Robustness of Medical Imaging Models With Physics-Based Image Quality Degradations

Tyler Ward TYLER.WARD@UKY.EDU
Abdullah Imran AIMRAN@UKY.EDU
University of Kentucky, Lexington, KY 40506, USA

Editors: Under Review for MIDL 2026

Abstract

Variability in the quality of medical images poses a significant challenge to the reliability of deep learning models employed for medical image analysis and computer-aided diagnostic tasks. Ideally, segmentation and classification networks would be trained on clean, well-curated datasets, but oftentimes real clinical data exhibits degradations caused by suboptimal acquisition conditions. To investigate the robustness of popular medical imaging models under such conditions, we synthetically corrupt publicly available X-ray and dermoscopic datasets with degradations based on real-world medical physics and develop a comprehensive evaluation framework with multiple segmentation and classification architectures. In addition, we introduce a corruption-level prediction objective to quantify models' ability to infer image quality directly from degraded inputs. Based on the models' performance across these tasks, we analyze how different corruption levels influence performance and how exposure to physics-informed degradations affects generalizability. Our results demonstrate that models trained with realistic corruption processes exhibit improved robustness in both segmentation and classification accuracy, while also achieving reliable prediction of corruption severity.

Keywords: Data augmentation, Deep learning, Dermoscopy, Image quality, Medical physics, X-ray

1. Introduction

Deep neural networks (DNNs) are commonly employed in modern medical image analysis, offering strong performance on tasks ranging from disease classification to organ segmentation. The success of such models, however, depends heavily on the quality of the images that make up their training data (Imran, 2020). Even slight degradations in image quality can lead to substantial drops in model performance (Sabottke and Spieler, 2020). This is problematic because real-world clinical data rarely conforms to the perfect conditions implicitly assumed by many training pipelines (Urbaniak, 2024), be that because of patient motion, outdated equipment, poor operator technique, or some other suboptimal acquisition parameter. As a result, this mismatch between common training data distributions and real-world image quality remains a fundamental obstacle that hampers the deployability of DNN models in clinical practice.

These concerns are particularly pronounced in X-ray imaging. Unlike natural images, radiographs are acquired through a chain of physical interactions involving the X-ray tube, patient anatomy, and detector system, each of which impacts the final quality of the image

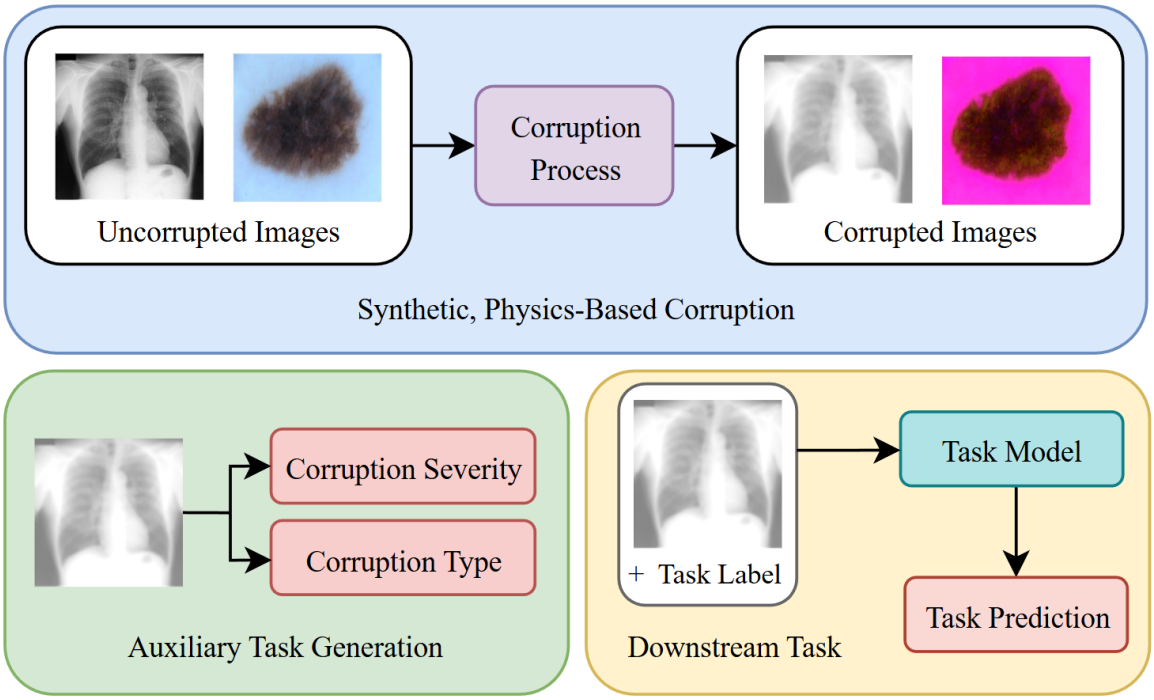

Figure 1: The general workflow of our proposed method. First, we subject both X-ray and dermoscopic images to a series of physics-based corruptions (visualizations shown in Appendix A and C, respectively.) Next, we generate auxiliary classification labels based on these corruptions. Finally, we train our segmentation and classification models on a training set that contains both corrupted and uncorrupted images to predict the desired task label.

in its own way (Tafti and Byerly, 2020). For example, beam energy determines the balance between contrast and penetration, detector characteristics, shape resolution and noise, focal spot size and geometric configuration influence sharpness and magnification, and scatter, collimation, and exposure alter image appearance in clinically impactful ways (Tompe and Sargar, 2020). While it has long been established that proper acquisition parameters are crucial for obtaining high-quality medical images, the influence of suboptimal acquisition parameters on the behavior of modern DNNs is much less known.

In general, data augmentation is one of the most widely used strategies for enhancing the robustness of DNNs to poor image quality (Mintun et al., 2021). Conventional augmentation pipelines apply simple transformations such as rotations, flips, random crops, blur, or brightness adjustments. While effective in enhancing the robustness to corruption in natural images, such methods struggle to capture the physical properties that can introduce quality defects in medical images (Kebaili et al., 2023). Recent work has shown that incorporating domain-specific knowledge into augmentation strategies can lead to more robust models compared to standard approaches (Athalye and Arnaout, 2023), but systematic investigation of physics-based augmentation in medical imaging remains limited.

Addressing this, we systematically explore in this work how physics-based degradations impact both image quality and DNN performance across various medical imaging tasks.

Using publicly available chest X-rays and dermoscopic images, we synthetically generate a variety of realistic, physics-based degradations. We then train and evaluate popular segmentation (U-Net (Ronneberger et al., 2015), Swin-Unet (Cao et al., 2022), SAM (Kirillov et al., 2023)) and classification (ResNet (He et al., 2016), ConvNeXt (Liu et al., 2022), ViT (Dosovitskiy, 2020)) networks on both the original and corrupted datasets to assess how such degradations influence performance. Evaluated tasks include organ and skin lesion segmentation, classification of lung nodules and skin lesions, and prediction of both image quality and corruption types. Our contributions are summarized as:

- Custom datasets of high-quality chest radiographs and dermoscopic with corrupted variants depicting a series of quality degradations inspired by real-world medical physics.
- Evaluation of the impact these physics-based degradations have on the accuracy and robustness of deep learning models across various medical imaging tasks.
- A quantitative comparison between the performance of medical imaging models when trained with traditionally augmented data (blur, noise, brightness) vs. with our physics-based augmentations.

