

BackMix: Mitigating Shortcut Learning in Echocardiography with Minimal Supervision

Kit M. Bransby^{1,2}, Arian Beqiri¹, Woo-Jin Cho Kim¹, Jorge Oliveira¹, Agisilaos Chartsias¹, Alberto Gomez¹

¹ Ultromics Ltd., Oxford, United Kingdom
alberto.gomez@ultromics.com

² Queen Mary University of London, United Kingdom

Abstract. Neural networks can learn spurious correlations that lead to the correct prediction in a validation set, but generalise poorly because the predictions are right for the wrong reason. This undesired learning of naive shortcuts (Clever Hans effect) can happen for example in echocardiogram view classification when background cues (e.g. metadata) are biased towards a class and the model learns to focus on those background features instead of on the image content. We propose a simple, yet effective random background augmentation method called BackMix, which samples random backgrounds from other examples in the training set. By enforcing the background to be uncorrelated with the outcome, the model learns to focus on the data within the ultrasound sector and becomes invariant to the regions outside this. We extend our method in a semi-supervised setting, finding that the positive effects of BackMix are maintained with as few as 5% of segmentation labels. A loss weighting mechanism, wBackMix, is also proposed to increase the contribution of the augmented examples. We validate our method on both in-distribution and out-of-distribution datasets, demonstrating significant improvements in classification accuracy, region focus and generalisability. Our source code is available at: <https://github.com/kitbransby/BackMix>

Keywords: shortcut learning · echocardiography · augmentation

1 Introduction

Echocardiography (echo) is one of the primary cardiovascular imaging modalities used to study the structures and function of the heart from a variety of cross-sectional views (Figure 1). Classification of the view is a necessary first step in automated echo analysis as views are not labelled during acquisition, and are required for reliable interpretation [17]. Furthermore, the development of an accurate automated classifier is challenging due to the extensive time required to manually label studies for training data, which are often several thousand frames in total. Several convolutional neural network based methods have been used to create echo view classifiers [17, 18, 10, 15] demonstrating excellent performance.

An echo video consists of ultrasound images framed within a triangle or

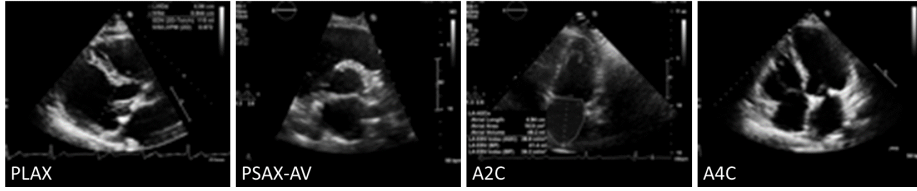


Fig. 1. Examples of echocardiograms: Unwanted text and measurement data can be seen in regions outside the ultrasound sector. View label superimposed on bottom left.

trapezoid known as the ultrasound “sector”, set against a black background with patient and acquisition metadata overlaid on top. Metadata outside of the sector can spuriously correlate with the ultrasound view classification label, leading neural networks to focus upon these features instead, as visualised in the GradCAM [16] feature attribution heatmaps of Figure 2. Such shortcuts allow learning of simple decision rules, thus limiting the classifier’s capacity to build accurate and trustworthy heart representations. This behaviour can be difficult to detect as test data used for model validation are typically drawn from the same distribution (i.d) as training data, and can therefore leverage shortcuts leading to artificially high performance. Once deployed in the wild however, the quality of the view classifier may deteriorate when it encounters out of distribution (o.o.d) images from different medical sites, acquisition protocols or scanner manufacturers, all of which are non-patient specific and can affect metadata.

A standard preprocessing step is to remove the area outside the sector using image segmentation, as for instance demonstrated in the CAMUS dataset [12]. However, this is limited by unreliable performance in edge cases, may require training additional networks, and adds significant computation time during inference. Another perspective involves leveraging attention, which implicitly learns to focus on image regions [4]. However without direct supervision, attention networks may also focus on spurious correlations to minimise the objective function during training. Ma et al. [14] address this in a natural image setting by supervising with saliency maps learnt from eye-gaze data. Despite excellent performance, this model requires an encoder-decoder as part of a multi-model pipeline, which adds to computational expense. Bassi et al. [3] integrate an interpretable layer-wise relevance propagation (LRP) [2] module, which estimates the contribution of each pixel by using back-propagated gradients. This approach does not change the architecture and adds minimal computation; however LRP attention maps are not perfect representations, and can be noisy and non-discriminative [9].

