

# INTERPRETING MEDICAL IMAGE CLASSIFIERS BY OPTIMIZATION BASED COUNTERFACTUAL IMPACT ANALYSIS

David Major\*, Dimitrios Lenis\*, Maria Wimmer, Gert Sluiter, Astrid Berg, and Katja Bühler

VRVis Zentrum für Virtual Reality und Visualisierung Forschungs-GmbH, Vienna, Austria

## ABSTRACT

Clinical applicability of automated decision support systems depends on a robust, well-understood classification interpretation. Artificial neural networks while achieving class-leading scores fall short in this regard. Therefore, numerous approaches have been proposed that map a salient region of an image to a diagnostic classification. Utilizing heuristic methodology, like blurring and noise, they tend to produce diffuse, sometimes misleading results, hindering their general adoption.

In this work we overcome these issues by presenting a model agnostic saliency mapping framework tailored to medical imaging. We replace heuristic techniques with a strong neighborhood conditioned inpainting approach, which avoids anatomically implausible artefacts. We formulate saliency attribution as a map-quality optimization task, enforcing constrained and focused attributions. Experiments on public mammography data show quantitatively and qualitatively more precise localization and clearer conveying results than existing state-of-the-art methods.

**Index Terms**— Classifier Decision Visualization, Image Inpainting, Mammography, Explainable AI

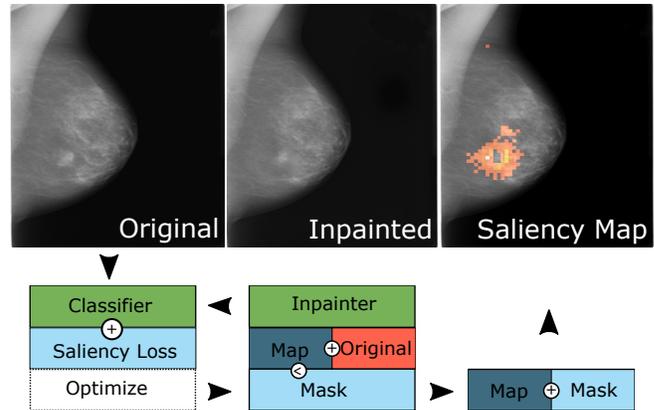
## 1. INTRODUCTION

Consulting radiologists will routinely back their findings by pinpointing and describing a specific region on a radiograph. Contrary, acting as a highly efficient black box, artificial neural networks (ANN) [1] fall short of this form of explanation for their predictions. ANNs' high dimensional, nonlinear nature, does not induce a canonical map between derived prediction and input image. Understandably, a plethora of approaches have been presented that try to derive a so called saliency map, that is, a robust mapping between pixel space and prediction class [2–10].

Most frequently this form of reasoning is based on *local explanations* (LE), i.e. on concrete maps for image-prediction

\*Equal contribution

VRVis is funded by BMVIT, BMDW, Styria, SFG and Vienna Business Agency in the scope of COMET - Competence Centers for Excellent Technologies (854174) which is managed by FFG. Thanks go to our project partner AGFA HealthCare for providing valuable input.



**Fig. 1:** Our saliency mapping framework: (i) classification score of an image is obtained, (ii) a hole mask is generated, inpainted and classified, (iii) saliency loss is computed based on the score difference of original/inpainted images and map quality, (iv) optimization is continued for a fixed number of steps and a result map/mask is derived.

pairs [3, 11]. A clinically applicable LE needs to be *informative* for the radiologists, that is, focusing on regions coinciding with medical knowledge [12]. Moreover, a methodologically sound LE is *faithful* to the classifier, i.e. dependent on architecture, parametrization, and its preconditions like training-set distribution [13].

*Direct approaches* efficiently utilize the assumed analytic nature or the layered architecture of an ANN classifier to derive the desired saliency map for a LE [5,6]. While frequently applied, the obtained results of this class are possibly incomplete, diffuse, hard to interpret, and as recent work shows misleading [2–4, 7, 13]. Thereby they violate both criteria, informativeness and faithfulness, hindering their general application in medical imaging.

