-world applications. Contrastive self-supervision (CSS) has recently been shown to boost performance of deep learning on downstream applications in several medical imaging tasks. However, it is unclear how much of these pre-trained representations are impacted by contextual cues, both known and unknown. In this work, we evaluate how CSS pre-training can produce not only more accurate but also more trustworthy and generalizable models for clinical imaging applications. Specifically, we evaluate the saliency and accuracy of deep learning models using CSS in contrast to end-to-end supervised training and conventional transfer learning from natural image datasets using an institutional specific and public cardiomyopathy cohorts. We find that CSS pre-training models not only improve downstream diagnostic performance in each cohort, but more importantly, also produced models with higher saliency in cardiac anatomy. Our code is available at https://github.com/makiya11/ssl_spur_cmr.

1. Introduction

The potential value of deep learning-based clinical decision support systems to improve outcomes, efficiency, quality, equity, and access is now widely acknowledged in the field of healthcare (Sutton et al., 2020; Miller, 2009). AI models have near or sometimes exceeded human cognitive performance in targeted clinical tasks such as anatomic segmentation (Naik et al., 2008), and disease classification (De Fauw et al., 2018). Despite successes *in silico*, deep learning-based models have not yet seen widespread clinical translation (Kelly et al., 2019) due in part to the continued skepticism in such models (Ghassemi et al., 2021; Schwartz et al., 2022).

One key aspect of that skepticism is the difficulty of elucidating the saliency or features used in such models. Conventional clinical decision support systems use well-validated biomarkers that have been proven to be causal or at the very least well-correlated with pathologies of interest. For example, the thickness of the myocardial ventricles is a known pathophysiologic response to cardiac amyloidosis (Maceira et al., 2008). When such models fail or do not generalize to new data, it is straightforward to diagnose the source of failure. However, it is not yet possible to explicitly learn such biologically connected features through deep learning. Therefore, when deep learning models fail, it is not always possible to leverage conventional clinical knowledge to understand the reasons for model failure.

In particular, models may leverage spurious contextual cues to achieve high performance during training, but will fail to generalize in practice. Oakden-Rayner et al. (2020) recently showed that models can use the presence of a chest drain to identify pneumothorax in the ChestX-Ray14 (Wang et al., 2017) dataset rather than true physiological changes in the lungs. Such bias towards contextual shortcuts are well-documented in natural image datasets where complex motions can be classified using very small number of images, contrary to our own understanding of motion (Xiao et al., 2020). With regards to clinical imaging, not all potential sources of bias are as clear as the presence of a chest drain. For example, Paschali et al. (2018) found that even when models trained using the same clin-

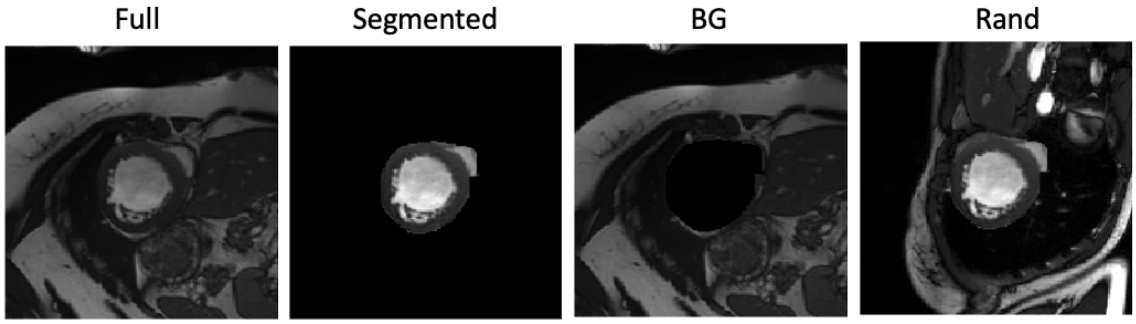

Figure 1: Example CMR images separating object of interest from background.

ical data, with comparable measured performance, may be sensitive to different patterns of adversarial attack (imperceptible noise). This gives evidence towards the fact that some shortcut cues are not always explainable via human perception.

Cardiac magnetic resonance imaging (CMR) encodes aspects of cardiac physiology, anatomy, and viability (Menacho et al., 2022). Like many natural images, CMR captures a significant amount of surrounding tissue as background to the central object of interest (the heart). However, cardiac imaging has significantly less variability in object characteristics, locations, and distribution compared to natural images. Therefore, it is not immediately clear how many unwanted contextual cues may be present in these surrounding tissue.

Generalizable Insights about Machine Learning in the Context of Healthcare

In this work, we aim to understand the importance of background tissue in CMR images for training deep neural networks given the potential of neural networks to leverage contextual features as shortcuts. We further investigate how the global representation of CMR trained using contrastive self-supervision (CSS)(Karthik et al., 2021) may be robust to such features.