We propose a simple yet effective random background augmentation method called BackMix, which encourages a classification network to focus on the area inside the ultrasound sector. We initially split a training image into sector and background regions, and then randomly replace the background with a background from another training example. By making the background uncorrelated to the outcome, the model learns to ignore the background and becomes invariant to the spurious regions. Our method has the advantage of not adding any

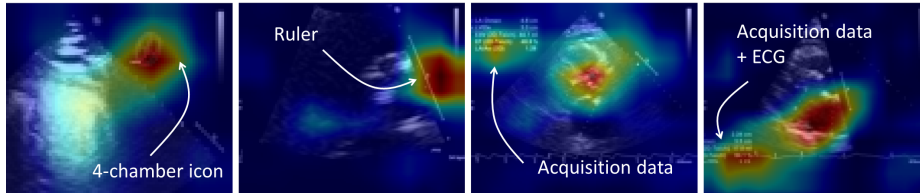


Fig. 2. Examples of shortcut learning in TMED dataset induced by metadata, which can be specific to the manufacturer, software and operator, but not the patient.

parameters or architectural changes, nor does it incur any additional training or inference time. The closest works to ours have applied background blurring [13], foreground in-painting for segmentation [5] and image synthesis [21] in natural image settings, however ours explores the impact on o.o.d performance and is the first to be applied to non-natural medical images.

Similarly to the above learning-based methods, training requires segmentation masks that separate ultrasound sectors from backgrounds prior to applying any augmentation. To avoid the need of acquiring segmentation masks for all training data, we extend our method to a semi-supervised setting where BackMix is only applied to a fraction of the training data. The positive effects of BackMix are maintained when as little as 5% of the training dataset is used in augmentation. Our methodology is strengthened by re-weighting the classification loss at an example level, so that a higher loss is assigned to examples that use BackMix. Despite such minimal supervision we achieve a significant improvement in performance on an o.o.d dataset against a baseline classifier trained without BackMix. We evaluate our hypothesis through GradCAM [16] analysis to show both quantitatively and qualitatively that our model focuses more on the ultrasound data within the sector and ignores spurious features.

Contributions: We identify that shortcut learning of background metadata harms generalisability in echocardiogram view classification and propose an effective background mixing augmentation called BackMix. We explore a semi-supervised setting and demonstrate that minimal numbers of segmentation masks are required for significant improvements in classification and focus metrics. Our method is strengthened with wBackMix, which emphasises examples with random backgrounds by appropriately re-distributing the loss. We show that our method removes the need for background removal in inference, a common and computationally expensive requirement. Finally, we propose two metrics to quantitatively evaluate how much the ultrasound sector affects the prediction label.

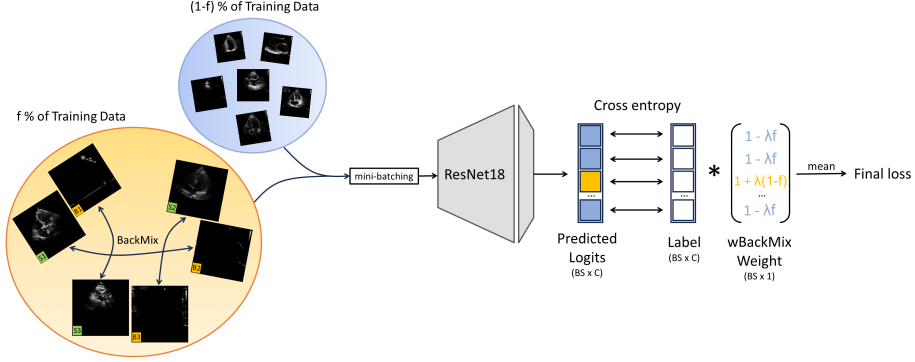


Fig. 3. Training schematic: Backgrounds are shuffled between a subset of training examples, and the prediction loss re-weighted in favour of these examples

2 Methods

2.1 BackMix

We implement an augmentation method, tailored to ultrasound data, called BackMix, which randomly swaps backgrounds between images in the training set. During training, we separate an image i into sector S_i and background B_i using a segmentation mask M_i and in-paint the empty areas with zeros (value of background). An additional frame j is then randomly sampled and the same process is applied to yield S_j and B_j . S_i is superimposed onto B_j to replace $S_i B_i$ and synthesise a new sample $S_i B_j$ that is used for training. This augmentation process is illustrated in Figure 3.