Contrary, *reference based* LE approaches [3] try to mitigate these issues by studying how the given classifier reacts to perturbations of the input image. Using the original input as a reference and marginalizing a dedicated image region's contribution, they estimate this region's effect on the classification score. Solutions mainly vary in the ways this marginalization is achieved. They range from heuristic approaches, e.g. blurring, noise, or graying out [3, 4], over lo-

cal neighbourhood conditioning [2], to utilizing strong conditional generative models [8, 10]. These methods address *informativeness*, however, applied to medical images, they introduce noise, possibly pathological indications, anatomical implausible tissue or other adversarial artefacts. By this, they amplify the out-of-distribution problem, similar to an adversarial attack: they expect a meaningful classification result for an image, that is not within the training-set distribution. Hence, they fall short of *faithfulness* for clinical applications.

Marginalization for medical imaging, i.e. the replacement of pathological regions with counterfactual healthy tissue, is being actively explored and addressed by generative adversarial network setups (GAN). Besides promising results, authors report resolution limitations, and the same underlying out-of-distribution issue [14–17].

**Contribution:** We address the open challenge of *faithful and informative* medical black-box classifier interpretation by expanding natural image classifier visualization approaches [2–4]. We propose a reference based optimization framework tailored to medical images, focusing on the interactions between original and marginalized image classification-scores, and map quality. To tackle anatomical correctness of marginalization in medical images, partial convolution inpainting [18] is adapted. Hence, instead of a globally acting GAN, we utilize local per-pixel reconstruction without sacrificing global image composition. We validate our approach on publicly available mammography data, and show quantitatively and qualitatively more precise localization, and clearer conveying results than existing state-of-the-art methods.

## 2. METHODS

Our goal is to estimate a *faithful and informative* saliency map between a medical image and its classification score: given an image, we search for and visually attribute the *specific* pixel-set that contributes towards a confident classification for a fixed class (see Fig. 1). Following [2, 4] we formulate the general problem as finding the *smallest deletion region* (SDR) of a class  $c$ , i.e. the pixel-set whose marginalization w.r.t. the classifier lowers the classification score for  $c$ .

**Image-wise Saliency Mapping:** Informally, we search for the smallest smooth map, that indicates the regions we need to change (inpaint) such that we get a *sufficiently healthy* image able to *fool the classifier*. We formalize the problem as follows: Let  $I$  denote an image of a domain  $\mathcal{I}$  with pixels on a discrete grid  $m_1 \times m_2$ ,  $c$  a fixed class, and  $f$  a classifier capable of estimating  $p(c|I)$ , the probability of  $c$  for  $I$ . Also let  $M$  denote the saliency mask for image  $I$  and class  $c$ , hence  $M \in M^{m_1 \times m_2}(\{0, 1\})$ . We use *total variation*  $tv(M)$  [4], and *size*  $ar(M)$ , to measure the mask’s shape. Note that *size* here is ambiguous. Experimentally we found dice overlap with regions-of-interest like organ masks to be favourable over the map’s average

pixel value [4]. With  $\odot$  denoting elementwise multiplication, and  $\pi(M)$  the inpainting result of a hole image  $I \odot M$ , we can define  $\phi(M) := -1 \cdot \log(p(c|\pi(M)))$  and  $\psi(M) := \log(\text{odds}(I)) - \log(\text{odds}(\pi(M)))$ , where  $\text{odds}(I) = \frac{p(c|I)}{1-p(c|I)}$ . Both,  $\phi$  and  $\psi$ , weigh the new probability of the inpainted image. If we assume class  $c$  to denote *pathological*, then healthy images, and large score differences will be favoured. With this preparation we define our desired optimization function as

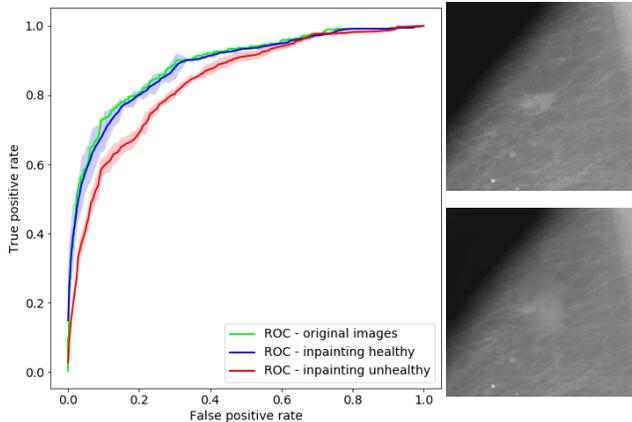
$$\mathcal{L}(M) := \lambda_1 \cdot (\phi(M) + \psi(M)) + \lambda_2 \cdot tv(M) + \lambda_3 \cdot ar(M)$$