Our work examines problems in the following ways:

- Demonstrate that CMR images contain contextual features outside of the heart may be used to shortcut learning for clinical tasks.
- Demonstrate saliency of the model can be used as a measure of the potential generalizability of the model.
- Demonstrate CSS increases saliency on organs of interest without explicit anatomic labels.

2. Related Work

Contextual Bias in Natural Images Implicit biases in data is a well-known problem in natural image machine learning datasets. Although models trained on biased datasets can yield high numerical metrics, poor generalizability in practice provides clear evidence that the model does not learn a meaningful representation of the intended class. For example, the classic problem of differentiating dogs versus wolves comes down to the snow in the background of wolves' pictures. Such models can have superhuman performance, but fail spectacularly when the object of interest is provided without the contextual cues

(snowy background) used in the training data. There has been extensive work to identify and mitigate such biases in natural images. Xiao et al. (2020) used semantic segmentation to randomly replace background to neutralize potential signals from correlated background information. This "randomization" of backgrounds as a form of data augmentation improved test accuracy and improved saliency on objects of interest. Mo et al. (2021) used a similar augmentation strategy except incorporating it within a CSS framework. Singh et al. (2020) takes a different approach, leveraging knowledge of the context to actively choose pairs of the same category but with different context to decouple object representation from its context. However, these methods assume that we explicitly understand background and foreground objects. Often, this is not immediately (Zhang et al., 2022) identifiable in medical imaging, where it is not unusual for radiologists to use contextual clues to make a diagnosis, nor is there a definitive object of interest.

Contextual Bias in Clinical Imaging There has been significant study on the sources of contextual bias in clinical imaging. As previously noted, Oakden-Rayner et al. (2020) recently showed that the presence of a chest drain can be used as a contextual shortcut to identify pneumothorax in chest X-rays, reducing the clinical utility where pneumothorax has not been previously detected. Unfortunately, the authors did not propose a method to address these shortcuts. Similarly, Jabbour et al. (2020) and Duffy et al. (2022) independently showed that deep learning models can accurately predict patient demographics (age, sex, body mass index) in chest X-rays and echocardiograms respectively. Such a finding would suggest that noncausal contextual cues for cardiac disease such as presence of high levels of subcutaneous fat or reduced bone density may impact the accuracy of disease discrimination. Resampling methods can be used to control for some of these biases (Reinhold et al., 2021) given that demographics are easily understood sources of confounding factors; however, other contextual biases such as the presence of chest tubes are more difficult to identify and account for.

3. Methods and Data

This work evaluated the impact of potential contextual cues from adjacent tissue in CMR on the performance of deep learning models. We then evaluated strategies to mitigate such problems. This work leveraged two CMR datasets comprising of short axis cine images targeted at left ventricular diseases. First, we created a cardiomyopathy (CM) dataset derived from patients from Cleveland Clinic main campus and have been previously studied without using CSS pre-training(Cockrum et al., 2022). Usage of this dataset for research purposes was approved by the Cleveland Clinic Institutional Review Board. Informed consent and Health and Insurance Portability and Accountability Act authorization were waived given the retrospective study design. Second, we validated the generalizability of our findings in the public Automated Cardiac Disease Challenge dataset which contained two classes of patients not seen in the CM dataset(Bernard et al., 2018).

3.1. Datasets

CM Dataset The **CM** dataset was constructed from adult patients who underwent a CMR exam between 2002 and 2021 at Cleveland Clinic main campus. All patients received

a standard CMR exam, with cine, and late gadolinium enhancement (LGE) imaging on a Phillips 1.5T Achieva or 3.0T Ingenia scanners, although only cine short axis images were included for this study. The dataset is comprised of patients with definitive diagnosis of a cardiomyopathy, which include undifferentiated non-ischemic cardiomyopathy (NICM), ischemic cardiomyopathy (ICM), cardiac amyloidosis (AMYL), and hypertrophic cardiomyopathy (HCM). The final diagnosis was identified through a chart review of all clinical data by clinical research fellows using the relevant clinical guidelines. A level 3 board-certified cardiologist reviewed the results for accuracy. There are in total of 1,742 studies included in this dataset; 412 ICM, 227 AMYL, 304 HCM, and 799 NICM. The mean age at the time of CMR was 56.57 ± 15.40. Within the 1,742 patients, 574 are female and 1,168 are male.