We apply BackMix augmentation after standard augmentations, such as random image rotations in the range $[-30^\circ, +30^\circ]$, brightness-contrast adjustment, and horizontal flipping. BackMix is evaluated when training a ResNet18 [6] network, that is selected due to its widespread use and computational efficiency, although there is no restriction in the choice of backbone network architecture. During inference there is no need for segmentation masks. As demonstrated in Section 3, training with BackMix augmentations encourages the classification network to focus in the ultrasound sector.

2.2 Semi-Supervised classification

Pixel-wise segmentation labels of the ultrasound sector are time consuming and difficult to obtain as the sector boundaries are not always well defined. We thus explore a semi-supervised approach, where only a fraction f of the training dataset has sector segmentation masks available. BackMix augmentation is performed only on the random $f\%$ sample of the training data, leaving the remain-

ing $(1 - f)\%$ of images untouched.³ The network’s focus is expected to correlate with f and the backgrounds pool size that participate in the augmentation. We observe this trend empirically, but find evidence that even with few segmentation masks, significant improvement in performance is seen when compared to networks where BackMix is disabled.

2.3 wBackMix

In scenarios where f is small and BackMix is only applied to few examples, the supervisory signal may be weak and overshadowed by the large quantity of examples with no BackMix. We address this by re-weighting the cross-entropy loss on an example-level to increase the contribution of augmented examples. Specifically, we devise a weighting, which scales non-augmented examples by a factor of $1 - \lambda f$ and augmented examples by $1 + \lambda(1 - f)$. Parameter λ is a constant that is found empirically. When λ is set to a value of zero, it weighs all examples equally, but when increased, it puts more weight on augmented examples. This weight formulation maintains the loss magnitude as the sum of the loss weight in a batch is equal to 1, mitigating any undesirable changes in the training dynamics of the model between experiments.

2.4 Evaluating Focus

To determine whether each model is attending to the pixels inside the sector when making a prediction, we devise two attention-based metrics: energy percentage $\%E$, and focus percentage $\%F$. Firstly, GradCAM class activation maps are calculated for every test image $i \in I$, which give an importance score $z_p \in [0, 1]$ to pixels $p \in i$ based on back-propagated gradients. For $\%E$, the corresponding sector mask $m \in M$ is used to compare the z_p values of pixels inside sector $m \odot i$ with the z_p values across the whole image i . This is formalised as follows, where N is the test set size:

$$\%E = \frac{1}{N} \sum_{i \in I, m \in M} \frac{\sum_{p \in m \odot i} z_p}{\sum_{p \in i} z_p}$$

Higher values of $\%E$ indicate that the network is attending more to pixels within the sector region. We also quantify the main regions of focus using $\%F$, which only considers highly activated pixels where $z_p > 0.5$. We determine the fraction of these pixels that intersects with the sector mask as follows:

$$z_h = \begin{cases} 1 & \text{if } z_p > 0.5 \\ 0 & \text{otherwise} \end{cases} \quad \%F = \frac{1}{N} \sum_{i \in I, m \in M} \frac{\sum_{p \in m \odot z_h} p}{\sum_{p \in z_h} p}$$

³ Note that during training, random backgrounds can only be sampled from frames within the $f\%$ subset and not outside.

Table 1. Comparison of Augmentation methods on TMED and WASE Normals dataset. **Bold** indicates best performance

	Accuracy	TMED (i.d)			WASE Normals (o.o.d)			
		F1	%E	%F	Accuracy	F1	%E	%F
Baseline	97.7	97.5	77.9	92.3	88.7	88.0	79.5	94.3
Black	96.2	95.7	81.8	97.4	89.8	89.4	80.8	96.8
Noise	95.5	95.0	83.7	97.4	89.3	88.8	81.9	96.5
Shuffle	96.6	96.2	82.6	96.8	89.9	89.4	82.1	96.2
Bokeh [13]	97.2	96.9	77.6	93.9	87.9	87.1	78.8	95.4
CutMix [20]	97.9	97.7	70.3	87.6	89.1	88.6	73.2	90.9
SMA [11]	97.2	96.9	82.8	95.3	88.0	86.8	81.9	96.5
BackMix	96.9	96.2	86.2	97.8	92.4	92.1	85.6	97.8

3 Experiments and Results

3.1 Datasets

We train a view classifier on TMED [8, 7] public dataset, a collection of echo studies acquired in the course of routine care from 2011–2020 at Tufts Medical Center, Boston, USA. A labelled subset of 24,964 frames from 1,266 patients was extracted. We extensively validate the classifier generalisability on TMED test set (in-distribution dataset), and WASE Normals [1], a large multi-site proprietary dataset (out-of-distribution dataset). WASE Normals contains 36,029 echo videos from 2,009 healthy volunteers acquired at 18 sites from 15 countries.

