Faithful, Interpretable Chest X-ray Diagnosis with Artifact-free B-cos Networks

Shreyash Arya[1*][0000−0002−2108−4579], Shashank Agnihotri[2*][0000−0001−6097−8551], Marcel Kleinmann[2][0009−0001−6770−231X], Bernt Schiele[1][0000−0001−9683−5237], and Margret Keuper[2,1][0000−0002−8437−7993]

[1] Max-Planck-Institute for Informatics, Saarland Informatics Campus, Germany
{sarya,schiele}@mpi-inf.mpg.de
[2] Machine Learning Group, University of Mannheim, Mannheim, Germany
{shashank.agnihotri,keuper}@uni-mannheim.de

Abstract. Faithfulness and interpretability are essential for deploying deep neural networks (DNNs) in safety-critical domains such as medical image analysis. B-cos networks modify the parameterization of convolutional and classification layers to measure class evidence via feature-weight alignment, enabling built-in, class-specific contribution maps without post-hoc explanations. While maintaining diagnostic performance competitive with state-of-the-art DNNs, standard B-cos networks exhibit severe aliasing artifacts in their explanation maps, rendering them unsuitable for clinical use, where clarity is essential. In this work, we address this limitation by introducing anti-aliasing strategies using ASAP and BlurPool (BP) to significantly improve explanation quality. Our experiments on chest X-ray datasets demonstrate that the modified B-cos$_{\text{ASAP}}$ and B-cos$_{\text{BP}}$ preserve strong predictive performance while providing faithful and artifact-free explanations suitable for clinical application in multi-class and multi-label settings. Code is available at: `https://github.com/shrebox/Artifact-free-B-cos-Networks`.

Keywords: Explainability · Interpretability · Signal Processing.

1 Introduction

Faithfulness and interpretability are critical prerequisites for deploying deep learning models in safety-critical domains such as healthcare, where decisions have direct implications for patient outcomes [33, 44]. In particular, clinical adoption demands models whose reasoning processes can be understood and verified by medical professionals. However, most existing approaches in medical image processing rely on architectures that lack inherent interpretability, offering limited insight into the basis of their predictions [27]. As emphasized in [44], knowing how a diagnosis was derived is highly valuable, especially in practice,

* Equal contribution.

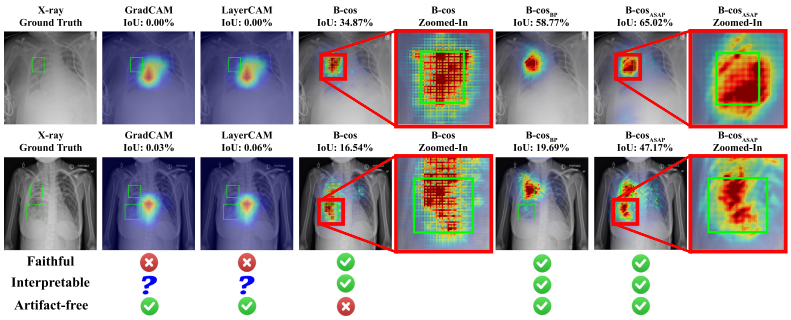

Fig. 1. Comparison of standard B-cos with our anti-aliased variants, B-cos$_{\text{ASAP}}$ and B-cos$_{\text{BP}}$, against post-hoc methods (GradCAM [34], LayerCAM [17]) on the Baseline (ResNet50). While B-cos explanations localize better than post-hoc maps, they exhibit severe grid artifacts that hinder clinical use; B-cos$_{\text{ASAP}}$ and B-cos$_{\text{BP}}$ remove these artifacts and improve localization (IoU with ground-truth boxes). Examples are from RSNA Pneumonia Detection; all methods correctly predict "pneumonia".

where AI tools are used to support, not replace, radiologists. In this work, we explore B-cos networks [7], which are inherently interpretable by design and generate class-specific contribution maps that visually indicate the evidence behind each prediction. Representative examples are shown in Fig. 1, which compares our artifact-free, faithful, and interpretable explanations to other common approaches such as GradCAM and LayerCAM while displaying the need for our extensions when working with B-cos networks and chest X-rays simultaneously.