ACDC The ACDC dataset Bernard et al. (2018) is a public CMR dataset comprising of 150 clinical CMRs acquired at the University Hospital of Dijon, France acquired over a 6 year period on either a 1.5T Siemens Area and 3.0T Siemens Trio scanner. Short axis cines were acquired with in-plane spatial resolution ranging from 1.37 to 1.68 mm^2, and slice thickness of 5-8mm. The dataset includes two sets of labels/tasks; balanced multi-class disease classification (myocardial infarction with systolic heart failure, dilated cardiomyopathy, hypertrophic cardiomyopathy, abnormal right ventricle, and normal) and semantic segmentation of cardiac anatomy.

Segmentation The ACDC dataset includes 150 manually drawn contours of the left myocardial wall, the left ventricular blood pool, and the right ventricle. A commercial product (CVI42, Circle Cardiovascular Imaging, Calgary, Canada) was used to automatically generate the left myocardial wall, left ventricular blood pool, and right ventricle contours for short-axis cine frames in the internal institutional dataset. In internal testing, we found the results to be operationally viable and achieve > 0.95 DSC on the mid-ventricular short axis slice. A board-certified cardiologist with 10+ years of CMR reading experience reviewed the segments for quality assurance. We combined these contours into a single cardiac mask against all other background tissue.

3.2. Contrastive Self-Supervised Learning for Pre-training

Background randomization requires that we have cardiac contours that are typically expensive to acquire and not realistic to use at scale. Self-supervised learning has been shown to achieve generalizable data embeddings in many settings (Azizi et al., 2021; Jing and Tian, 2020; Chen et al., 2020; Kahn et al., 2018). However, it is not clear how such embeddings may be robust to contextual cues from the background. Therefore, we aim to demonstrate its application in clinical imaging.

Early methods of self-supervision relied on pretext tasks based on pseudo-labels. For example, one could use masked autoencoders to recover an image which has been perturbed in some way. However, it is sometimes difficult to identify a suitable task for a specific downstream problem and some tasks may introduce hallucination artifacts into the model (Cohen et al., 2018). CSS operates directly on the latent space by maximizing agreement between the learned representation of images from the dataset and augmented versions (e.g. crop, rotation, etc.). In general, any chosen network can be used as an encoder ($h_i = f(x_i)$). A projection head $g(h_i)$ is used to reduce the potential loss of information

induced by contrastive loss. The output of the projection head is then compared with the augmented version.

We evaluated three CSS frameworks for pre-training deep learning models: SimCLR (Chen et al., 2020), momentum contrast (MoCo) (He et al., 2020), and bootstrap your own latent (BYOL) (Grill et al., 2020). Both SimCLR and MoCo use a regularized cross-entropy loss as a measure of positive pair similarities as given below:

$$\mathcal{L}_{i,j} = -\log \frac{exp(cos(z_i, \hat{z_j})/\tau)}{\sum_{k=1}^{2N} exp(cos(z_i, \hat{z_j})/\tau)} \tag{1}$$

The cosine similarities of the projected features of the positive pair in the numerator are normalized by the sum of the similarities between all negative pairs (all other samples in the mini-batch) in the denominator. τ is the temperature hyperparameter which controls local separation and global uniformity. MoCo leverages the same overall learning paradigm but avoids the need for large batch sizes by keeping prior embeddings in a dictionary. BYOL makes no comparison with other samples in each mini batch (and therefore does not assume any negative samples), but uses an exponential moving average and a separate projection head to ensure the two encoders do not arbitrarily converge.

4. Experiments

4.1. Model Training and Evaluation

All DL models were developed in Python (3.9) using PyTorch (Paszke et al., 2017). We randomly split each dataset into 70/30 train and test sets. All models were trained using Adam optimizer with a weight decay of $1e^{-4}$, and sparse categorical cross-entropy loss. We used the TorchVision (maintainers and contributors, 2016) implementation of DenseNet121.

The accuracy of the model was evaluated using micro area under the receiver operator curve (AUC). AUC confidence interval was provided using bootstrapping. We evaluated the overlap of the model features with anatomy using GradCAM to visualize the saliency of the model (Selvaraju et al., 2017). We measure the overlap of model saliency with anatomy using both Dice-Sorensen Score (DSC) and the following proposed metric which we termed the Anatomic Knowledge Score (AKS):

$$AKS_{x,y} = \frac{\sum x \cdot y}{\sum x} \tag{2}$$

Where x is the continuous GradCAM output and y is the binary mask of the heart. The AKS is zero when there is no overlap, and one if model saliency falls entirely within the boundaries of the heart. We report this metric rather than just the Dice-Sorensen Score (DSC) (Sorensen, 1948) or the Jaccard score (Jaccard, 1908) as we did not want to penalize cases where model activations focused on a small area in the heart, such as the common case of focal ischemic lesions.

4.2. Evaluating Impact of Background Tissue on Classification

We first examined whether the background tissue in the CMR images may contain contextual clues and evaluated the magnitude of impact on the accuracy of the classification.