Both datasets were filtered to retain the shared view labels, PLAX, PSAX-AV, A2C and A4C, and split into train (80%), validation (10%) and test (10%) sets at a patient level. A single random frame was sampled from each WASE Normals video, resulting in a final dataset of 14,569 train (TMED), 1,670 validation (TMED), 1,815 i.d test (TMED), and 2,565 o.o.d test (WASE Normals) frames. All images have resolution 112×112. Segmentation masks of the ultrasound sector were automatically generated and checked for quality manually⁴.

3.2 Implementation & Training

The baseline ResNet18 and training were implemented in PyTorch and BackMix in Numpy. All models were trained for 100 epochs on a NVIDIA GeForce RTX 2080 Ti with Adam optimiser, batch size of 64, and learning rate of 1e-3. Weights from the epoch with the highest validation accuracy were saved. For reliable performance estimates, all models were trained 3 times with 3 random seeds (shared across experiments), the mean scores were used for quantitative analysis and the model with median performance was used for qualitative analysis. Hyperparameters were tuned on a held-out validation set and set for all experiments. Alongside the attention-based metrics, we also report mean accuracy, precision, recall and F1-score to evaluate classification performance.

⁴ The automatic mask generation was implemented by an in-house proprietary software that uses classical image processing techniques.

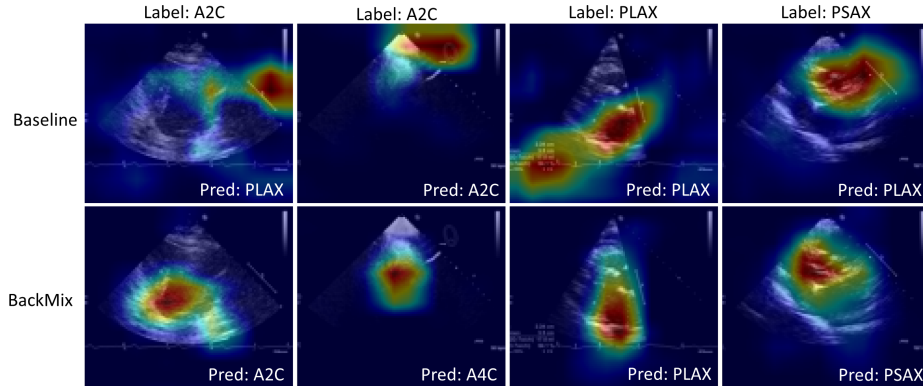


Fig. 4. Qualitative results on TMED with GradCAM heatmaps.

3.3 Comparison to Existing Methods and Ablation Study

We validated BackMix by comparing to various background-based configurations and augmentations in literature. These are: (1) ‘Black’, where the background is filled with zero values; (2) ‘Noise’, where the background is filled with random noise sampled from a uniform distribution; (3) ‘Shuffle’, where the background pixels are randomly arranged; (4) Bokeh [13], a method using background blur; (5) CutMix [20], an augmentation strategy where pairs of images are mixed with a soft label; (6) SMA [11], a contrastive learning method which separates object and background in feature space without segmentation masks. Quantitative and qualitative results are presented in Table 1 and Figure 4, respectively. We focus on a single architecture (ResNet18) because the data used in this work is typically paired with that model [8, 19].

The decreased classification performance on TMED for background augmentation methods is expected as the shortcuts aiding performance are not learnt due to improved sector attention (higher %E and %F). Augmentation methods which significantly alter images (‘Black’, ‘Noise’) fare worst in both i.d and o.o.d

Table 2. Semi-supervised classification for different amounts of supervision on an out-of-distribution dataset (Wase Normals).