2. Related Work

Because of the large number of parameters that go into their acquisition, X-ray images are a natural choice for evaluating medical physics-based approaches. Generative approaches specifically have shown promise in enhancing the generalizability of medical AI models. For example, SyntheX (Gao et al., 2023) demonstrated that combining realistically simulated X-ray images from human models with contemporary domain generalization techniques was an effective method of teaching ML models to perform reliably on real-world data. This is reinforced by additional work (Prakash et al., 2025) that demonstrated the use of a latent diffusion model to generate synthetic chest X-rays conditioned on text or segmentation masks, which were then used as a form of data augmentation to achieve a 15% improvement on classification tasks, and a 14% improvement on segmentation tasks. Non-generative, but still physics-based approaches, have also been proven effective. It has been shown that forward-projecting 3D CT scans to create synthetic X-rays with controlled geometric distortions improves segmentation performance by 3–15% on real X-rays when these synthetic images are used as augmentation, outperforming models trained only on the real-world data (Fok et al., 2024). Other works have simulated specific radiographic artifacts to increase robustness, such as a realistic pipeline to simulate collimator cut-off and shadows in X-rays to train models that are unfazed by varying field-of-view restrictions (El-Zein et al., 2023).

Similar synthetic corruption methods have also been used to aid in dermoscopic imaging. EnhanceNet-U (Al Mazed et al., 2025) makes use of synthetic degradation datasets that model Gaussian noise, brightness/contrast variations, Gaussian blur, and resolution degradation to better enhance dermoscopic images compared to non-augmented baselines. Another method, DermaSR-GAN (Branciforti et al., 2025), trained on dermoscopic images synthetically corrupted with optical aberrations, sensor noise, and JPEG compression artifacts to improve super-resolution performance. Other work utilized Bézier curves and conditional GANs to improve dermoscopic imaging via the removal of hair from regions of interest (Talavera-Martinez et al., 2020).

As evidenced by the literature, evaluation of synthetically corrupted dermoscopic images for more standard diagnostic classification or segmentation tasks is underexplored. Also underexplored is using classification models to predict the type and level of corruption in X-ray and dermoscopic images. Such predictions could be useful when training generative models to enhance image quality. In our work, we aim to address these underexplored challenges.

3. Methods

We propose a rigorous image corruption pipeline (see Fig. 1) that enables evaluation of model robustness across a variety of downstream tasks. Below, we detail our proposed approach in full.

3.1. Radiograph Corruptions

We design our radiograph corruptions in a way that exploits common vulnerabilities associated with the use of X-ray imaging, as identified by (Tompe and Sargar, 2020). Preliminaries on the physics behind radiograph acquisition which guided our methodology are discussed in Appendix B. All corruptions were implemented in Python using the NumPy and OpenCV libraries.

Acquisition Geometry: We simulated corrupted images, $I_c(x, y)$, that had been affected by suboptimal acquisition geometry by synthetically combining various levels of blurring, magnification, and contrast attenuation. Mathematically, the process we used can be written as:

$$I_c(x, y) = \mathcal{E}_\lambda\big[\,\alpha\,\mathcal{M}_s\big((I * h_\sigma)(x, y)\big) + (1 - \alpha)\,I(x, y)\big]\,, \tag{1}$$

where h_σ is a Gaussian point spread function (PSF) with standard deviation $\sigma \propto U_g$ representing focal spot blur; $\mathcal{M}_s(\cdot)$ represents the process of isotropically scaling then cropping back to original dimensions, simulating varying source-image distances (SIDs) and object-image distances (OIDs); α modulates global contrast to emulate changes in scatter and detector response; and $\mathcal{E}_\lambda[\cdot]$ describes mild edge-enhancement operator parameterized by λ to simulate automatic sharpness correction for mild degradation.

Anti-scatter Grid: In order to simulate a misconfigured anti-scatter grid, we considered scatter contamination, primary signal attenuation, artifacts, and stochastic noise. As a function, our simulation process here can be described by:

$$I_c(x, y) = \eta\Big[(1 - \gamma)I(x, y) + \gamma(I * h_\sigma)(x, y)\Big]\big[1 - \delta\,G_\theta(x, y)\big] + \mathcal{N}(0, \sigma_n^2), \tag{2}$$

where h_σ is a broad Gaussian kernel modeling the low-frequency scatter field, γ controls the proportion of scatter mixed with the primary signal, η is a multiplicative attenuation term representing the absorption of photons by the grid, δ controls the amplitude of the artifact, and $\mathcal{N}(0, \sigma_n^2)$ represents Gaussian noise due to dose-related signal-to-noise (SNR) degradation. $G_\theta(x, y)$ in Equation 2 is calculated as:

$$G_\theta(x, y) = \sin^2\Big(\frac{2\pi}{\Delta}(x\cos\theta - y\sin\theta)\Big). \tag{3}$$

This function encodes the grid artifact pattern on the images with random orientation and spatial frequency.

Beam Energy and pKV: To simulate too much penetration by beam energy, we induce a series of contrast compression, scatter-induced blurring, and brightness scaling. We do this by:

$$I_c(x, y) = \beta((I(x, y) - \bar{I})\alpha + \bar{I}) * h_\sigma, \tag{4}$$

where $I(x, y)$ is the normalized, uncorrupted radiograph and $\bar{I}$ is its mean intensity.

Collimation: To emulate suboptimal collimation, we synthetically add low-frequency scatter contamination, reduce contrast, and inject Gaussian noise. The corrupted image is defined as:

$$I_c(x, y) = [(1 - \gamma)I(x, y) + \gamma(I * h_\sigma)(x, y)]_\alpha + \mathcal{N}(0, \sigma_n^2). \tag{5}$$

Detector Performance: The simulation of declining detector performance consists of three main components: spatial-resolution reduction, modulation transfer function (MTF) attenuation, and additive noise. These components and how they influence the final corrupted images can be represented by:

$$I_c(x, y) = [\mathcal{F}^{-1}(\mathcal{F}(\mathcal{R}_p)[I(x, y)])\mathrm{MTF}_\lambda(u, v))] + \mathcal{N}(0, \sigma_n^2), \tag{6}$$

where $\mathcal{R}_p$ denotes resampling, which simulates enlargement of the detector element;

$$\mathrm{MTF}_\lambda(u, v)$$

is a Butterworth low-pass filter, and $\mathcal{F}/\mathcal{F}^{-1}$ denote forward and inverse Fourier transforms. Together, these variables capture both the fall-off in spatial frequency and stochastic degradation common in underperforming flat-panel detectors.