This was done by comparing models trained using the whole image against those with backgrounds masked using the cardiac mask and also a model trained using only the background tissue. We then used previously proposed background swapping method proposed by Xiao et al. (2020) to minimize the effect of contextual cues background signal by randomly switching backgrounds within a mini-batch as a form of data augmentation in the training process. We further evaluated the impact of contextual clues on the trained representation on the generalizability of the model on the public ACDC dataset with similar but still unique classes. We used GradCAM (Selvaraju et al., 2017) to investigate the effect of such efforts on model saliency. We quantified the change in overlap between model saliency and heart to quantify the potential amount of discriminative information represented in tissue outside the heart.

4.3. Linking Model Saliency to Model Performance

Next, we investigated how the magnitude of overlap between model saliency and the heart may inform classification performance. Model saliency has been proposed as a way to provide explainability to downstream users. We hypothesize that a good representation of the underlying data will result in increased saliency of the object of interest. Therefore, we evaluated how model saliency is linked with model performance in CMR tasks to examine if models with significant saliency outside of the heart have systematically lower measures of accuracy. Specifically, we compare the average overlap of our trained models against model AUC in test sets. We also look to examine how well models with different levels of average saliency overlap may help inform generalizability to unseen data. This was done by correlating in-domain (CM) overlap with out-of-domain (ACDC) AUC.

4.4. Evaluate Effects of Pretraining on Robustness Against Contextual Cues

We finally examined how CSS may impact the model's robustness to contextual cues. We trained models on the CM data using SimCLR, MoCo, and BYOL frameworks described previously. We then evaluated how these pre-trained embeddings generalize to both in-domain (CM) and out-of-domain (ACDC) data. The saliency of these CSS pretrained models was also compared against their fully-supervised analogs using AUC and overlap metrics.

5. Results

5.1. Existence of Contextual Bias in CMR

There are significant discriminative features in adjacent tissue Table 1 shows the results of training models in the full images, just the heart, just the background, and images with randomized backgrounds compared to linear probe using just ImageNet pretrained weights. Unsurprisingly, the models trained using just the heart evaluated on the in-set test set on average achieved higher AUC. The model pre-trained using the CM dataset achieved better generalizability to ACDC dataset compared to ImageNet pre-training. Models trained using the full image yielded lower generalization performance compared to those trained using only heart, which suggests that adjacent tissue can be a source of misclassification. The results of models trained using only background images give further evidence

Table 1: AUC, AKS, and DSC metrics on supervised classification experiments. Zero - Linear probe on ImageNet weights; Full - Finetuned using full image; Seg - Finetuned using only the heart; BG - Finetuned using only the backgrounds; Rand - models finetuned with randomized backgrounds. Heart only images (Seg) produces higher AUCs, suggesting the existence of contextual cues in the background. This is supported by the results using only BG images which achieved lower, but non-random classification results. Randomizing (Rand) the background is not effective for inset classification performance, but does improve generalizability. AKS and DSC metrics positively follow generalizability to ACDC.

Metric	Dataset	Zero	Full	Seg	BG	Rand
AUC	CM	0.728±0.058	0.833±0.047	0.806±0.050	0.784±0.053	0.801±0.052
	ACDC	0.701±0.189	0.838±0.144	0.990±0.010	0.746±0.183	0.714±0.147
	CM→ACDC	0.869±0.130	0.850±0.133	0.884±0.115	0.773±0.176	0.884±0.108
AKS	CM	0.076±0.053	0.115±0.072	0.162±0.095	0.074±0.057	0.123±0.077
	ACDC	0.053±0.056	0.100±0.096	0.102±0.109	0.093±0.097	0.103±0.108
	CM→ACDC	0.059±0.076	0.072±0.087	0.095±0.105	0.076±0.086	0.132±0.122
DSC	CM	0.119±0.011	0.173±0.014	0.243±0.014	0.120±0.012	0.184±0.015
	ACDC	0.095±0.014	0.168±0.029	0.168±0.043	0.153±0.036	0.164±0.048
	CM→ACDC	0.098±0.037	0.118±0.041	0.150±0.049	0.126±0.036	0.206±0.048

that there are significant contextual cues contained in the adjacent tissue. The influence of the surrounding tissue can be large ($\Delta 0.124$) as demonstrated by the difference between full images and segmented images.

Conventional methods of addressing contextual cues are not effective Our attempts to de-correlate any potential background signal with the anatomy by randomizing (Rand) the background during training achieved mostly negative results. Neither the models for the CM or ACDC were able to outperform training using the full image. However, the Rand model did provide better generalizability to the ACDC dataset. We hypothesize two reasons for the overall poor performance of background randomization. First, there are not significant differences in textural features between the heart and its adjacent tissue like there typically is in natural images given the bias convolutional architectures have towards global textural features (Geirhos et al., 2018; Li et al., 2020). Second, radiologists may use perceptual cues in the background tissue for diagnostic purposes Geirhos et al. (2020); Williams and Drew (2019). For instance, a radiologist may leverage the swirling patterns in the blood pool to diagnose aortic valve disease despite those patterns being a result of a magnetic resonance physics artifact. Removing these potential cues may negatively impact the model's ability to learn as well.