where  $\lambda_i \in \mathbb{R}$  are regularization parameters, and search for  $\arg \min_M \mathcal{L}(M)$ . There are two collaborating parts in  $\mathcal{L}$ . The first term enforces the class probability to drop, the latter two emphasize an informative mask. Focusing on medical images,  $\mathcal{L}$  directly solves the SDR task, thereby minimizing medically implausible and adversarial artefacts caused by inpainting of large classifier-neutral image regions, as observable in [3, 4, 18]. The optimization problem is solved by local search through stochastic gradient descent, starting from a regular grid initialization. By design, no restrictions are applied on the classifier  $f$ . For optimization we relax the mask’s domain to  $M^{m_1 \times m_2}([0, 1])$ , and threshold at  $\theta \in (0, 1)$ .

**Image Inpainting with Partial Convolutions:** For marginalization, we want to emphasize local context, while still considering global joint region interaction, and thereby favor a globally sound anatomy. Therefore, we adapt the U-Net like architecture of [18], capable of handling masks with irregular shapes, fitting our optimization requirements for pathological regions of different sizes and shapes. The chosen architecture consists of eight partial convolution layers on both encoding and decoding parts. It takes an image with holes  $I \odot M$  and the hole mask  $M$  as an input, and outputs the inpainted image  $\pi(M)$ . The partial convolution layers insert only the convolution result of the current sliding convolution-window when image information is present. The convolution filter  $W$  is applied on the features  $X$  using the binary mask  $M$  and yields new features  $x'$  the following way:

$$x' = \begin{cases} W^T(X \odot M) \frac{1}{\text{sum}(M)} + b, & \text{if } \text{sum}(M) > 0 \\ 0, & \text{otherwise} \end{cases}$$

where  $b$  is the bias term. The convolution operation is scaled by  $\frac{1}{\text{sum}(M)}$  according to the amount of information available in the current sliding window. Moreover a new mask is passed to the next layer which is updated by setting its values to 1 in the sliding window if  $\text{sum}(M) > 0$ . The authors of [18] propose to train the network with a loss function concentrating on both per-pixel reconstruction performance of the hole/non-hole regions and on overall appearance of the image. To improve the overall appearance a perceptual loss and a style loss are applied which match images in a mapped feature space. Total variation is used as a last loss component to ensure a smooth transition between hole regions and present image regions.



**Fig. 2:** Left: Comparison of classifier performance without inpainting (green), with inpainting in healthy tissue (blue) and in mass tissue (red) over 10 random runs (shaded). Right: Original image with mass (top), inpainted with replaced healthy texture (bottom).

P	$D_{ours}$	$D_{cam}$	$D_{sal}$
50	<b>137.1±69.5</b>	200.1±65.3	164.2±36.0
75	<b>137.1±69.5</b>	182.1±78.9	162.9±36.9
90	<b>137.1±69.5</b>	144.1±89.3	166.3±44.0
Ablation	163.3±35.9	-	-

P	$H_{ours}$	$H_{cam}$	$H_{sal}$
50	<b>227.1±82.1</b>	359.3±83.9	371.3±61.5
75	<b>227.1±82.1</b>	308.5±103.4	350.3±63.1
90	<b>227.1±82.1</b>	232.2±121	328.9±64.5
Ablation	328.1±69.5	-	-

**Table 1:** Results of the saliency mapping on test data with masses for different percentiles  $P$ :  $D...$  average Euclidean,  $H...$  average Hausdorff distances, between result mask connected components and ground-truth annotations (in pixels).

### 3. EXPERIMENTAL SETUP

**Dataset:** In this work the Database for Screening Mammography (DDSM) [19] and the Curated Breast Imaging Subset of DDSM (CBIS-DDSM) [20] were used, downsampled to a resolution of 576x448 pixels. Data was split into 1231 scans containing masses and 2000 healthy samples for training, and into 334 mass and 778 healthy scans for testing. Scans with masses contain pixel-wise ground-truth annotation (GT).

**Image Classifier:** The basis of our saliency detection framework is a MobileNet [21] binary classifier to categorize images either as healthy or as a sample with masses. The network was trained on all training scans with batch size of 4 using the Adam optimizer with a learning rate ( $lr$ ) of  $1e-5$  for 250 epochs using early stopping. Rotation, zoom, horizontal and vertical flips were used for data augmentation. It was pre-trained by 50k 224x224 pixel patches from the training data with the task of classifying background vs. masses.