	f	λ	Accuracy	Precision	Recall	F1	%E	%F
Baseline	0	-	88.7	91.2	87.1	88.0	79.5	94.3
+ BackMix	1	0	92.4	92.9	91.7	92.1	85.6	97.8
	0.5	0	91.7	92.9	90.6	91.4	84.9	98.0
	0.2	0	91.2	92.3	90.1	90.7	84.4	97.4
	0.1	0	90.4	92.2	89.5	90.2	84.2	97.0
	0.05	0	90.8	92.0	89.5	90.2	83.0	97.4
	0.01	0	90.3	91.7	89.2	89.6	81.4	96.4
+ wBackMix	0.05	1	91.4	92.4	90.5	91.1	83.3	97.2
	0.05	2	91.3	92.5	90.2	90.9	83.7	97.5

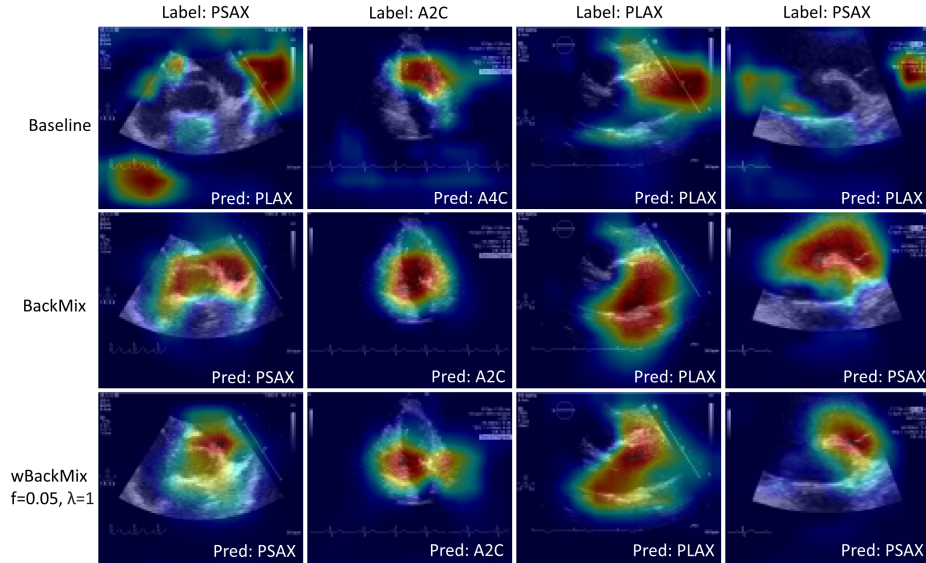


Fig. 5. Qualitative results on WASE with GradCAM heatmaps.

test sets due to the incurred distribution shift. BackMix reduces this distribution shift as all backgrounds contain similar patterns and pixel intensities. In comparison to other methods, BackMix attends the sector best (high %E and %F), enabling learning generalisable representations of the heart. This is reflected in the highest classification performance on the o.o.d dataset.

In Table 2, we validate BackMix and wBackMix in semi-supervised classification under different amounts of supervision on the o.o.d dataset, and present a variety of examples in Figure 5. As the proportion of data to which BackMix is applied decreases, the performance naturally decreases. High performance is maintained when using 5-10% supervision with an accuracy under 91% and an F1 score over 90%. Performance increases when wBackMix is applied at 5% supervision, achieving an accuracy and F1 score comparable to BackMix at 20% supervision. The weighting value used appears to not have a significant impact, and we would recommend a grid search to identify the best configuration.

To further assess our semi-supervised approach, we explore in an ablation study to what extent the random selection of supervised training samples impacts accuracy and focus. We run 5 BackMix experiments ($f=0.05$) with a fixed random seed, but different non-overlapping supervised samples. We find standard deviations of ± 0.55 for accuracy, ± 0.55 for F1, ± 1.01 for %E, and ± 0.70 for %F suggesting that the choice of samples has minimal impact on performance.