5.2. Saliency and Generalizability

Supervised learning results in poor saliency with anatomy No models achieved good correlation between model saliency and anatomy (Table 1) as evaluated by either AKS or DSC metrics. However, both metrics of overlap between model saliency and anatomy (AKS and DSC) followed the general trend of positive correlation with AUC. Of interest is

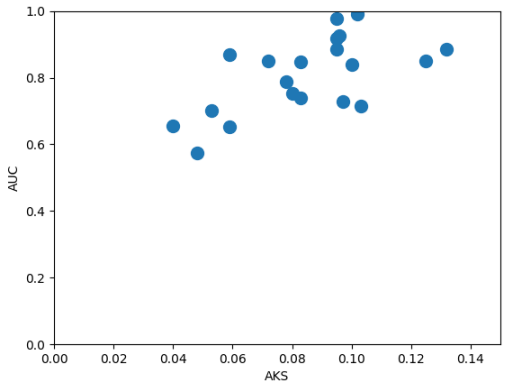

Figure 2: Correlation as a measure of generalizability. Correlation between AKS and AUC in the out-of-set test set (ACDC).

that AKS is well correlated with the AUC in out-of-domain dataset (Figure 2). This trend suggests that saliency can be used as another metric to evaluate generalizability.

5.3. Contrastive Learning to Reduce Contextual Bias

CSS improves classification accuracy over supervised learning We report the impact of CSS frameworks on model accuracy and saliency in Table 2. We found that CSS frameworks were able to achieve higher AUCs compared to using either the full image or segmented image. Furthermore, we found that CSS was able to achieve better generalizability to the ACDC dataset compared to supervised pre-training despite large similarities in classes. There was not a consistent best CSS paradigm although all CSS paradigms showed improved results compared to end-to-end training. This result would suggest that CSS may do a better job of learning anatomically important features compared to naïve end-to-end training.

Table 2: Pretrained models with full images can achieve similar results to non-pretrained models using segmented images. There is clear improvement using in-domain CSS pretraining compared to end-to-end supervised models. The improvement holds true for both our proposed overlap metric and DSC.

	Dataset	**Full**	**Seg**	**SimCLR**	**MoCo**	**BYOL**
AUC	CM	0.833±0.047	0.806±0.050	0.875±0.042	0.835±0.047	0.863±0.043
	CM→ACDC	0.850±0.133	0.884±0.115	0.978±0.022	0.919±0.078	0.925±0.075
AKS	CM	0.115±0.072	0.162±0.095	0.094±0.062	0.102±0.066	0.087±0.058
	CM→ACDC	0.072±0.087	0.095±0.105	0.095±0.102	0.095±0.098	0.096±0.102
DSC	CM	0.173±0.014	0.243±0.014	0.144±0.012	0.155±0.013	0.133±0.012
	CM→ACDC	0.118±0.041	0.150±0.049	0.158±0.039	0.158±0.035	0.158±0.040

