Abbreviated Title Page

Title:

Evaluation of Resolution-Aware Training Strategies for Deep Learning Detection of Calcaneus Fractures on X-Ray

Authors:

Nicholas J. Yee MD[1,4,5,6], Atta Taseh MD[1], Samir Ghandour MD[1], Evan Sirls[1], Mansur Halai MBChB[3,4], Cari Whyne PhD[4,5,6], Christopher W. DiGiovanni MD[1,2], John Y. Kwon MD[1,2], Soheil Ashkani-Esfahani MD MPH[1,2]

Affiliations:

1- Foot & Ankle Research and Innovation Lab (FARIL), Department of Orthopaedic Surgery, Mass General Brigham, Harvard Medical School, Boston, MA
2- Foot and Ankle Division, Department of Orthopaedic Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, MA
3- Division of Orthopaedic Surgery, St. Michael's Hospital, University of Toronto, Toronto, ON
4- Division of Orthopaedic Surgery, Department of Surgery, Temerty Faculty of Medicine, University of Toronto, Toronto, ON
5- Institute of Biomedical Engineering, Faculty of Applied Science and Engineering, University of Toronto, Toronto, ON
6- Orthopaedic Biomechanics Lab, Holland Bone and Joint Program, Sunnybrook Research Institute, University of Toronto, Toronto, ON

Article Type:

Original Research

Summary statement:

Different deep learning training strategies affect performance in detecting calcaneus fractures on radiographs across in- and out-of-distribution image resolutions, with a multi-scale augmentation strategy conferring the greatest overall performance improvement in a single model.

Key points:

1. Training strategies addressing differences in radiograph image resolution (or pixel dimensions) could improve deep learning performance.
2. The highest average performance across different image resolutions in a single model was achieved by multi-scale augmentation, where the sampled training dataset is uniformly resized between square resolutions of 128×128 to 900×900.
3. Compared to model training on a single image resolution, sequentially training on increasingly higher resolution images up to 900x900 (i.e., curriculum learning) resulted in higher fracture detection performance on images resolutions between 128x128 and 2048x2048.

Abstract

Purpose:

To evaluate convolutional neural network (CNN) model training strategies that optimize the performance of calcaneus fracture detection on radiographs at different image resolutions.

Materials and Methods:

This retrospective study included foot radiographs from a single hospital between 2015 and 2022 for a total of 1,775 x-ray series (551 fractures; 1,224 without) and was split into training (70%), validation (15%), and testing (15%). ImageNet pre-trained ResNet models were fine-tuned on the dataset. Three training strategies were evaluated: 1) single size: trained exclusively on 128x128, 256x256, 512x512, 640x640, or 900x900 radiographs (5 model sets); 2) curriculum learning: trained exclusively on 128x128 radiographs then exclusively on 256x256, then 512x512, then 640x640, and finally on 900x900 (5 model sets); and 3) multi-scale augmentation: trained on x-ray images resized along continuous dimensions between 128x128 to 900x900 (1 model set). Inference time and training time were compared.

Results:

Multi-scale augmentation trained models achieved the highest average area under the Receiver Operating Characteristic curve of 0.938 [95% CI: 0.936 - 0.939] for a single model across image resolutions compared to the other strategies without prolonging training or inference time. Using the optimal model sets, curriculum learning had the highest sensitivity on in-distribution low-resolution images (85.4% to 90.1%) and on out-of-distribution high-resolution images (78.2% to 89.2%). However, curriculum learning models took significantly longer to train (11.8 [IQR: 11.1–16.4] hours; $P<.001$).

Conclusion:

While 512x512 images worked well for fracture identification, curriculum learning and multi-scale augmentation training strategies algorithmically improved model robustness towards different image resolutions without requiring additional annotated data.

Main Body

Introduction

Calcaneus fractures are challenging orthopaedic injuries to diagnose on radiographs due to the complex anatomy of the hindfoot and potential for subtle or minimally displaced fracture patterns. Conventional x-rays remain the primary imaging modality in acute settings, yet their diagnostic performance can be limited by the complex hindfoot osteology. Recent developments in using deep learning (DL) to detect fracture patterns on radiographs have demonstrated performance comparable to diagnostic radiologists, suggesting a potential opportunity for an automated radiology support tool in identifying calcaneus fractures (1–3).

The image dimensions of diagnostic radiographs are typically 2000 to 3000 pixels (4). Radiograph dimensions and their associated resolution depend on the x-ray device quality and its usage setting, and configuration. The source-to-object and object-to-detector distances will affect the radiograph resolution. DL model performance is dependent on the radiograph image resolutions where higher resolution images enable the detection of fine-grained features, which can be essential in identifying subtle cortical disruptions in minimally displaced fractures. However, increasing resolution imposes greater computational costs (i.e. graphics processing unit memory) and model training time (5–7). Sabottke and Spieler found that the diagnostic performance of convolutional neural networks (CNN), which is a popular DL architecture that hierarchically extracts spatial features from input images by applying a series of sliding kernels, generally improves with increasing resolution with diminishing returns beyond 448×448 pixels for thoracic pathologies (5–7). Low-resolution images obscure the details, limiting diagnostic interpretation. To our knowledge, there are no studies reporting the optimal image resolution or exploring image resolution-aware training strategies for CNN-based fracture identification on radiographs.

Given the anatomical complexity of the hindfoot and the diagnostic challenge posed by subtle calcaneus fractures, DL algorithms are assumed to enable diagnosis with a higher accuracy and consistency. Thus, investigating methods to improve the performance of DL algorithms could bring about a significant added value. In this study, we hypothesized that there are optimal training strategies that address the diversity of image resolutions to improve DL models in identifying calcaneus fractures. The purpose of this study is to investigate three training strategies to balance the trade-offs between the model performance across image resolutions and training/inference computational resources.