4 Conclusion

Echocardiograms contain imaging data in a sector, and non-imaging features that may induce shortcut learning outside the sector. Typically, image analysis methods first need to pre-process and remove background features by applying masks. However this is prone to errors, requires labels and is often computationally expensive. In this paper, we propose BackMix and wBackMix, two augmentation methods which encourage any classification network to focus on imaging data, without the need for a mask during inference. Our results demonstrate that networks trained with BackMix are able to focus more on the sector and ignore spurious correlations in the background, even when augmentation is applied to as few as 5% of training examples. We aim to extend BackMix in the future by performing augmentation in feature space, without needing sector masks.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Asch, F.M., Banchs, J., Price, R., Rigolin, V., Thomas, J.D., Weissman, N.J., Lang, R.M.: Need for a global definition of normative echo values—rationale and design of the world alliance of societies of echocardiography normal values study (wase). *Journal of the American Society of Echocardiography* **32**(1), 157–162 (2019)
2. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* **10**(7), e0130140 (2015)
3. Bassi, P.R., Dertkigil, S.S., Cavalli, A.: Improving deep neural network generalization and robustness to background bias via layer-wise relevance propagation optimization. *Nature Communications* **15**(1), 291 (2024)
4. Guo, M.H., Xu, T.X., Liu, J.J., Liu, Z.N., Jiang, P.T., Mu, T.J., Zhang, S.H., Martin, R.R., Cheng, M.M., Hu, S.M.: Attention mechanisms in computer vision: A survey. *Computational visual media* **8**(3), 331–368 (2022)
5. Hasan, S.K., Simon, R.A., Linte, C.A.: Segmentation and removal of surgical instruments for background scene visualization from endoscopic/laparoscopic video. In: *Medical Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling*. vol. 11598, pp. 55–61. SPIE (2021)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 770–778 (2016)
7. Huang, Z., Long, G., Wessler, B., Hughes, M.C.: A new semi-supervised learning benchmark for classifying view and diagnosing aortic stenosis from echocardiograms. In: *Machine Learning for Healthcare Conference*. pp. 614–647. PMLR (2021)
8. Huang, Z., Long, G., Wessler, B., Hughes, M.C.: Tmed 2: a dataset for semi-supervised classification of echocardiograms. In: *In DataPerf: Benchmarking Data for Data-Centric AI Workshop* (2022)
9. Jung, Y.J., Han, S.H., Choi, H.J.: Explaining cnn and rnn using selective layer-wise relevance propagation. *IEEE Access* **9**, 18670–18681 (2021)

10. Kusunose, K., Haga, A., Inoue, M., Fukuda, D., Yamada, H., Sata, M.: Clinically feasible and accurate view classification of echocardiographic images using deep learning. *Biomolecules* **10**(5), 665 (2020)
11. Kwon, J., Lee, E., Cho, Y., Kim, Y.: Learning to detour: Shortcut mitigating augmentation for weakly supervised semantic segmentation. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. pp. 819–828 (2024)
12. Leclerc, S., Smistad, E., Pedrosa, J., Østvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E.A.R., Jodoin, P.M., Grenier, T., et al.: Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE transactions on medical imaging* **38**(9), 2198–2210 (2019)
13. Li, H., Liu, C., Basu, A.: Semantic segmentation based on depth background blur. *Applied Sciences* **12**(3), 1051 (2022)
14. Ma, C., Zhao, L., Chen, Y., Guo, L., Zhang, T., Hu, X., Shen, D., Jiang, X., Liu, T.: Rectify vit shortcut learning by visual saliency. *IEEE Transactions on Neural Networks and Learning Systems* (2023)
15. Madani, A., Arnaout, R., Mofrad, M., Arnaout, R.: Fast and accurate view classification of echocardiograms using deep learning. *NPJ digital medicine* **1**(1), 6 (2018)
16. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE international conference on computer vision*. pp. 618–626 (2017)
17. Vaseli, H., Liao, Z., Abdi, A.H., Girgis, H., Behnami, D., Luong, C., Dezaki, F.T., Dhungel, N., Rohling, R., Gin, K., et al.: Designing lightweight deep learning models for echocardiography view classification. In: *Medical Imaging 2019: Image-Guided Procedures, Robotic Interventions, and Modeling*. vol. 10951, pp. 93–99. SPIE (2019)
18. Wegner, F.K., Benesch Vidal, M.L., Niehues, P., Willy, K., Radke, R.M., Garthe, P.D., Eckardt, L., Baumgartner, H., Diller, G.P., Orwat, S.: Accuracy of deep learning echocardiographic view classification in patients with congenital or structural heart disease: importance of specific datasets. *Journal of Clinical Medicine* **11**(3), 690 (2022)
19. Wessler, B.S., Huang, Z., Long Jr, G.M., Pacifici, S., Prashar, N., Karmiy, S., Sandler, R.A., Sokol, J.Z., Sokol, D.B., Dehn, M.M., et al.: Automated detection of aortic stenosis using machine learning. *Journal of the American Society of Echocardiography* **36**(4), 411–420 (2023)
20. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 6023–6032 (2019)
21. Zhong, Y., Li, X., Xie, J., Zhang, J.: A lightweight automatic wildlife recognition model design method mitigating shortcut learning. *Animals* **13**(5), 838 (2023)