P	$A_{ours}$	$A_{cam}$	$A_{sal}$	$O_{cam}$	$O_{sal}$
50	<b>.02±.02</b>	.43±.10	.50±.0	.60±.33	.86±.14
75	<b>.02±.02</b>	.24±.03	.25±.0	.44±.32	.64±.22
90	<b>.02±.02</b>	.10±.0	.10±.0	.23±.25	.39±.23

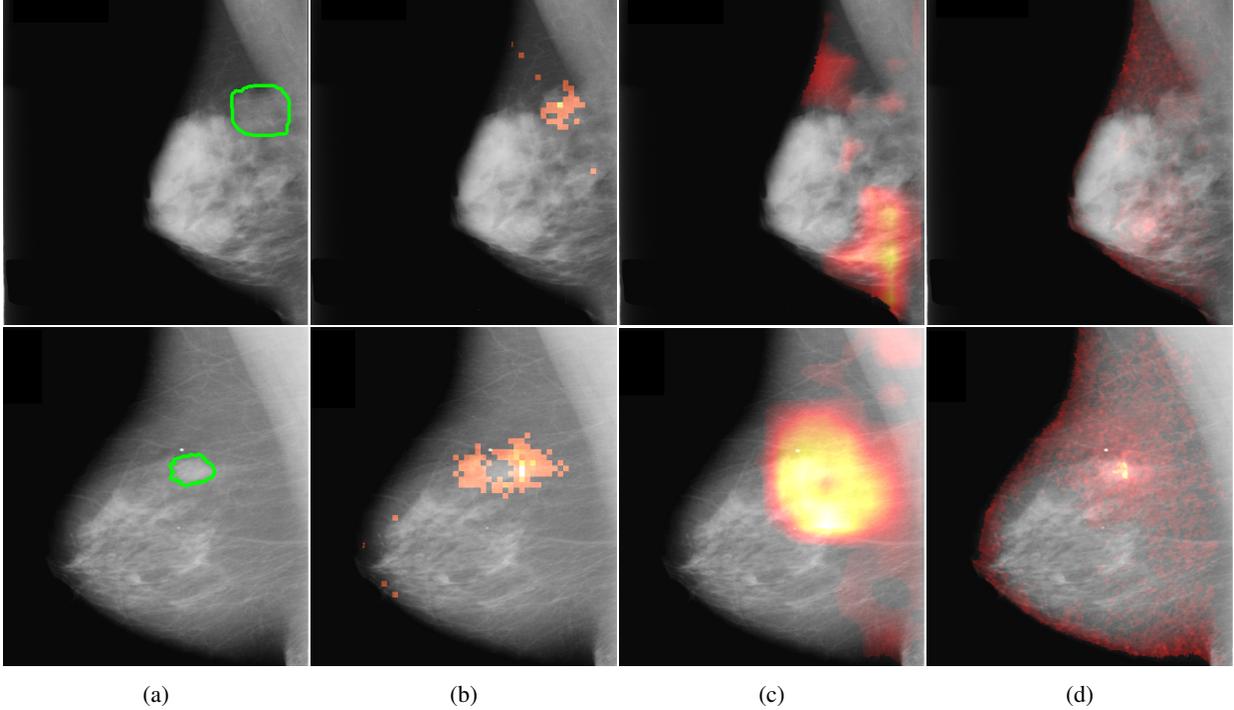
**Table 2:** For three threshold-levels, chosen by percentiles of the output-distribution of CAM resp. SAL:  $A$ : Average overlaps between attributions and breast area;  $O$ : Overlap between our attribution and CAM resp. SAL.

**Inpainting:** The inpainter was trained on the healthy training samples with a batch size of 1 in two phases [18]. The first phase was set up with batch normalization (BN) and  $lr = 1e-5$  for 100 epochs, the second without BN in the encoder part and with  $lr = 1e-6$  for 50 epochs. For each image up to 400 8x8 pixel holes were generated at random positions, where both single small holes and larger clusters were simulated to mimic configurations during optimization.

The inpainter has the task to change the classification score of an image towards healthy when replacing mass tissue, no considerable change should happen otherwise. To demonstrate that, we computed (i) a ROC curve using the classifier on all test samples without any inpainting, (ii) ROC curves for inpainting only in healthy tissue over 10 runs with randomly sampled holes and (iii) ROC curves for inpainting of mass tissue in unhealthy scans over 10 runs (Fig. 2 left).