Focal Spot Size: We simulate realistic variations in focal spot size by modeling the corrupted image as a combination of geometric unsharpness, attenuated edge response, and motion blur:

$$I_c(x, y) = [(I * h_\sigma)(x, y) - \eta\nabla^2 I(x, y)] * M_{\theta,L}, \tag{7}$$

where $\nabla^2 I$ is a Laplacian operation that models reduction in edge definition, and $M_{\theta,L}$ is a motion-blur kernel of length L and direction θ, meant to simulate patient movement from longer exposure.

Tube Current–Exposure Time Product (mAs): To simulate increased mAs due to a reduced tube current, we corrupt the images as follows:

$$I_c(x, y) = [\mathcal{P}(s \cdot I(x, y)) + \mathcal{N}(0, \sigma^2] * M_L, \tag{8}$$

where s is the signal reduction factor simulation photon flux drop; $\mathcal{P}(\cdot)$ denotes Poisson noise, which approximates photon-counting variation; $\mathcal{N}(0, \sigma^2)$ is additive Gaussian noise, and M_L is a horizontal motion blur kernel of size L.

3.2. Dermoscopy Corruptions

As with the radiograph corruptions, we design our dermoscopic corruptions in a manner that exploits common physics-based degradations in dermoscopic image quality, as identified by (Avanaki et al., 2020). The corruptions were implemented in Python's NumPy and OpenCV libraries, are are discussed below.

Blur: Oftentimes in dermoscopy images, blur can arise due to the camera being out of focus or because of the dermatoscope or skin motion. To simulate this, we do:

$$I_c(x, y) = \alpha(I * h_\sigma)(x, y) + (1 - \alpha)(I * M_{\theta,L}(x, y). \tag{9}$$

For higher severity levels, we also introduce a spatially varying blur field by defining a radial mask $w(x, y)$ that increases towards the periphery, which reflects degradation from focal-plane curvature.

Color Reproduction Accuracy: In dermoscopic imaging, color reproduction accuracy is defined as the differences between the colors directly measured on the skin and the corresponding colors rendered on the visualization device in CIE ΔE2000. To corrupt the dermoscopic images around this fact, we introduce controlled shifts to hue, saturation, lightness, and channel mixing to simulate deviations between true skin color and device-rendered appearance. The transformations we employ are expressed:

$$I_c = T_{HSV}(MI_{RGB}), \tag{10}$$

where T_{HSV} encodes nonlinear hue and saturation shifts, M represents a linear color transformation to the RGB channels of the original image (I_{RGB}).

Focal Plane Flatness: To enforce regions with differing sharpness due to curvature or irregularity in the focal surface, we model the corrupted image as:

$$I_c(x, y) = (I * h_{\sigma(x,y)}(x, y). \tag{11}$$

At mild and moderate levels of corruption, $\sigma(x, y)$ follows a radial profile that increases with normalized distance $d(x, y)$ from the image center, simulating a smoothly curved focal place with a focused central region. At higher levels of corruption, $\sigma(x, y)$ incorporates a stochastic component $n(x, y)$ blended with the radial baseline to produce uneven focal zones, and a locally focused mask reduces blur within a random region to emulate nonuniform lens-to-skin distance.

Resolution: To model reduced spatial resolution, we first downsample the original image by a severity-dependent factor r, then upsample back to the original dimensions. Mathematically, the operation is represented as:

$$I_c = U(D_r(I)) \tag{12}$$

where $D_r(\cdot)$ reduces the image size by r, and U resotres it to the original resolution via bilinear interpolation. Doing this emulates loss of fine details associated with the limited sensor pixel density of optical resolving power.

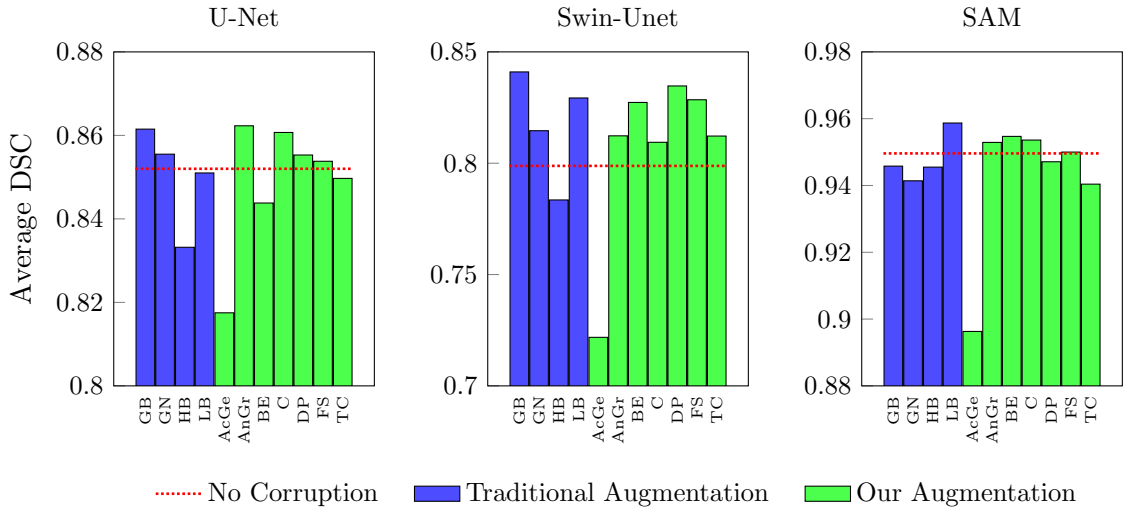

Figure 2: Visual comparison of the segmentation performance of physics-based augmentation vs. standard data augmentation on the JSRT chest x-ray dataset. Acronyms —GB: Gaussian blur, GN: Gaussian noise, HB: High brightness, LB: Low brightness, AcGe: Acquisition geometry, AnGr: Antiscatter grid, BE: Beam energy, C: Collimation, DP: Detector performance, FS: Focal spot, TC: Tube current.