These models can directly show which part of the image caused network activations, enabling an understanding of why a classification was made [7]. Unlike post-hoc approaches such as GradCAM [34] or LayerCAM [17], which provide coarse and sometimes misleading heatmaps that have insufficient mechanistic explanation of the decision-making process [10, 32], B-cos networks offer inherently interpretable, class-specific contribution maps that yield clearer, more faithful visual explanations. This can help radiologists verify predictions, improve clinical workflows, and reduce diagnostic errors [39, 40]. However, standard B-cos networks face a critical limitation: the generated explanation maps often exhibit grid artifacts, making them visually unreliable for clinical use. In this work, we address these limitations by incorporating anti-aliasing techniques such as ASAP and BlurPool to improve visual clarity. Additionally, while the original B-cos networks support multi-label classification, they were proposed for multi-class classification. Thus, we use the framework from [31], to obtain multi-label classification since chest X-rays commonly involve multiple co-occurring conditions [26]. We refer to the resulting aliasing-free model as B-cos$_{\text{ASAP}}$ and B-cos$_{\text{BP}}$. While B-cos$_{\text{ASAP}}$ combines B-cos networks with ASAP [12], B-cos$_{\text{BP}}$ combines it with BlurPool [43], both effective anti-aliasing techniques. These anti-aliasing techniques help improve the explanation maps from the B-cos network, since they replace the artifact-producing downsampling operations with an

artifact-free downsampling path. Thus, removing the artifacts in the explanation maps of B-Cos networks. Our main contributions are:

- Application-specific modifications to B-cos networks by integrating anti-aliasing methods (ASAP and BlurPool) to produce spectral artifact-free explanations suitable for diagnostic use.
- A practical evaluation of B-cos networks for interpretable and clinically relevant chest X-ray disease detection, showing that interpretability does not impede performance.
- We show that our proposed framework can be adopted for both multi-class and multi-label classification, proving useful in critical medical applications.

2 Related Work

We briefly review (i) clinically relevant evaluation criteria, (ii) deep learning for chest X-ray diagnosis and the datasets used in this work, and (iii) explainability and interpretability requirements for medical AI.

Clinical Evaluation Metrics. While many medical imaging works report standard machine learning metrics [20], clinical studies primarily assess diagnostic tools using sensitivity and specificity [22]. Sensitivity (recall) measures the fraction of true positives detected and is critical for early disease detection, because missed diagnoses can severely impact outcomes [19], with high reported mortality for childhood pneumonia [25]. However, sensitivity often trades off against specificity, and low specificity can lead to unnecessary follow-up and treatment [25]. In this work, we therefore report recall prominently while also tracking specificity-relevant behavior through complementary performance metrics, since reducing false positives is important for clinical trust.

Deep learning for chest X-ray diagnosis. Automated abnormality detection in chest X-rays predates deep learning, with early rule-based and statistical systems reported already in the 1960s [44]. Modern chest X-ray analysis is dominated by deep neural networks, enabled by large-scale datasets and architectural advances, and can approach expert-level performance in several settings [23, 44]. CNN backbones such as ResNet [13] are widely used on large datasets like CheXpert [16] and NIH ChestX-ray14 [42], with strong reported performance in chest X-ray classification [14]. DenseNet-based CheXNet [30, 15] is a prominent example for pneumonia detection. More recently, vision transformers have been applied to pneumonia detection [38], including medical variants such as IEViT [28]. Most importantly for this paper, B-cos aligned transformers have been demonstrated in the medical domain, supporting the relevance of B-cos style inherent explanations beyond natural images [40].

Datasets with localization supervision. Large-scale chest X-ray benchmarks such as NIH ChestX-ray14 [42], CheXpert [16], and MIMIC-CXR [18] have driven progress in classification, but most do not provide region-level supervision for rigorous quantitative evaluation of explanation localization. Since our goal is to assess whether attribution maps align with pathology regions,

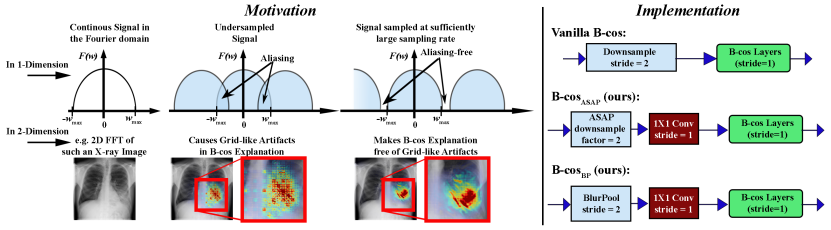

Fig. 2. Method overview: Grid artifacts in B-cos explanations arise from aliasing while downsampling (undersampling). We mitigate them by replacing stride=2 convolutions with anti-aliasing operations like ASAP or BlurPool.

we focus on datasets with bounding boxes. For binary pneumonia detection, we use the RSNA Pneumonia Detection Challenge dataset [36], which provides image-level labels and bounding boxes. For multi-label diagnosis, we use the VinBigData Chest X-ray Abnormalities Detection dataset (VinDr-CXR based) [26], which offers bounding boxes for 14 thoracic findings with radiologist annotation, enabling evaluation under realistic co-occurring conditions.