**Saliency Mapping:** Parametrization was experimentally chosen based on grid-search, restricted by  $\lambda_i \in [0, 1]$ , for  $i = 1, 2, 3$ . Similar to [3, 8] we found the resulting masks to be especially sensitive to  $\lambda_2$ . This smoothness controlling term, balances between noisy result-maps and compression induced information-loss. We exemplify this behaviour with an ablation study, where contributions of smoothing and sizing are set to zero (cf. Table 1). The final optimization results were derived in 100 steps per image, with  $lr = 2e-3$ ,  $\theta = 0.5$  and setting  $\lambda_1 = 1.0$ ,  $\lambda_{2,3} = 0.1$ .

We compared our approach against two established methods based on widespread adaptation in medical imaging [15, 22], and inherent validity [13]. We chose the gradient based *Saliency Map* [5] (SAL) and the network-derived *Cam* [6] (CAM) visualizations. As our domain prohibits the utilization of blurring, noise, etc. we could not meaningfully test against reference based methods [2–4]. For evaluation, we derived four measures, that (i) relate the result masks (RM) to the mass annotations (GT), and (ii) compare the result masks to each other (cf. Table 1, 2). Particularly for (i) we studied  $D$  the average of Euclidean distances between the centres of GT masks and RMs’ connected components, and  $H$  the average Hausdorff distances between GT masks and RM. Here, lower values indicate better localization, i.e. a better vicinity to the pathology. Conversely, for (ii) we calculated  $A$  the ratio between derived RMs and the organ masks (the breast area), and  $O$  the overlap coefficients between our RMs and those of CAM and SAL. Here lower values of  $A$  indicate a more com-



**Fig. 3:** Result saliency heatmaps: (a) depicts the original image overlaid with mass annotation contours, (b) shows results of our saliency mapping framework. (c) is the CAM of the classifier and (d) visualizes SAL, both thresholded by 50th-percentile.

compact map, i.e. better visual distinction.  $O$  scores describe the correlation level between result masks. Statistically significant difference between all resulting findings was formalized using Wilcoxon signed-rank tests, for  $\alpha < 0.01$ .

All measurements were performed on binary masks, for which CAM and SAL had to be thresholded as those maps’ noise covers the complete image. We therefore chose the 50/75/90th percentiles, i.e. 50/25/10 percent of the map-points. Where multiple masses, or mapping results occurred we used their median for a robust estimation per image.

#### 4. RESULTS AND CONCLUSION

**Inpainting Evaluation:** The ROC curves in Fig. 2 represent an AUC of 0.89 for original images (green), average AUCs of 0.88 for inpainting tissue in healthy cases (blue) and 0.83 for inpainting only in masses for pathological cases (red). Besides the AUCs, the visual separability of the green/blue curves from the red one indicates that the inpainter behaves correctly and introduces significant changes only when replacing mass tissue w.r.t. the classifier. The inpainting quality of replacing mass with healthy tissue is visible in Fig. 2 right.

**Saliency Evaluation:** *Quantitatively*, our framework yields saliency masks significantly closer to GT masks based on Euclidean distances  $D$  and Hausdorff distances  $H$ , both substantiated by p-values below  $2e-12$  for all tested percentiles (cf. Table 1). Considering the map sizes  $A$  (Table 2), we report overall significantly smaller masks, and again p-

values below  $2e-12$  for all percentiles. This behavior changes when the shape-specific regularization parameters  $\lambda_{2,3}$  are relaxed, as exemplified by the ablation study. As shown in the last row of both parts of Table 1, our feature attributions become scattered and noisy. Close inspection of the overlap-values, esp.  $O$  in Table 2, reveals that on average our method’s attributions have a higher overlap with SAL than CAM. This indicates that our results tend to adhere to the dense localization spots of SAL, but alleviate the latter’s noise and interpretation issue described in [2, 3].

*Qualitatively*, as depicted in Fig. 3 (b), our salient regions appear at the circumference of masses which is reasonable w.r.t. the fact that this is the discriminative region for the presence of masses. This is in line with [16], which reports on injection of poorly *circumscribed*, malignant looking masses while transforming healthy cases into pathological ones using a GAN variant. In addition, our method yields more accurate visualizations than CAM and SAL (Fig. 3 first row), i.e. it has a smaller, more precise and more informative feature attribution than these standard visualization methods (Fig. 3 (b)-(d)).