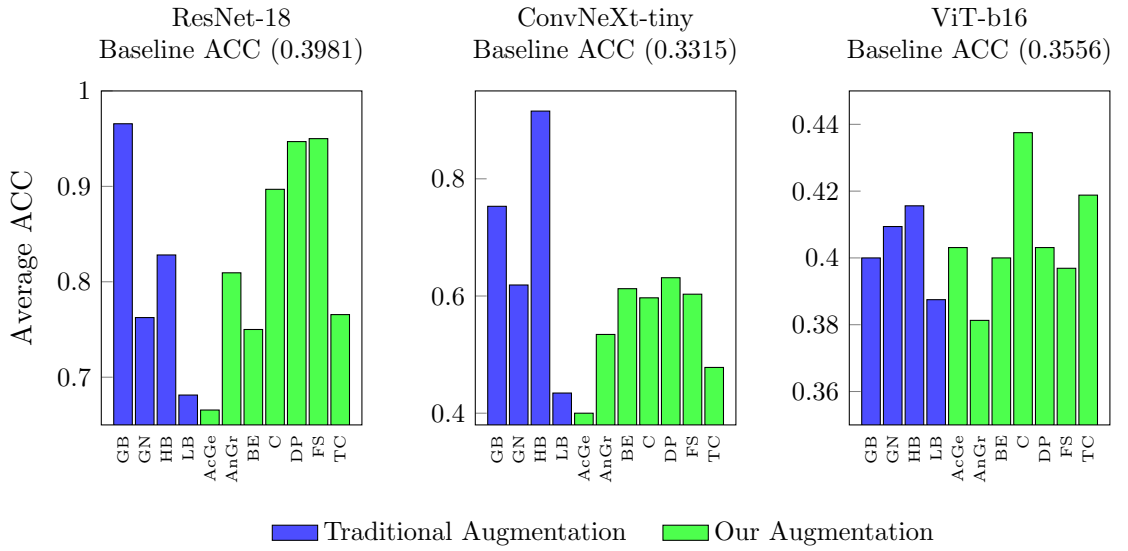

Figure 3: Visual comparison of the classification performance of physics-based augmentation vs. standard data augmentation on the JSRT chest x-ray dataset. Acronyms —GB: Gaussian blur, GN: Gaussian noise, HB: High brightness, LB: Low brightness, AcGe: Acquisition geometry, AnGr: Antiscatter grid, BE: Beam energy, C: Collimation, DP: Detector performance, FS: Focal spot, TC: Tube current.

Table 1: Performance evaluation of the classifiers in predicting corruption severity and type on the JSRT dataset.

Corruption Severity Prediction			
Corruption Severity	ResNet-18	ConvNeXt-tiny	ViT-b16
No Corruption	0.5466	0.000	0.0000
Level 1	0.8034	0.9451	0.6850
Level 2	0.8670	0.5289	0.2659
Level 3	0.8797	0.7139	0.2052
Level 4	0.9983	0.6734	0.5434
Overall	0.8190	0.5723	0.3399
Corruption Type Prediction			
Corruption Type	ResNet-18	ConvNeXt-tiny	ViT-b16
No Corruption	0.5466	0.0000	0.0000
Acquisition Geometry	0.8745	0.7650	0.4750
Antiscatter Grid	1.0000	0.9742	0.5515
Beam Energy	0.9585	0.9594	0.4924
Collimation	0.8654	0.8594	0.3177
Detector Performance	0.2398	0.8594	0.2500
Focal Spot	0.7702	0.2537	0.2780
Tube Current	0.8016	0.9750	0.6100
Overall	0.8445	0.6283	0.3718

4. Experimental Evaluation

4.1. Implementation Details

We validate our proposed approach on the publicly available Standard Digital Image Database of chest radiographs with and without a lung nodule from the Japanese Society of Radiological Technology (JSRT) (Shiraishi et al., 2000), as well as the 2016 International Skin Imaging Collaboration (ISIC) Challenge Dataset (Gutman et al., 2016). Additional dataset details can be found in Appendix D.To evaluate the robustness of medical imaging models with physics-based image quality degradations, we trained a variety of popular segmentation and classification models.

For segmentation, we utilized U-Net (Ronneberger et al., 2015) and a successful variant, Swin-Unet (Cao et al., 2022) to represent segmentation models based on CNNs and transformers, respectively. We also explored the performance of segmentation foundation models via the use of SAM (Kirillov et al., 2023). For classification, to showcase the performance of CNN-based methods, we trained ResNet (He et al., 2016) and ConvNeXt (Liu et al., 2022). To gauge the performance of transformers, we used ViT (Dosovitskiy, 2020).

All classification models as well as the non-SAM segmentation models were trained for 25 epochs. For the experiment with SAM, we trained for just 10 epochs due to the demonstrated ability of SAM to converge faster than non-foundation models (Ward and Imran, 2025; Ward et al., 2025). The models are trained on an *Intel (R) Xeon (R) w7-2475X, 2600MHz* machine with a dual *NVIDIA A4000X2* GPU (32GB). For the segmentation

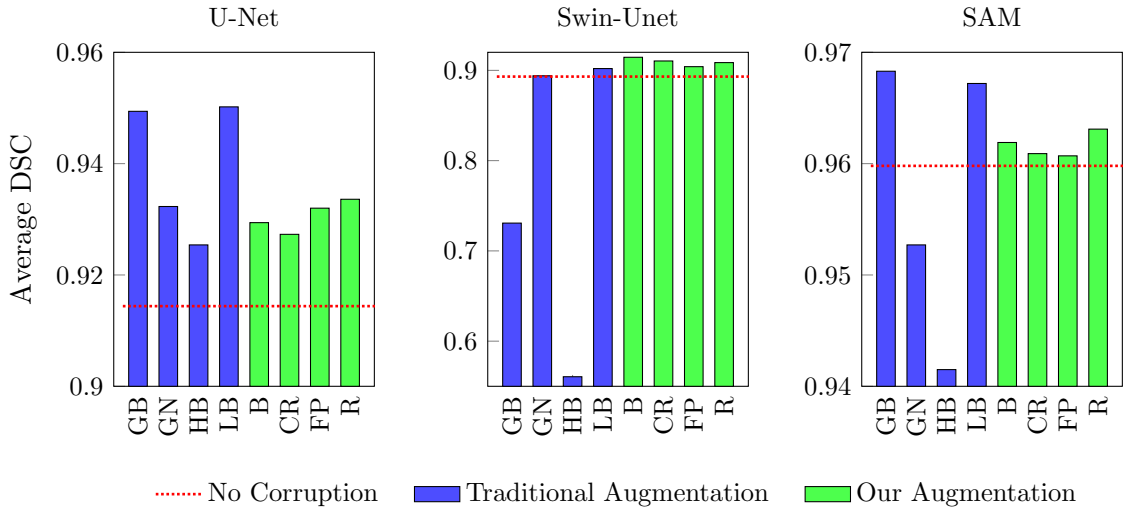

Figure 4: Visual comparison of the segmentation performance of physics-based augmentation vs. standard data augmentation on the ISIC dermoscopy dataset. Acronyms —GB: Gaussian blur, GN: Gaussian noise, HB: High brightness, LB: Low brightness, B: Blur, CR: Color reproduction, FP: Focal plane, R: Resolution.