Explainability and inherent interpretability. Explainability aims to make model decisions understandable [11, 3], and is particularly important in medical imaging where clinicians must be able to verify and trust model outputs [8]. Regulatory discussions further motivate decision traceability in medical AI, including references to European data protection frameworks such as GDPR [37]. A central distinction is between post-hoc explanations and inherently interpretable models [6]. Post-hoc heatmaps can be coarse and may not faithfully reflect the model decision process, which limits their reliability in high-stakes settings [10]. This concern is echoed by arguments that high-stakes applications should prefer interpretable models over post-hoc explanations [32]. In this work, we follow this paradigm and evaluate B-cos networks [7, 4] for chest X-ray diagnosis, focusing on making their built-in contribution maps clinically usable by removing aliasing artifacts while preserving predictive performance.

3 Methodology

3.1 Why do B-cos explanations have Grid-like Artifacts?

It is essential to understand why strided downsampling introduces grid-like artifacts in B-cos explanations, and why ASAP mitigates them. A B-cos neuron admits the scaled linear form,

$$\text{B-cos}(x; w) \;=\; \hat{w}^\top x \cdot \big|c(x, \hat{w})\big|^{B-1}, \qquad \hat{w} = \frac{w}{\|w\|}, \;\; c(x, \hat{w}) = \frac{\hat{w}^\top x}{\|x\|}, \tag{1}$$

which makes explicit that responses depend on both the linear projection $\hat{w}^\top x$ and an input-dependent gain $|c(x, \hat{w})|^{B-1}$. Crucially, B-cos explanations are de-

fined via the induced linear map $W_{1\to l}(x)$ and the input contribution scores

$$s_j^l(x) \;=\; \big[W_{1\to l}(x)\big]_j^\top \;\odot\; x, \tag{2}$$

with pixelwise scores obtained by summing over channels at each spatial location. Therefore, any structured distortion that propagates into $W_{1\to l}(x)$ appears directly and faithfully in the contribution map $s_j^l(x)$, as seen in Fig. 1. A key source of such distortions is improper sampling at resolution changes. Let x be a 2D feature map, h a kernel and $v = x * h$. A stride-d convolution equals a convolution with a subsequent downsampling where every dth element is sampled

$$v_\downarrow[m,n] \;=\; (x*h)[m,n]\cdot\sum_{k,\ell}\delta[m-dk,n-d\ell] = v[dm,dn], \tag{3}$$

where $\sum_{k,\ell}\delta[m-dk,n-d\ell]$ is a 2D dirac impulse comb, i.e. the sampling pattern.

3.2 Anti-aliasing Using Nyquist Rate

In the model forward pass, undersampling when downsampling can lead to aliasing, as shown by Fig. 2 [12, 1, 2]. Specifically, the pointwise multiplication in the spatial domain (c.f. Eq. (3)) corresponds to a convolution in the Fourier domain, i.e., after applying the 2D DFT according to

$$V[\omega_x,\omega_y] = \frac{1}{MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} v[m,n]\, e^{-j\left(\frac{\omega_x}{M}m+\frac{\omega_y}{N}n\right)}, \tag{4}$$

for v of size $M\times N$, the downsampling is given through [35]

$$V_\downarrow[\omega_x,\omega_y] \;=\; V * \frac{1}{d^2}\sum_{k,\ell}\delta\left[\omega_x-\frac{2\pi k}{d},\omega_y-\frac{2\pi\ell}{d}\right] = \frac{1}{d^2}\sum_{k,\ell}V\left[\omega_x-\frac{2\pi k}{d},\omega_y-\frac{2\pi\ell}{d}\right]. \tag{5}$$

The $k,\ell \neq \mathbf{0}$ terms are alias replica that fold energy outside the reconstructable region into the baseband whenever $V[\omega_x,\omega_y]$ is not sufficiently low-pass, i.e. is sampled below the Nyquist rate. Equivalently, alias-free, i.e. distortion-free downsampling requires an anti-aliasing pre-filter such that

$$V[\omega_x,\omega_y]\approx 0 \quad \text{for} \quad \omega_x,\omega_y \notin [-\pi/d,\pi/d)^2. \tag{6}$$

When this condition is violated, periodic grid-like structure is injected into intermediate activations; in B-cos, such structure affects both the projection $\hat{w}^\top x$ and the gain term $|c(x,\hat{w})|^{B-1}$ in Eq. (1), thereby modulating the induced map $W_{1\to l}(x)$ and appearing as spectral/grid artifacts in $s_j^l(x)$ via Eq. (2).