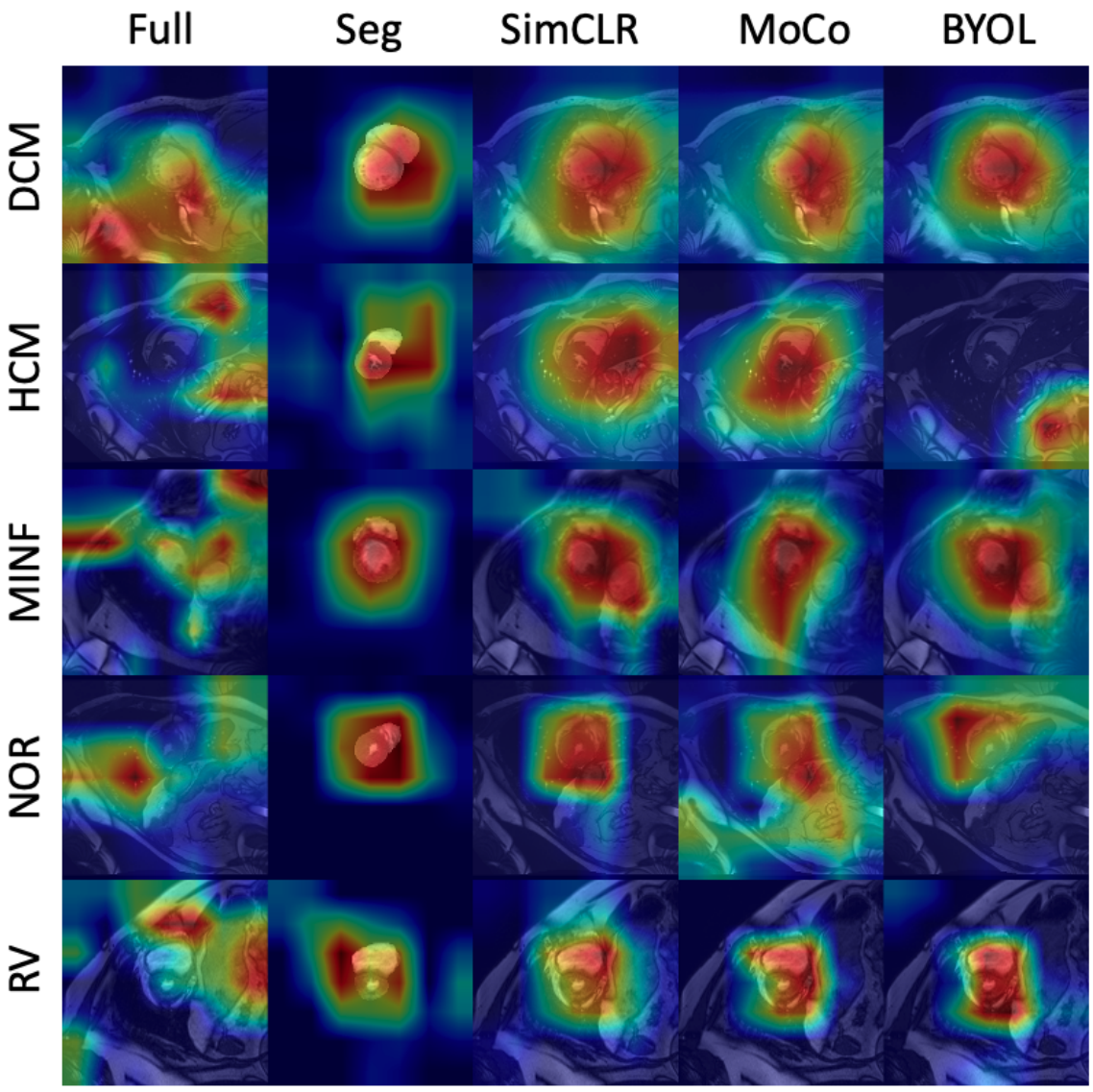

Figure 3: Example GradCAM saliency of CSS models compared with models trained end-to-end on the ACDC dataset. There are less activations outside of the heart with all 3 CSS paradigms (SimCLR, MoCo, BYOL) compared to end-to-end supervised training (None and ImageNet).

Overlap between model saliency and anatomy increased with CSS The improvement in model salience is correlated with the improvement in classification performance (Figure 2). The aggregated AKS and DSC showed improvement with CSS compared to the supervised pre-training using the full image in the external ACDC dataset, which is also visually demonstrated by the improved saliency in Figure 3. However, CSS yielded lower AKS and DSC values in the internal CM dataset at first glance. However, this result may be due to the unique pathophysiology of one of the classes of the cohort. Specifically, AMYL is a disease which results from deposition of misformed proteins systemically. Therefore, it is reasonable to expect that a model may leverage information outside of the heart to discriminate between classes. We find this to be the case when examining class specific AKS and DSC metrics in Table 3. The AMYL class reduced AKS and DSC metrics following CSS compared to the other classes. The combination of improved classification performance and improved model saliency overlap with anatomy in out-of-set tasks would suggest that CSS is heavily responsible for improved model saliency. We postulate that such training can make a model both more generalizable and more trustworthy.

Table 3: Overlap metrics by disease in the CM cohort.

Metric	Disease	Full	Seg	SimCLR	MoCo	BYOL
AKS	NICM	0.115±0.132	0.196±0.118	0.125±0.109	0.125±0.118	0.116±0.112
	ICM	0.101±0.120	0.159±0.107	0.092±0.102	0.112±0.112	0.082±0.086
	AMYL	0.113±0.053	0.131±0.097	0.052±0.059	0.080±0.094	0.054±0.053
	HCM	0.071±0.046	0.091±0.102	0.042±0.040	0.038±0.049	0.033±0.052
DSC	NICM	0.182±0.177	0.286±0.157	0.186±0.153	0.187±0.161	0.172±0.148
	ICM	0.166±0.166	0.241±0.160	0.140±0.145	0.170±0.160	0.129±0.128
	AMYL	0.172±0.081	0.211±0.151	0.087±0.097	0.125±0.119	0.091±0.086
	HCM	0.068±0.077	0.143±0.161	0.072±0.070	0.065±0.080	0.054±0.084

6. Discussion

In this work, we show that CSS not only improves overall trained model accuracy, but also improves model saliency to correlate better with physiology in downstream tasks (and thereby more closely associated with human saliency) in two different CMR datasets with three different CSS paradigms. Despite the published successes of many deep learning algorithms for clinical tasks, there remains significant skepticism around the application of such models in clinical practice (Miotto et al., 2018). Part of the skepticism revolves around the difference in how deep learning algorithms arrive at decisions compared to human decision making.