**Conclusion:** We presented a novel, model agnostic, optimization-based, accurate saliency visualization approach tailored to medical images. Image region marginalization was solved by partial convolution based inpainting, treating medical images correctly, thus overcoming the out-of-distribution problem. Our method showed informative and faithful results on public mammography data, and is therefore suitable as a classification interpretation tool in radiological workflows.

## 5. REFERENCES

- [1] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A.W.M. van der Laak, B. van Ginneken, and C.I. Sánchez, “A survey on deep learning in medical image analysis,” *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.
- [2] L.M. Zintgraf, T.S. Cohen, T. Adel, and M. Welling, “Visualizing deep neural network decisions: Prediction difference analysis,” in *Proceedings of ICLR*, 2017.
- [3] R.C. Fong and A. Vedaldi, “Interpretable explanations of black boxes by meaningful perturbation,” in *Proceedings of ICCV*, 2017, pp. 3429–3437.
- [4] P. Dabkowski and Y. Gal, “Real time image saliency for black box classifiers,” in *Proceedings of NIPS*, 2017, pp. 6967–6976.
- [5] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” *arXiv preprint arXiv:1312.6034*, 2013.
- [6] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *Proceedings of CVPR*, 2016, pp. 2921–2929.
- [7] A. Shrikumar, P. Greenside, and A. Kundaje, “Learning important features through propagating activation differences,” in *Proceedings of ICML*, 2017, pp. 3145–3153.
- [8] C.-H. Chang, E. Creager, A. Goldenberg, and D. Duvenaud, “Explaining image classifiers by counterfactual generation,” in *Proceedings of ICLR*, 2019.
- [9] V. Petsiuk, A. Das, and K. Saenko, “RISE: Randomized input sampling for explanation of black-box models,” in *BMVC*, 2018.
- [10] H. Uzunova, J. Ehrhardt, T. Kepp, and H. Handels, “Interpretable explanations of black box classifiers applied on medical images by meaningful perturbations using variational autoencoders,” in *SPIE Medical Imaging*, 2019, vol. 10949, pp. 264–271.
- [11] Z. Lipton, “The mythos of model interpretability,” *ACM Queue*, vol. 16, no. 3, pp. 30:31–30:57, 2018.
- [12] T. Lombrozo, “The structure and function of explanations,” *Trends in Cognitive Sciences*, vol. 10, no. 10, pp. 464–70, 2006.
- [13] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, “Sanity checks for saliency maps,” in *Proceedings of NIPS*, 2018, pp. 9505–9515.
- [14] C. Bermudez, A.J. Plassard, L.T. Davis, A.T. Newton, S.M. Resnick, and B.A. Landman, “Learning implicit brain MRI manifolds with deep learning,” in *SPIE Medical Imaging*, 2018, vol. 10574, pp. 408–414.
- [15] C.F. Baumgartner, L.M. Koch, K.C. Tezcan, J.X. Ang, and E. Konukoglu, “Visual feature attribution using Wasserstein GANs,” in *Proceedings of CVPR*, 2017, pp. 8309–8319.
- [16] A.S. Becker, L. Jendele, O. Skopek, N. Berger, S. Ghafoor, M. Marcon, and E. Konukoglu, “Injecting and removing suspicious features in breast imaging with CycleGAN: A pilot study of automated adversarial attacks using neural networks on small images,” *European Journal of Radiology*, vol. 120, pp. 108649, 2019.
- [17] S. Andermatt, A. Horváth, S. Pezold, and P. Cattin, “Pathology segmentation using distributional differences to images of healthy origin,” in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. 2019, vol. 11383 of *LNCS*, pp. 228–238, Springer Int. Publishing.
- [18] G. Liu, F.A. Reda, K.J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, “Image inpainting for irregular holes using partial convolutions,” in *Proceedings of ECCV*, 2018, pp. 85–100.
- [19] M. Heath, K. Bowyer, D. Kopans, R. Moore, and W.P. Kegelmeyer, “The digital database for screening mammography,” in *Proceedings of IWDM*, 2000, pp. 212–218.
- [20] R.S. Lee, F. Gimenez, A. Hoogi, and D. Rubin, “Curated breast imaging subset of DDSM,” *The Cancer Imaging Archive*, vol. 8, 2016.
- [21] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “MobileNets: Efficient convolutional neural networks for mobile vision applications,” *arXiv preprint arXiv:1704.04861*, 2017.
- [22] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Y. Ding, A. Bagul, C. Langlotz, K.S. Shpanskaya, M.P. Lungren, and A.Y. Ng, “CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning,” *CoRR*, vol. abs/1711.05225, 2017.