ASAP replaces aliasing-prone downsampling by explicit bandlimited Fourier-domain downsampling. To this end, ASAP computes the 2D DFT to downsample by factor d, ASAP enforces the new Nyquist limit by cropping (zeroing) coefficients outside the admissible band $|\omega_x|,|\omega_y| \le \frac{\pi}{d}$ (thus $\frac{\pi}{2}$ for $d=2$), which suppresses the alias-replica contributions in Eq. (5). To reduce ringing from a

hard spectral cut, ASAP applies a Hamming window in the Fourier domain (2D via outer product) and uses stabilization variants (transpose and/or asymmetric padding for even sizes), before inverse transforming to obtain the $\frac{M}{d} \times \frac{N}{d}$ output. Such a low-pass filter in the frequency domain corresponds to blurring in the spatial domain. Thus, we also implement BlurPool [43], which blurs in the spatial domain while downsampling.

3.3 Making B-cos Aliasing-Free

In our B-cos backbones, we therefore replace each stride=2 downsampling convolution with anti-aliasing operations while downsampling with d=2, followed by a stride=1 convolution:

$$\text{Conv}(\text{stride} = 2) \implies \text{ASAP}_{d=2} \circ \text{Conv}(\text{stride} = 1). \tag{7}$$

We additionally use $\text{BlurPool}_{d=2}$, which executes a spatial blurring operation before downsampling, as an anti-aliasing operation in place of $\text{ASAP}_{d=2}$. Both mitigate aliasing source at resolution transitions, yielding markedly cleaner, grid-like artifacts-free B-cos contribution maps as shown by Fig. 2.

4 Experiments

We run all experiments with three fixed random seeds. We evaluate on two chest X-ray datasets with bounding box annotations: RSNA Pneumonia Detection (binary: pneumonia vs. no pneumonia) [36] and VinBigData (multi-label classification of 14 thoracic findings) [26]. Bounding boxes enable quantitative evaluation of explanation localization in addition to classification performance.

Following established protocols [22], we use 5-fold cross-validation [21] with a 20% held-out test split to capture clinical variability in pneumonia presentation [5]. All baselines and B-cos variants are initialized from ImageNet-pretrained weights and fine-tuned on the target datasets. As commonly used architectures, we study ResNet-50 [13], ConvNeXt-tiny, and ConvNeXt-base [24].

Explanation metrics. We use Energy-based Pointing Game (EPG) scores inspired by ScoreCAM [41]. Let $L^c(i,j)$ denote the attribution at pixel (i,j) for class (label) c, with positive and negative contributions $L^c_p = \max(L^c, 0)$ and $L^c_n = \min(L^c, 0)$. For a ground-truth bounding box set bbox,

$$\text{EPG}(A, B) = \frac{\sum_{(i,j)\in\text{bbox}} A(i,j)}{\sum_{(i,j)\in\text{bbox}} A(i,j) + \sum_{(i,j)\notin\text{bbox}} B(i,j)}. \tag{8}$$

Standard EPG is $\text{EPG}(L^c, L^c)$. We report $\text{EPG}_{\text{Precision}} = \text{EPG}(L^c_p, L^c_p)$, which focuses on how much positive evidence is attributed to within bbox (only B-cos explanations allow analysis with this metric). In contrast, $\text{EPG}_{\text{Recall}} = \text{EPG}(L^c_p, |L^c_n|)$ measures negative evidence inside the box. We average EPG over classes; for VinBigData, we additionally report $\text{EPG}^{\text{micro}}$ (mean is weighted by per-label sample counts), mAP, Recall, F1, and AUROC.

Table 1. Empirical Results on the RSNA Pneumonia Detection Challenge Dataset using ResNet50 and ConvNeXt-base.

Architecture	Model	Accuracy (%)	AUROC	F1 Score	EPG_{Recall}	$EPG_{Precision}$	Recall	mIoU
ResNet50	Baseline	83.5 ± 0.4	83.3 ± 0.6	58.0 ± 1.3	–	–	50.6 ± 1.7	–
	B-cos	80.5 ± 0.8	86.7 ± 0.7	63.3 ± 1.1	100.0 ± 0.0	24.1 ± 0.8	74.8 ± 2.2	4.7 ± 0.5
	B-cos_{BP}	79.2 ± 0.9	87.1 ± 0.7	63.2 ± 0.9	98.7 ± 1.4	20.1 ± 1.9	79.2 ± 1.7	10.7 ± 2.0
	$\text{B-cos}_{\text{ASAP}}$	79.9 ± 0.6	87.1 ± 0.5	63.5 ± 0.6	98.6 ± 0.9	21.1 ± 1.3	77.5 ± 1.1	9.4 ± 1.8
ConvNeXt-base	Baseline	81.0 ± 0.5	81.7 ± 0.6	51.5 ± 1.7	–	–	44.8 ± 2.3	–
	B-cos	79.4 ± 0.6	88.2 ± 0.5	63.6 ± 0.9	97.8 ± 0.6	18.6 ± 1.1	79.9 ± 1.6	1.4 ± 0.5
	B-cos_{BP}	79.1 ± 0.5	88.1 ± 0.4	63.5 ± 0.7	93.9 ± 1.5	20.1 ± 1.6	80.7 ± 1.4	2.1 ± 0.6
	$\text{B-cos}_{\text{ASAP}}$	79.3 ± 1.0	88.0 ± 0.4	63.5 ± 0.8	94.4 ± 1.7	21.9 ± 2.8	80.2 ± 2.3	2.0 ± 0.3