Contextual cues can have particularly severe impact on the generalizability of a model due to shortcut learning. We find that many models, especially ones trained using small to moderate volumes of healthcare data have high levels of saliency outside of what normally considered to be physiologically plausible. This would suggest that the models are learning contextual cues rather than true features discriminative of disease. One avenue to reduce the impact of these contextual cues is by explicitly incorporating prior knowledge of anatomy through segmentation. However, acquiring such data is laborious. This work gives evidence

that CSS may reduce the impact of these contextual cues within CMR images compared to conventional supervised learning. The more focused saliency of CSS trained models would suggest these models are not as dependent on contextual cues lying within the surrounding tissue.

However, focusing model saliency on cardiac tissue is not always desirable as the pathophysiology of many diseases are much wider in scope. Although CMR is used to prognosticate the effects of amyloidosis on heart function, the pathophysiology of amyloidosis effects all organs in the body. Therefore, it is no surprise that extra-cardiac tissue may be strongly discriminative. This would suggest that making indiscriminate background augmentations to maximize focus on the heart would be harmful in these systemic diseases. Just as in natural images where humans also use contextual clues for object discrimination, there needs to be an open discussion on how contextual clues outside of biological understanding may impact clinical decision making.

Limitations One major limitation of model saliency methods is their inadequacy in providing true explainability for computer vision models. Various works (Arun et al., 2021; Adebayo et al., 2018) have shown a wide variety of saliency methods to be insensitive (or perhaps overly sensitive) to either the model or training process. The inadequacy of existing model explainability frameworks may call into question the true magnitude of CSS on model trustworthiness. However, we believe that the proposed evaluation framework may introduce another facet of trust in future pretraining methods, particularly within the medical imaging applications. More studies are needed with respect to the generalizability of the findings and to validate the linkage between improved saliency and real physiologic features. Overall, the findings in this work suggest the usefulness of CSS in not only improving macro performance, but also improving more subtle aspects of model performance such as robustness to shortcut learning, which is crucial to adoption of and trust in deep learning-based clinical decision support systems.

Acknowledgments

This work was supported by a kind gift from AKASA Inc. We would also like to thank Dr. Richard Grimm and Charles and Loraine Moore Endowed Chair in Cardiovascular Imaging for funding the powerful infrastructure necessary for this work and the Software Development/Data Science team in Imaging Informatics for facilitating transfer of imaging studies to our computational platform.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

Nishanth Arun, Nathan Gaw, Praveer Singh, Ken Chang, Mehak Aggarwal, Bryan Chen, Katharina Hoebel, Sharut Gupta, Jay Patel, Mishka Gidwani, et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence*, 3(6):e200267, 2021.

Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, et al. Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3478–3488, 2021.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11): 2514–2525, 2018.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

Joshua Cockrum, David Chen, Makiya Nakashima, Joseph Mauch, Mazen A Hanna, Kaia Kanj, Basnet Ramesh, Mitchel Benovoy, Samir R Kapadia, Lars G Svensson, et al. Deep learning analysis using cardiovascular magnetic resonance imaging for risk prediction in cardiac amyloidosis. *Journal of the American College of Cardiology*, 79(9_Supplement): 1193–1193, 2022.

Joseph Paul Cohen, Margaux Luck, and Sina Honari. Distribution matching losses can hallucinate features in medical image translation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pages 529–536. Springer, 2018.

Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342–1350, 2018.

Grant Duffy, Shoa L. Clarke, Matthew Christensen, Bryan He, Neal Yuan, Susan Cheng, and David Ouyang. Confounders mediate ai prediction of demographics in medical imaging. *npj Digital Medicine*, 5(1):188, December 2022. ISSN 2398-6352. doi: 10.1038/s41746-022-00720-8. URL https://doi.org/10.1038/s41746-022-00720-8.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L. Beam. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11):e745–e750, 2021. ISSN 2589-7500. doi: 10.1016/S2589-7500(21)00208-9.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, and Mohammad Gheshlaghi Azar. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

Sarah Jabbour, David Fouhey, Ella Kazerooni, Michael W Sjoding, and Jenna Wiens. Deep learning applied to chest x-rays: Exploiting and preventing shortcuts. In *Machine Learning for Healthcare Conference*, pages 750–782. PMLR, 2020.