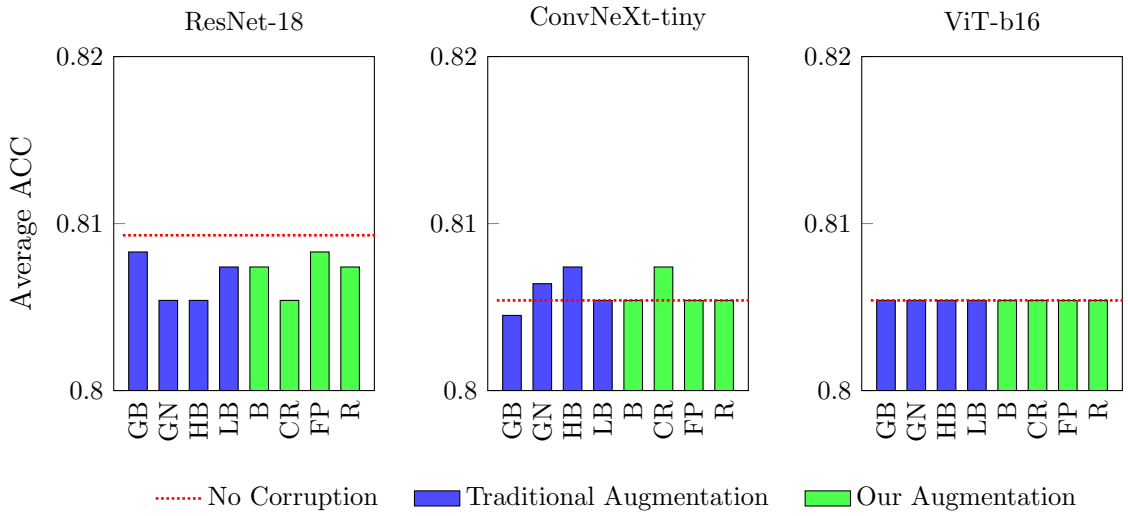

Figure 5: Visual comparison of the classification performance of physics-based augmentation vs. standard data augmentation on the ISIC dermoscopy dataset. Acronyms —GB: Gaussian blur, GN: Gaussian noise, HB: High brightness, LB: Low brightness, B: Blur, CR: Color reproduction, FP: Focal plane, R: Resolution.

tasks, we employ the Dice-Sørensen coefficient as the evaluation metric. For the classification tasks, accuracy is used as the evaluation metric.

4.2. Results and Discussion

Segmentation Performance: To evaluate the influence of physics-based degradations on segmentation robustness, we examined the multi-class segmentation performance on the JSRT dataset of the three segmentation models previously discussed. Figure 2 provides a visual depiction of the impact of adding corrupted images to the training data compared to when only uncorrupted images are used in the training set. When no corrupted images are present, the model achieves average Dice scores of 0.8520 (U-Net), 0.7988 (Swin-Unet), and 0.9496 (SAM).

As evidenced by Figure 2, the performance of the model generally increases as corrupted data is added to the training set, even as the corruption level increases. The performance improvements offered by our physics-based augmentations are most apparent for Swin-Unet, although that is also the case for the traditional augmentations. The difference between traditional augmentation and physics-based augmentation is clearer on U-Net and SAM, where the physics-based augmentation generally outperforms traditional augmentation, particularly on SAM, where three out of four of the traditional augmentations fail even to surpass the baseline model trained on only uncorrupted images. These results are validated by the performance of physics-based augmentation vs. traditional methods on the dermoscopic dataset, whose results are shown in Figure 4.

Classification Performance: Figure 3 shows the performance of ResNet-18, ConvNeXt-tiny, and ViT-b16 when trained on uncorrupted, traditional, and physics-augmented x-ray images. Adding corrupted images to the training set boosts the accuracy of the models, especially for ResNet-18, which performs the best. However, the physics-based augmentations struggle compared to traditional methods on ConvNeXt-tiny, probably because the traditional methods' corruptions are so slight that the model is essentially just being fed additional training samples it has already seen. Table 1 shows the performance of the classifiers on the auxiliary classification tasks. ResNet-18 demonstrates strong performance on both tasks, while ConvNeXt-tiny and ViT-b16 struggle. Our hypothesis for this is that the latter two models are much bigger than ResNet-18, and they may struggle to converge as quickly as a result. Figure 5 shows the performance of the classifiers on the ISIC dataset. Performance is remarkably consistent across all of the experiments, with no clear winner among the models. The reason for this is the heavy class imbalance, with there being far more benign samples in the dataset than malignant, so each of the models performs very well on the benign samples, and very poorly on the malignant ones.

5. Conclusions

In this work, we have systematically evaluated how physics-based image degradations affect the robustness of segmentation foundation models in medical imaging. By generating synthetic corruptions that mimic real-world variations in image quality across both chest radiographs and dermoscopic images, we have demonstrated that degradation-aware training can enhance the generalization ability of both segmentation and classification models.

References

Abdullah Al Mazed, Md Faiyaj Ahmed Limon, Shahidul Haque Thouhid, Md Fazle Hasan Shiblee, Shubradeb Das, Md Shahid Iqbal, and Debojyoti Biswas. A comprehensive deep learning approach for dermoscopic image enhancement. *Franklin Open*, page 100405, 2025.

Chinmayee Athalye and Rima Arnaout. Domain-guided data augmentation for deep learning on medical imaging. *PLOS One*, 18(3):e0282532, 2023.

A Avanaki, K Espig, Albert Xthona, D Brooks, J Young, and Tom Kimpe. Perceptual image quality in digital dermoscopy. In *Medical Imaging 2020: Image Perception, Observer Performance, and Technology Assessment*, volume 11316, pages 235–241. SPIE, 2020.

Francesco Branciforti, Kristen M Meiburger, Elisa Zavattaro, Paola Savoia, and Massimo Salvi. Leveraging synthetic degradation for effective training of super-resolution models in dermatological images. *Electronics*, 14(15):3138, 2025.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European Conference on Computer Vision*, pages 205–218. Springer, 2022.

G Alm Carlsson, CA Carlsson, B Nielsen, and J Persliden. Generalised use of contrast degradation and contrast improvement factors in diagnostic radiology. application to vanishing contrast. *Physics in Medicine & Biology*, 31(7):737, 1986.

Nikunj Desai, Abhinav Singh, and Daniel J Valentino. Practical evaluation of image quality in computed radiographic (cr) imaging systems. In *Medical Imaging 2010: Physics of Medical Imaging*, volume 7622, pages 1498–1507. SPIE, 2010.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Benjamin El-Zein, Dominik Eckert, Thomas Weber, Maximilian Rohleder, Ludwig Ritschl, Steffen Kappler, and Andreas Maier. A realistic collimated X-ray image simulation pipeline. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 137–145. Springer, 2023.

Wai Yan Ryana Fok, Andreas Fieselmann, Christian Huemmer, Ramyar Biniazan, Marcel Beister, Bernhard Geiger, Steffen Kappler, and Sylvia Saalfeld. Adversarial robustness improvement for X-ray bone segmentation using synthetic data created from computed tomography scans. *Scientific Reports*, 14(1):25813, 2024.

Cong Gao, Benjamin D Killeen, Yicheng Hu, Robert B Grupp, Russell H Taylor, Mehran Armand, and Mathias Unberath. Synthetic data accelerates the development of generalizable learning-based algorithms for X-ray image analysis. *Nature Machine Intelligence*, 5(3):294–308, 2023.

Edward Gumbrell, JM McNaney, CM Huntington, AG Krygier, and H-S Park. Characterizing the modulation transfer function for x-ray radiography in high energy density experiments. *Review of Scientific Instruments*, 89(10), 2018.