Table 2. Empirical Results on the VinBig Dataset using ResNet50, ConvNeXt-tiny, and ConvNeXt-base as the backbones.

Architecture	Model	mAP	AUROC	F1 Score	EPG^{micro}_{Recall}	$EPG^{micro}_{Precision}$	Recall	mIoU
ResNet50	Baseline	50.6 ± 1.4	80.8 ± 1.1	46.2 ± 1.9	–	–	38.6 ± 1.8	–
	B-cos	53.5 ± 1.4	91.9 ± 0.6	44.5 ± 1.5	97.7 ± 0.4	9.4 ± 0.2	38.6 ± 1.5	2.3 ± 0.2
	B-cos_{BP}	47.8 ± 1.3	87.7 ± 0.9	42.3 ± 1.6	92.3 ± 1.2	9.0 ± 0.2	37.3 ± 1.6	5.4 ± 0.7
	$\text{B-cos}_{\text{ASAP}}$	47.6 ± 1.3	87.3 ± 0.8	42.2 ± 1.2	92.4 ± 1.3	9.9 ± 0.4	37.0 ± 1.1	5.2 ± 0.9
ConvNeXt-tiny	Baseline	52.0 ± 1.0	90.2 ± 0.7	45.1 ± 1.1	–	–	39.4 ± 1.0	–
	B-cos	57.9 ± 1.5	90.1 ± 1.1	53.4 ± 1.2	94.9 ± 0.6	12.3 ± 0.4	47.6 ± 1.3	2.3 ± 0.3
	B-cos_{BP}	55.9 ± 1.6	90.0 ± 0.7	49.4 ± 1.3	93.7 ± 0.5	14.6 ± 0.4	42.1 ± 1.3	4.6 ± 0.4
	$\text{B-cos}_{\text{ASAP}}$	50.6 ± 1.4	90.9 ± 0.6	40.0 ± 1.8	93.6 ± 1.1	11.4 ± 1.0	33.4 ± 1.7	2.3 ± 0.5
ConvNeXt-base	Baseline	51.6 ± 0.9	89.7 ± 0.6	44.7 ± 0.8	–	–	39.0 ± 0.9	–
	B-cos	58.4 ± 1.5	92.2 ± 0.5	51.3 ± 1.4	95.4 ± 0.6	13.2 ± 0.4	44.8 ± 1.4	1.6 ± 0.1
	B-cos_{BP}	54.9 ± 1.8	90.3 ± 1.1	47.5 ± 1.4	90.7 ± 1.0	16.2 ± 0.7	40.7 ± 1.6	4.6 ± 0.4
	$\text{B-cos}_{\text{ASAP}}$	49.8 ± 1.3	91.7 ± 0.5	37.7 ± 1.9	93.1 ± 1.0	13.4 ± 1.1	31.5 ± 1.7	2.8 ± 0.4

4.1 Quantitative Results

We observe in Fig. 1 that modifying the strided convolution operations in B-cos with ASAP or BlurPool provides significantly better and aliasing-free explanations than the inherent B-cos explanations. We also observe that these explanations are significantly better localized than the post-hoc explanations like GradCAM and LayerCAM on the Baseline architectures. It is paramount that interpretability does not come at the cost of performance. Thus, in Tab. 1 and Tab. 2, we report the empirical performance over 5-fold and 3 random seeds on the RSNA Pneumonia Detection Challenge Dataset [36] and the VinBigXray Dataset [26] using prominent architectures like ResNet50, ConvNeXt-tiny, and ConvNeXt-base, trained on the respective datasets. We observe that the performance of B-cos, $\text{B-cos}_{\text{ASAP}}$, and B-cos_{BP} is comparable to the baseline architectures. This is reflected in the significantly higher than baseline Recall for the B-cos variants on the RSNA Pneumonia Detection Challenge Dataset. This is desirable as a higher recall is often warranted for medical diagnosis.