Paul Jaccard. Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.*, 44:223–270, 1908.

Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43 (11):4037–4058, 2020.

Gregory Kahn, Adam Villaflor, Bosen Ding, Pieter Abbeel, and Sergey Levine. Self-supervised deep reinforcement learning with generalized computation graphs for robot navigation. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 5129–5136. IEEE, 2018.

Ananya Karthik, Mike Wu, Noah Goodman, and Alex Tamkin. Tradeoffs between contrastive and supervised learning: An empirical study. *arXiv preprint arXiv:2112.05340*, 2021.

Christopher J. Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(1):195, October 2019. ISSN 1741-7015. doi: 10.1186/s12916-019-1426-2. URL https://doi.org/10.1186/s12916-019-1426-2.

Yingwei Li, Qihang Yu, Mingxing Tan, Jieru Mei, Peng Tang, Wei Shen, Alan Yuille, and Cihang Xie. Shape-texture debiased neural network training. *arXiv preprint arXiv:2010.05981*, 2020.

Alicia M. Maceira, Sanjay K. Prasad, Philip N. Hawkins, Michael Roughton, and Dudley J. Pennell. Cardiovascular magnetic resonance and prognosis in cardiac amyloidosis. *Journal of Cardiovascular Magnetic Resonance*, 10(1):54, November 2008. ISSN 1532-429X. doi: 10.1186/1532-429X-10-54.

TorchVision maintainers and contributors. Torchvision: Pytorch's computer vision library, 2016.

Katia Devorha Menacho, Sara Ramirez, Aylen Perez, Laura Dragonetti, Diego Perez de Arenaza, Diana Katekaru, Violeta Illatopa, Sara Munive, Bertha Rodriguez, and Ana

Shimabukuro. Improving cardiovascular magnetic resonance access in low-and middle-income countries for cardiomyopathy assessment: rapid cardiovascular magnetic resonance. *European Heart Journal*, 2022.

Randolph A. Miller. Computer-assisted diagnostic decision support: history, challenges, and possible paths forward. *Advances in Health Sciences Education*, 14(1):89–106, September 2009. ISSN 1573-1677. doi: 10.1007/s10459-009-9186-y.

Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T. Dudley. Deep learning for healthcare: review, opportunities and challenges. *Brief Bioinform*, 19(6): 1236–1246, November 2018. ISSN 1477-4054 (Electronic) 1467-5463 (Linking). doi: 10.1093/bib/bbx044.

Sangwoo Mo, Hyunwoo Kang, Kihyuk Sohn, Chun-Liang Li, and Jinwoo Shin. Object-aware contrastive learning for debiased scene representation. *Advances in Neural Information Processing Systems*, 34:12251–12264, 2021.

Shivang Naik, Scott Doyle, Shannon Agner, Anant Madabhushi, Michael Feldman, and John E. Tomaszewski. Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology. pages 284–287, May 2008. ISBN 1945-8452. doi: 10.1109/ISBI.2008.4540988.

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, pages 151–159, 2020.

Magdalini Paschali, Sailesh Conjeti, Fernando Navarro, and Nassir Navab. Generalizability vs. robustness: adversarial examples for medical imaging. *arXiv preprint arXiv:1804.00504*, 2018.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Jacob C Reinhold, Aaron Carass, and Jerry L Prince. A structural causal model for mr images of multiple sclerosis. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pages 782–792. Springer, 2021.

Jessica M. Schwartz, Maureen George, Sarah C. Rossetti, Patricia C. Dykes, Simon R. Minshall, Eugene Lucas, and Kenrick D. Cato. Factors influencing clinician trust in predictive clinical decision support systems for in-hospital deterioration: Qualitative descriptive study. *JMIR Hum Factors*, 9(2):e33960, May 2022. ISSN 2292-9495. doi: 10.2196/33960.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Krishna Kumar Singh, Dhruv Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feiszli, and Deepti Ghadiyaram. Don't judge an object by its context: Learning to overcome contextual bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11070–11078, 2020.

Thorvald A. Sorensen. 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.*, 5:1–34, 1948.

Reed T. Sutton, David Pincock, Daniel C. Baumgart, Daniel C. Sadowski, Richard N. Fedorak, and Karen I. Kroeker. An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digital Medicine*, 3(1):17, February 2020. ISSN 2398-6352. doi: 10.1038/s41746-020-0221-y.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

Lauren H Williams and Trafton Drew. What do we know about volumetric medical image interpretation?: A review of the basic science and medical image perception literatures. *Cognitive Research: Principles and Implications*, 4:1–24, 2019.

Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020.

Shijie Zhang, Lanjun Wang, Lian Ding, Senhua Zhu, and Dandan Tu. Intrinsic bias identification on medical image datasets. *arXiv preprint arXiv:2203.12872*, 2022.