David Gutman et al. Skin lesion analysis toward melanoma detection: A challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC). *arXiv:1605.01397*, 2016.

Pascal Hannequin and Jacky Mas. Statistical and heuristic image noise extraction (SHINE): A new method for processing poisson noise in scintigraphic images. *Physics in Medicine & Biology*, 47(24):4329, 2002.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Abdullah-Al-Zubaer Imran. From fully-supervised, single-task to scarcely-supervised, multi-task deep learning for medical image analysis, 2020.

Aghiles Kebaili et al. Deep learning approaches for data augmentation in medical imaging: A review. *Journal of Imaging*, 9(4):81, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

Jeonghee Lee, Chang Hwy Lim, Jong-Won Park, Ik-Hyun Kim, Myung Kook Moon, and Yong-Kon Lim. The effect of grid ratio and material of anti-scatter grid on the scatter-to-primary ratio and the signal to-noise ratio improvement factor in container scanner X-ray imaging. *Journal of Radiation Protection and Research*, 42(4):197–204, 2017.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

Albert Macovski. *Medical Imaging Systems*. Information and Systems Science Series. Prentice-Hall, Inc., Englewood Cliffs, New Jersey 07632, 1983.

Cynthia H McCollough. The AAPM/RSNA physics tutorial for residents. x-ray production. *Radiographics*, 17(4):967–984, 1997.

Eric Mintun, Alexander Kirillov, and Saining Xie. On interaction between augmentations and corruptions in natural corruption robustness. *Advances in Neural Information Processing Systems*, 34:3571–3583, 2021.

Eva Prakash, Jeya Maria Jose Valanarasu, Zhihong Chen, Eduardo Pontes Reis, Andrew Johnston, Anuj Pareek, Christian Bluethgen, Sergios Gatidis, Cameron Olsen, Akshay S Chaudhari, et al. Evaluating and improving the effectiveness of synthetic chest X-rays

for medical image analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4413–4421, 2025.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

Carl F Sabottke and Bradley M Spieler. The effect of image resolution on deep learning in radiography. *Radiology: Artificial Intelligence*, 2(1):e190015, 2020.

Beth A. Schueler. Clinical applications of basic X-ray physics principles. *Radiographics*, 18 (3):731–744, 1998.

Junji Shiraishi et al. Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *Am. J. Roentgenol.*, 174(1):71–74, 2000.

Amin Tafti and Doug W Byerly. X-ray image acquisition. 2020.

Lidia Talavera-Martinez, Pedro Bibiloni, and Manuel Gonzalez-Hidalgo. Hair segmentation and removal in dermoscopic images using deep learning. *IEEE Access*, 9:2694–2704, 2020.

F Timischl. The contrast-to-noise ratio for image quality evaluation in scanning electron microscopy. *Scanning*, 37(1):54–62, 2015.

Aparna Tompe and Kiran Sargar. X-ray image quality assurance. 2020.

Ilona Anna Urbaniak. Using compressed JPEG and JPEG2000 medical images in deep learning: A review. *Applied Sciences*, 14(22):10524, 2024.

Tyler Ward and Abdullah Al Zubaer Imran. Annotation-efficient task guidance for medical segment anything. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–4. IEEE, 2025.

Tyler Ward, Meredith K Owen, O'Kira Coleman, Brian Noehren, and Abdullah-Al-Zubaer Imran. Autoadaptive medical segment anything model. *arXiv preprint arXiv:2507.01828*, 2025.

Mark B Williams, Elizabeth A Krupinski, Keith J Strauss, William K Breeden III, Mark S Rzeszotarski, Kimberly Applegate, Margaret Wyatt, Sandra Bjork, and J Anthony Seibert. Digital radiography image quality: Image acquisition. *Journal of the American College of Radiology*, 4(6):371–388, 2007.

Amirkoushyar Ziabari, Mohamed Hakim Bedhief, Obaidullah Rahman, Singanallur Venkatakrishnan, Paul Brackman, and Peter Katuch. Combining deep learning and *scatterControl* for high-throughput X-ray CT based non-destructive characterization of large-scale casted metallic components. *Journal of Nondestructive Evaluation*, 44(4):148, 2025.

Appendix A. JSRT Dataset Corruptions

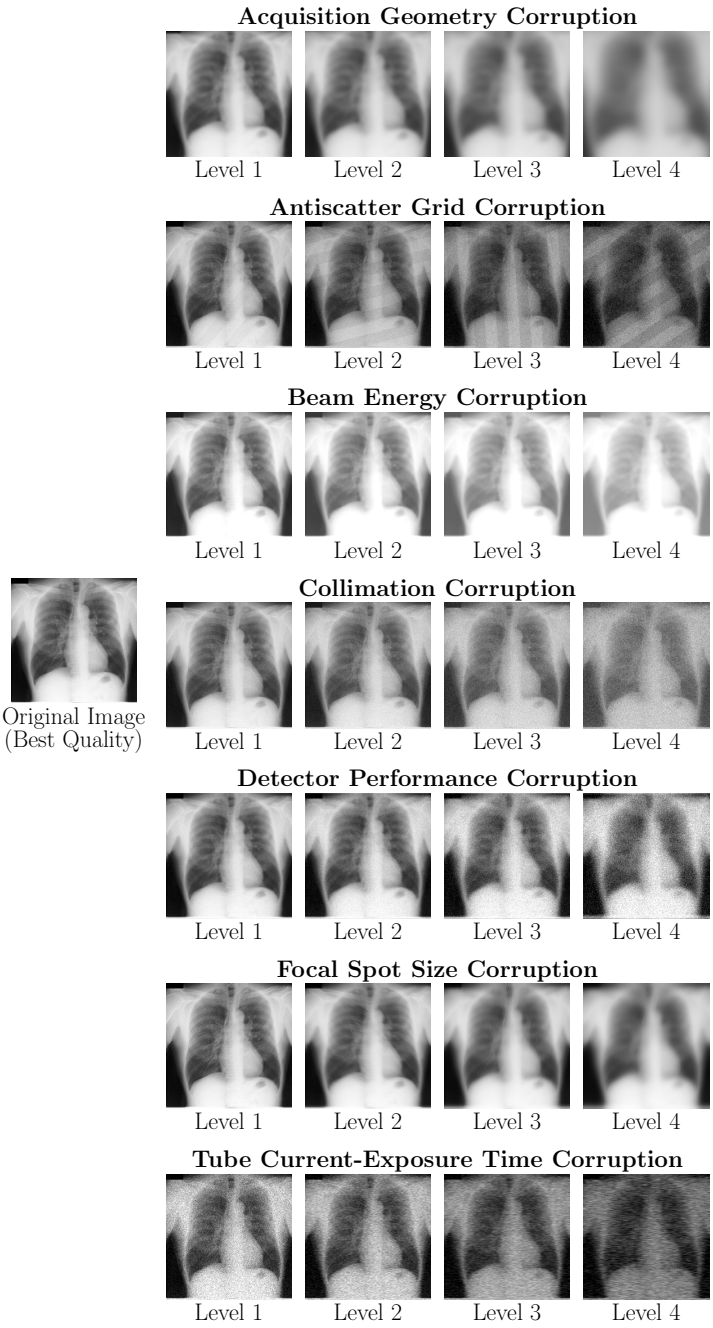

Figure 6: Examples of the physics-based image quality degradations applied to chest radiographs. The first row shows the original high-quality image, followed by four levels of increasing corruption severity (1–4) for each degradation.