5 Discussion

This work explores the potential of B-cos networks for chest X-ray classification. We show that they provide faithful, high-quality explanations grounded in the model architecture [7], while achieving competitive classification performance.

Applying B-cos networks to the RSNA Pneumonia Detection Challenge [36] and VinBigData [26] datasets revealed one critical limitation: explanation maps

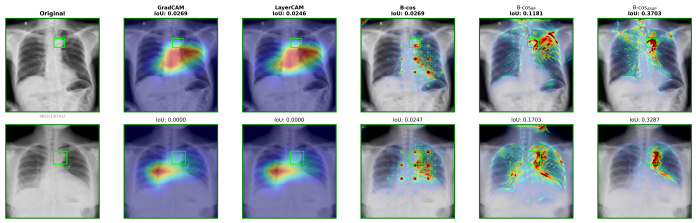

Fig. 3. Visualizing explanations from Baseline and B-cos networks using post-hoc attribution methods and inherent explanations respectively, for ResNet50 architecture on the VinBig Chest X-Ray dataset, for ground truth label: Aortic Enlargement.

are impaired by aliasing artifacts from strided convolutions. Replacing these operations with anti-aliasing layers, namely BlurPool [43] and ASAP [12], removes these artifacts and substantially improves interpretability.

With these modifications, B-cos networks provide faithful, high-resolution explanations aligned with disease-relevant regions, while preserving performance competitive with baseline models. Our precision-based energy pointing game (EPG) metrics show that over 50% of high-confidence activations fall inside bounding boxes, although these cover only 11.65% of the image area. This confirms that the models focus on clinically meaningful regions. Importantly, our goal is not to improve classifier performance, but to obtain faithful, aliasing-free B-cos explanations without sacrificing it. This is reflected in the results: on RSNA with ConvNeXt-B, B-cos$_{\text{ASAP/BP}}$ are only about 2% lower in accuracy than the baseline, but improve recall from 45% to ≈80% and AUROC from 82% to 88%, while adding inherent explanations. On VinBigXray with ConvNeXt-B, B-cos$_{\text{BP}}$ is on par with or better than the baseline in accuracy, recall, precision, and AUROC, while B-cos$_{\text{ASAP}}$ shows a recall trade-off but maintains high AUROC and improves mIoU. A stronger encoder further improves metrics, as shown by ConvNeXt-B compared to ResNet50.

All metrics are reported as mean ± standard deviation over 3 seeds and 5 folds. Paired t-tests over matched seed/fold runs show that B-cos$_{\text{ASAP}}$ and B-cos$_{\text{BP}}$ significantly improve mIoU over B-cos in all dataset-backbone comparisons, with maximum $p = 7.0\text{e}{-3}$. On RSNA, neither variant is significantly worse than B-cos in Accuracy, Recall, Precision, or AUROC, indicating improved localization without loss of predictive performance. While B-cos$_{\text{ASAP}}$ and B-cos$_{\text{BP}}$ perform similarly overall, with different trade-offs, ASAP gives slightly better visual explanations, as shown in Fig. 3.

Some limitations remain. We did not optimize for high recall alone, since this can increase false positives and reduce clinician trust [8]; instead, we aimed for balanced performance across metrics. Moreover, although B-cos was evaluated on multiple architectures, interpretability for vision transformers [9] was not analyzed in depth, and the effect of anti-aliasing in these models remains open. While this work focuses on medical applications, the methodology is broadly applicable, for example, to industrial anomaly detection [29].

6 Conclusion

This work is a comprehensive analysis of B-cos networks for chest X-ray classification, targeting the key limitation that has restricted their application in medical imaging: aliasing artifacts in explanation maps. We address them by integrating anti-aliasing techniques, ASAP, and BlurPool, which improve the clarity and reliability of the explanations. We demonstrate that the anti-aliasing methods work with both multi-class and multi-label B-cos networks to produce contribution maps, making the method suitable for realistic diagnostic tasks. Our experiments on the RSNA Pneumonia Detection Challenge and VinBigData Chest X-ray Abnormalities Detection dataset show that B-cos networks provide competitive classification performance while offering inherently interpretable, high-fidelity explanations. These explanations align closely with disease-relevant regions, as demonstrated both qualitatively and through strong energy-based pointing game scores, outperforming post-hoc attribution methods like LayerCAM and GradCAM. By enhancing the interpretability and practical applicability of B-cos networks, this work takes an important step toward transparent and trustworthy AI-assisted diagnosis in clinical settings [44], and sets the stage for future work on scalable, interpretable deep learning for medical imaging.

Acknowledgement. S.A. and M.K. acknowledge support by DFG Research Unit 5336 – Learning to Sense.

Disclosure of Interests. There are no conflicting interests.

References

[1] S. Agnihotri, J. Grabinski, J. Keuper, and M. Keuper. Beware of aliases–signal preservation is crucial for robust image restoration. *arXiv preprint arXiv:2406.07435*, 2024.

[2] S. Agnihotri, J. Grabinski, and M. Keuper. Improving feature stability during upsampling–spectral artifacts and the importance of spatial context. In *European Conference on Computer Vision*, pages 357–376. Springer, 2024.

[3] E. Alshami, S. Agnihotri, B. Schiele, and M. Keuper. Aim: Amending inherent interpretability via self-supervised masking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 993–1003, 2025.

[4] S. Arya, S. Rao, M. Böhle, and B. Schiele. B-cosification: Transforming deep neural networks to be inherently interpretable. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 62756–62786. Curran Associates, Inc., 2024.

[5] G. Balachandran. *Interpretation of Chest X-ray: An Illustrated Companion*. Jaypee Brothers Medical Publishers, New Delhi, 1 edition, 2014. ISBN 9789351521723.

[6] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, June 2020. ISSN 1566-2535. https://doi.org/10.1016/j.inffus.2019.12.012.

[7] M. Böhle, M. Fritz, and B. Schiele. B-cos networks: Alignment is all we need for interpretability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10329–10338, 2022.

[8] H. Chen, C. Gomez, C.-M. Huang, and M. Unberath. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. *npj Digital Medicine*, 5(1):1–15, Oct. 2022. ISSN 2398-6352. https://doi.org/10.1038/s41746-022-00699-2. Publisher: Nature Publishing Group.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. arXiv:2010.11929 [cs].

[10] S. Ghosh, K. Yu, F. Arabshahi, and K. Batmanghelich. Bridging the gap: From post hoc explanations to inherently interpretable models for medical imaging. In *Proceedings of the 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH) at ICML 2023*, 2023.

[11] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. Explaining Explanations: An Overview of Interpretability of Machine Learning, Feb. 2019. arXiv:1806.00069 [cs].
[12] J. Grabinski, J. Keuper, and M. Keuper. Fix your downsampling ASAP! Be natively more robust via Aliasing and Spectral Artifact free Pooling, July 2023. arXiv:2307.09804 [cs].
[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition, Dec. 2015. arXiv:1512.03385 [cs].
[14] M. B. Hossain, S. M. H. S. Iqbal, M. M. Islam, M. N. Akhtar, and I. H. Sarker. Transfer learning with fine-tuned deep CNN ResNet50 model for classifying COVID-19 from chest X-ray images. *Informatics in Medicine Unlocked*, 30:100916, Jan. 2022. ISSN 2352-9148. https://doi.org/10.1016/j.imu.2022.100916.
[15] G. Huang, Z. Liu, L. v. d. Maaten, and K. Q. Weinberger. Densely Connected Convolutional Networks, Jan. 2018. arXiv:1608.06993 [cs].
[16] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison, Jan. 2019.
[17] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, and Y. Wei. LayerCAM: Exploring Hierarchical Class Activation Maps for Localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021. ISSN 1057-7149, 1941-0042. https://doi.org/10.1109/TIP.2021.3089943.
[18] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, Dec. 2019. ISSN 2052-4463. https://doi.org/10.1038/s41597-019-0322-0. Publisher: Nature Publishing Group.
[19] J. Juan, E. Monsó, C. Lozano, M. Cufí, P. Subías-Beltrán, L. Ruiz-Dern, X. Rafael-Palou, et al. Computer-assisted diagnosis for an early identification of lung cancer in chest X rays. *Scientific Reports*, 13(1):7720, May 2023. ISSN 2045-2322. https://doi.org/10.1038/s41598-023-34835-z. Publisher: Nature Publishing Group.
[20] M. Keuper, T. Schmidt, M. Temerinac-Ott, J. Padeken, P. Heun, O. Ronneberger, and T. Brox. Blind deconvolution of widefield fluorescence microscopic data by regularization of the optical transfer function (otf). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2179–2186, 2013.
[21] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*, IJCAI'95, pages 1137–1143, San Francisco, CA, USA, Aug. 1995. Morgan Kaufmann Publishers Inc. ISBN 978-1-55860-363-9.
[22] R. Kundu, R. Das, Z. W. Geem, G.-T. Han, and R. Sarkar. Pneumonia detection in chest X-ray images using an ensemble of deep learning models. *PLOS ONE*, 16(9):e0256630, Sept. 2021. ISSN 1932-6203. https://doi.org/10.1371/journal.pone.0256630. Publisher: Public Library of Science.
[23] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, et al. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, Dec. 2017. ISSN 1361-8415. https://doi.org/10.1016/j.media.2017.07.005.
[24] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A ConvNet for the 2020s, Mar. 2022. arXiv:2201.03545 [cs].
[25] E. Naydenova, A. Tsanas, S. Howie, C. Casals-Pascual, and M. De Vos. The power of data mining in diagnosis of childhood pneumonia. *Journal of The Royal Society Interface*, 13(120): 20160266, July 2016. https://doi.org/10.1098/rsif.2016.0266. Publisher: Royal Society.
[26] H. Q. Nguyen, K. Lam, L. T. Le, H. H. Pham, D. Q. Tran, D. B. Nguyen, D. D. Le, et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations, 2020.
[27] L. Oakden-Rayner. Exploring large scale public medical image datasets, July 2019. arXiv:1907.12720 [eess].
[28] G. I. Okolo, S. Katsigiannis, and N. Ramzan. IEViT: An enhanced vision transformer architecture for chest X-ray image classification. *Computer Methods and Programs in Biomedicine*, 226:107141, Nov. 2022. ISSN 0169-2607. https://doi.org/10.1016/j.cmpb.2022.107141.
[29] N. Poggi, S. Agnihotri, and M. Keuper. Smart eyes for silent threats: Vlms and in-context learning for thz imaging. *arXiv preprint arXiv:2507.15576*, 2025.
[30] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning, Dec. 2017. arXiv:1711.05225 [cs].
[31] S. Rao, M. Böhle, A. Parchami-Araghi, and B. Schiele. Studying how to efficiently and effectively guide models with explanations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1922–1933, 2023.
[32] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.
[33] S. N. Saw and K. H. Ng. Current challenges of implementing artificial intelligence in medical imaging. *Physica Medica: European Journal of Medical Physics*, 100:12–17, Aug. 2022. ISSN 1120-1797, 1724-191X. https://doi.org/10.1016/j.ejmp.2022.06.003. Publisher: Elsevier.
[34] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *Interna-*