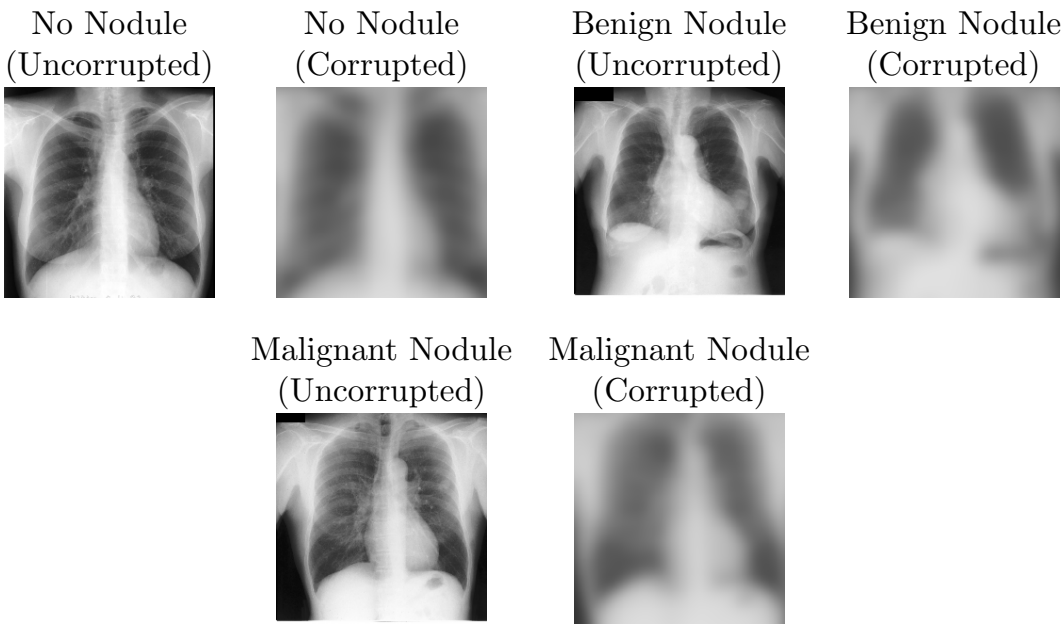

Figure 7: Example class-wise visualizations from the JSRT dataset, showing both uncorrupted and corrupted variants of the images. The corrupted images represent the effects of the most severe acquisition geometry corruption, arbitrarily selected because it is the most visually apparent.

Appendix B. Preliminaries — X-Ray Corruptions

B.1. Physical Properties of X-Ray Radiographs

B.1.1. Fundamentals

X-ray projection imaging is governed by the Beer–Lambert law, which describes exponential photon attenuation as X-rays traverse matter:

$$I(x, y) = I_0 \exp\left[-\int \mu(x, y, z)\, dz\right], \tag{13}$$

where I_0 is the incident photon fluence, and $\mu(x, y, z)$ is the linear attenuation coefficient of the tissue at location (x, y, z) (Macovski, 1983). Detectors measure the cumulative attenuation along each ray path, forming a two-dimensional projection image. Image quality depends on multiple physical parameters, including the beam geometry, spectrum, scatter control, and detector response, which jointly influence the spatial resolution, contrast, and signal-to-noise ratio (SNR) of the resulting image (Macovski, 1983; Williams et al., 2007).

B.1.2. Imaging Parameters

Acquisition Geometry: The geometric configuration of an X-ray system determines the image's magnification, spatial resolution, and geometric distortion (Schueler, 1998). Important metrics in the geometry of the system are the object-image distance (OID), source-image distance (SID), and source-object distance (SOD), which are the distances between the patient and the image receptor, focal spot and image receptor, and focal spot and the patient, respectively. The magnification, M, of the image is defined as the quotient of the SID and SOD:

$$M = \frac{SID}{SOD}, \tag{14}$$

while the amount of blur, Bf_i, in the image is defined by:

$$Bf_i = F \times \frac{OID}{SOD}. \tag{15}$$

In clinical practice, the goal is to achieve the best balance between these factors to achieve the best image quality possible.

Anti-scatter Grid: During imaging, X-ray photons undergo Compton scattering within the patient, producing secondary photons that reach the detector from off-axis directions (Macovski, 1983). This scatter degrades image contrast by introducing a nearly uniform background intensity. It can be mitigated via the use of an anti-scatter grid, which preferentially transmits primary radiation and absorbs obliquely scattered photons.

The ability of this grid to reduce scatter is characterized by the grid ratio (R):

$$R = \frac{h}{w}, \tag{16}$$

where h and w are the grid height and width, respectively (Lee et al., 2017). A contrast improvement factor (CIF), defined by:

$$CIF = \frac{1 + q_s/q_p}{1 + q_{sg}/q_{pg}} \frac{C_{pg}}{C_p}, \tag{17}$$

quantifies the gain in subject contrast achieved by the grid (Carlsson et al., 1986). Here, C represents contrast, q is a physical quantity, p and s refer to primary and scattered radiation, and g refers to the grid. Improper alignment of the grid can introduce artifacts and quantum mottle to the images.

Beam Energy and Peak Kilovoltage (kVp): The beam energy spectrum of an X-ray system is primarily determined by the applied kVp (Williams et al., 2007; McCollough, 1997). Higher kVp increases photon energy, enhancing tissue penetration but reducing contrast because the attenuation coefficients of different tissues become similar at higher energies (Schueler, 1998). Conversely, lower kVp produces stronger differential attenuation and higher contrast at the expense of dose and noise. In practice, the beam spectrum is filtered to remove low-energy photons that contribute to dose but not diagnostic signal. The optimal kVp depends on the anatomical region, balancing penetration, contrast, and patient exposure.

Collimation: Beam collimation confines the X-ray field to the region of diagnostic interest. By restricting the irradiated area, collimation reduces patient dose and the amount of scatter reaching the detector (Schueler, 1998). Scatter intensity approximately scales with the irradiated volume, meaning that tighter collimation also improves contrast-to-noise ratio (CNR), defined as:

$$CNR = \frac{S_A - S_B}{\sigma_o}, \tag{18}$$

where S_A and S_B are signal intensities for signal producing structures A and B in the region of interest and σ_o is the standard deviation of pure image noise (Desai et al., 2010; Timischl, 2015).