tional Journal of Computer Vision, 128(2):336–359, Feb. 2020. ISSN 0920-5691, 1573-1405. https://doi.org/10.1007/s11263-019-01228-7. arXiv:1610.02391 [cs].

[35] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

[36] G. Shih, C. C. Wu, S. S. Halabi, M. D. Kohli, L. M. Prevedello, T. S. Cook, A. Sharma, et al. Augmenting the National Institutes of Health Chest Radiograph Dataset with Expert Annotations of Possible Pneumonia. *Radiology: Artificial Intelligence*, 1(1):e180041, Jan. 2019. https://doi.org/10.1148/ryai.2019180041. Publisher: Radiological Society of North America.

[37] A. Singh, S. Sengupta, and V. Lakshminarayanan. Explainable Deep Learning Models in Medical Image Analysis. *Journal of Imaging*, 6(6):52, June 2020. ISSN 2313-433X. https://doi.org/10.3390/jimaging6060052.

[38] S. Singh, M. Kumar, A. Kumar, B. K. Verma, K. Abhishek, and S. Selvarajan. Efficient pneumonia detection using Vision Transformers on chest X-rays. *Scientific Reports*, 14(1):2487, Jan. 2024. ISSN 2045-2322. https://doi.org/10.1038/s41598-024-52703-2. Publisher: Nature Publishing Group.

[39] S. Sun, S. Woerner, A. Maier, L. M. Koch, and C. F. Baumgartner. Attri-net: A globally and locally inherently interpretable model for multi-label classification using class-specific counterfactuals. *arXiv preprint arXiv:2406.05477*, 2024.

[40] M. Tran, A. Lahiani, Y. Dicente Cid, M. Boxberg, P. Lienemann, C. Matek, S. J. Wagner, F. J. Theis, E. Klaiman, and T. Peng. B-cos aligned transformers learn human-interpretable features. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 514–524. Springer, 2023.

[41] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu. Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 111–119, Seattle, WA, USA, June 2020. IEEE. ISBN 978-1-72819-360-1. https://doi.org/10.1109/CVPRW50498.2020.00020.

[42] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471, July 2017. https://doi.org/10.1109/CVPR.2017.369. arXiv:1705.02315 [cs].

[43] R. Zhang. Making convolutional networks shift-invariant again. In *International conference on machine learning*, pages 7324–7334. PMLR, 2019.

[44] E. Çallı, E. Sogancioglu, B. van Ginneken, K. G. van Leeuwen, and K. Murphy. Deep learning for chest X-ray analysis: A survey. *Medical Image Analysis*, 72:102125, Aug. 2021. ISSN 1361-8415. https://doi.org/10.1016/j.media.2021.102125.