Detector Performance: Detectors convert incident X-ray photons into measurable electronic signals. Such detectors govern both the spatial resolution and noise characteristics of the image. Achieving the optimal performance of a detector relies on several factors, including the size of the detector element, the resolution of the detector, and the detector's SNR performance (Tompe and Sargar, 2020). Smaller detector elements improve sampling resolution, but may reduce the photon capture efficiency, which in turn has a detrimental effect on SNR. Maintaining a high SNR is preferred because it indicates a clearer signal compared to background noise (Ziabari et al., 2025). The MTF, which can be expressed as the image contrast values for spatial frequencies over the native object contrast (Gumbrell et al., 2018), is the primary measure of detector resolution. Higher MTF values at larger frequencies indicate stronger abilities of the detector to reproduce fine details.

Focal Spot Size: The focal spot is the region on the anode where electrons strike to produce X-rays. Larger focal spots increase geometric unsharpness and degrade spatial resolution, while smaller focal spots improve sharpness but limit tube loading capacity, at the same time increasing susceptibility to motion blur (Schueler, 1998).

Tube Current–Exposure Time Product (mAs): The mAs product determines the total number of photons contributing to image formation. The detected signal follows Poisson statistics, where photon noise variance scales with photon fluence (Hannequin and Mas, 2002). Reducing mAs decreases radiation dose but increases noise, producing grainier images. When tube current is limited, longer exposure times may be used to maintain photon fluence, at the risk of motion blur due to patient movement (Schueler, 1998).

Appendix C. ISIC Dataset Corruptions

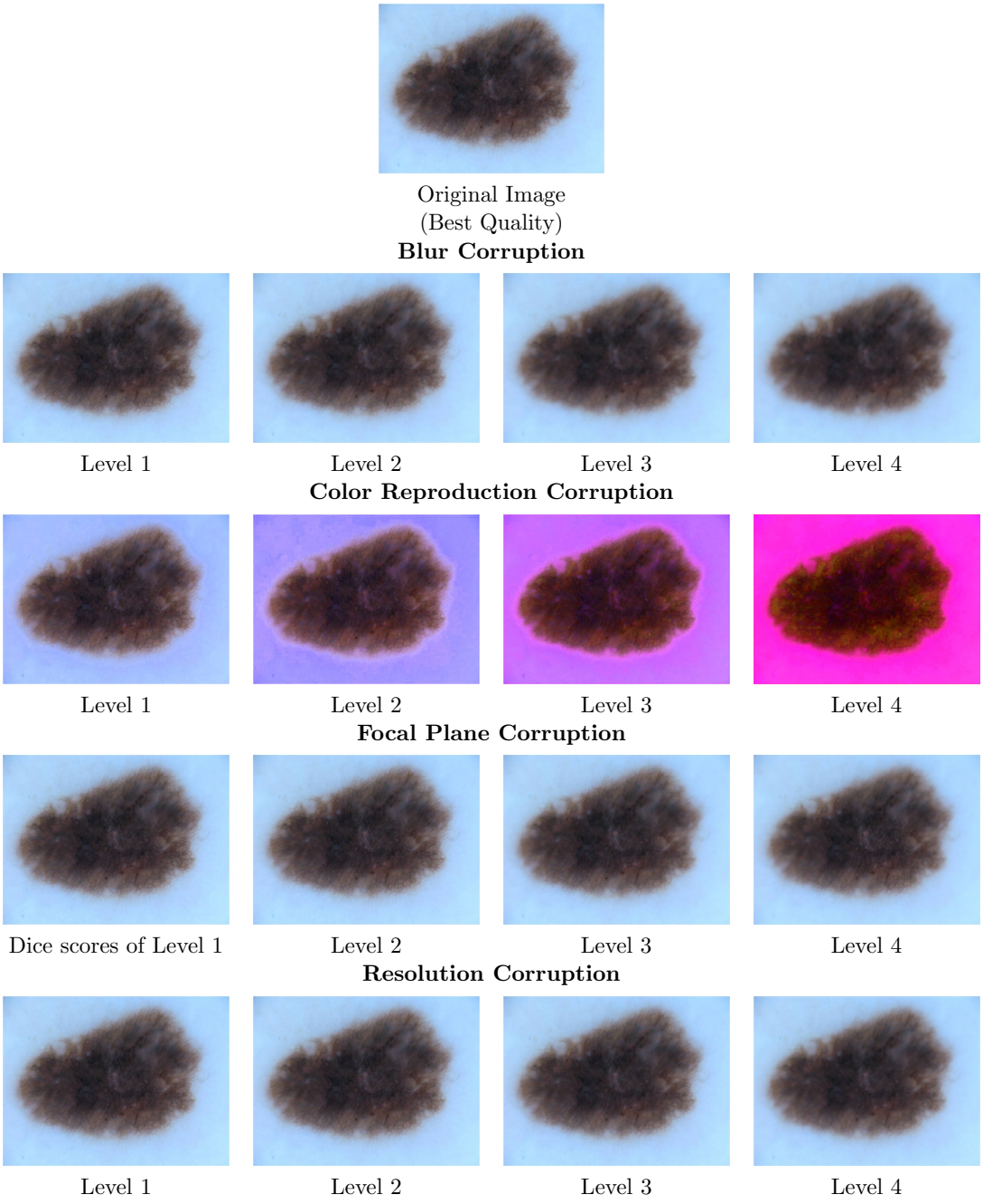

Figure 8: Examples of the physics-based image quality degradations applied to dermoscopic images. The first row shows the original high-quality image, followed by four levels of increasing corruption severity (1–4) for each degradation.

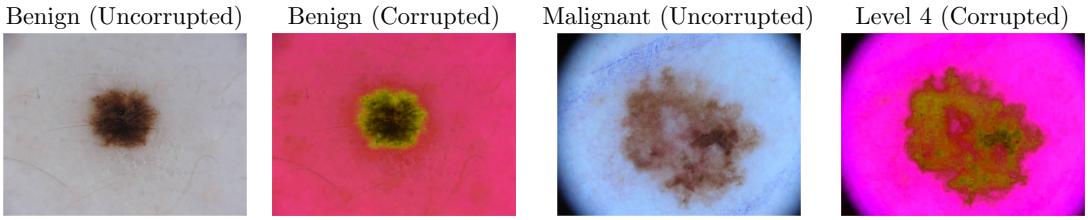

Figure 9: Example class-wise visualizations from the ISIC dataset, showing both uncorrupted and corrupted variants of the images. The corrupted images represent the effects of the most severe color reproduction corruption, arbitrarily selected because it is the most visually apparent.

Appendix D. Additional Data Details

The JSRT dataset contains a total of 247 chest radiographs with expert-labeled annotations that can be used to train both classification (no lesion, benign lesion, malignant lesion) and segmentation (heart, left clavicle, left lung, right clavicle, right lung) models. We separate these radiographs into train/val/test sets using a split of 200/7/40, employing a simple sequential split for the segmentation task, and a balanced stratified sampling strategy for the classification task to ensure that each class is represented in each split. The ISIC 2016 dataset contains a total of 1,279 demoscopic lesion images in JPEG format, which we divided into train/val/test sets using a 895/128/256 split. These dermoscopic images came with ground truth segmentation masks for the lesions, as well as image-level benign vs. malignant information.