Materials and Methods

This was a retrospective study conducted at a single academic medical center. Institutional review board approval was received (IRB no. 2015P000464). Anonymized diagnostic radiology imaging data were collected from the institution's imaging archive between 2015 and 2022 with a total of 1,775 foot X-ray series (AP, oblique, and lateral) were included: 551 series with a calcaneus fracture and 1,224 without a fracture. Females represented 43% of the patients (60 ± 17 years old) and males was 57% (53 ± 15 years old). The foot radiograph dimensions were 1942 ± 992 pixels in height by 2114 ± 767 pixels in width.

The classification of fracture status was based on the radiology reports by board-certified diagnostic radiologists. DL models were pre-trained on ImageNet and fine-tuned on the calcaneus training dataset (8). The dataset was partitioned randomly by image series (70% training, 15% validation, 15% test). Training image augmentation included random horizontal flipping (probability, P = 0.5); random rotation up to ±90° (P = 0.25); random affine transformation including rotation (±10°), scaling (0.9–1.1×), translation (up to 5% in both axes), and shear (up to 5°) (P = 0.3); Gaussian noise addition with mean 0.0

and standard deviation 0.1 (P = 0.3); and color jittering adjusting brightness, contrast, and saturation (±0.2) and hue (±0.1) (P = 0.3).

A model set included ResNet-18, -34, -50, and -101 and trained using PyTorch (version 2.4.1+cu121; https://pytorch.org/) and TorchVision (version 0.19.1+cu121; https://docs.pytorch.org/vision) on either a NVIDIA RTX 4090 or the high-performance computing cluster (HPC) with NVIDIA Tesla V100 graphics processing units (NVIDIA, California, USA). The Python version used was 3.10.16 distributed by conda-forge (https://anaconda.org/conda-forge/python). Model weights initialization used TorchVision v2 weights for Resnet-50 and -101 and used v1 weights for ResNet-18 and -34. Inputs used the AP, lateral, and oblique views as a 3-channel image. Images were resized with bilinear interpolation. The final classification layer was replaced with a He-initialized multilayer perceptron for binary classification. Models were trained end-to-end with no frozen layers using binary cross entropy loss with weight decay, label smoothing, and early stopping.

Three training strategies were investigated: (1) single size images, where 5 sets of ResNet models were trained on images resized to 128x128, 256x256, 512x512, 640x640, or 900x900; (2) curriculum learning, where the models were progressively trained from lower to higher image resolutions through image downsizing with bilinear interpolation to gradually increase the number of input features (i.e. trained with images with resolutions 128x128 then 256x256 then 512x512 then 640x640 and finally with 900x900); and (3) multi-scale augmentation, where models were trained using randomly resized square images between 128×128 and 900×900 to achieve partial scale invariance (Figure 1). All models were trained with 15 random hyperparameter sets with the best hyperparameters selected using the highest area under the Receiver Operating Characteristic curve (AUROC) on the validation dataset. The hyperparameter sets sampled from predefined ranges: learning rate from $10^{-4.1}$ to 10^{-3}, weight decay

from 10^{-6} to 10^{-2}, label smoothing from 0.1 to 0.3, and seed from 0 to 10,000. Batch size was optimized based on model architecture and image resolutions (Table 1).

Fracture detection performance on the held-out test dataset was reported by averaging the ResNet models' sensitivity/recall, specificity, and AUROC using 500 iterations of bootstrapping. For each bootstrapping iteration, a half-sized test subset was randomly sampled with replacement from the held-out test dataset. Inference time per image across ResNet architectures and image resolutions was calculated using a batch size of one on the workstation equipped with a NVIDIA RTX 4090. The training times was reported for the different training strategies on the HPC using a single NVIDIA V100 and compared using a Kruskal-Wallis test with post-hoc Dunn's test.

Results

We assessed the average performance of the DL models on the test images resized to the square dimensions 128x128, 256x256, 512x512, 640x640, 900x900, 1024x1024, and 2048x2048. When the performance was averaged across all those image resolutions, models trained using multi-scale augmentation achieved the highest average AUROC of 0.938 [95% CI: 0.936 - 0.939] compared to the other training strategies (Figure 2; Table 2). Among the curriculum learning models, the model set trained on images up to 512x512 had the highest average AUROC of 0.914 [0.913 - 0.915]. The models trained with the single size strategy on images of 512x512 or 640x640 had the highest average AUROC of 0.903 [0.901 - 0.905] and 0.903 [0.900 - 0.905], respectively.

The best performing models on 128x128 test images were the models trained with the single size strategy on 128x128 images with an AUROC of 0.954 [0.953 - 0.955]. On 256x256 test image, the best models were trained using single size training on 256x256 images with an AUROC of 0.970 [0.970 -

0.971], however, the curriculum learning models trained on images up to 256x256 was close with an AUROC of 0.967 [0.966–0.967], and the multi-scale augmentation models had an AUROC of 0.960 [0.959–0.960]. On 512x512 images, the highest AUROC of 0.981 [0.981–0.982] was achieved with the models using single size strategy on 640x640 images (and single size trained models on 512x512 images achieved 0.979 [0.979–0.979]). The best curriculum learning models was trained on images up to 512x512 with 0.976 [0.976–0.976], and multi-scale augmentation models was 0.978 [0.977–0.978]. On 640x640 images, the best AUROC of 0.985 [0.985–0.986] was achieved by the models using single size training on 640x640 images. The best for the other two strategies had curriculum training models on images up to 512x512 with 0.981 [0.981–0.981] and multi-scale augmentation models with 0.980 [0.980–0.980]. On 900x900 images, the best AUROC of 0.983 [0.983–0.984] was observed with single size training models on 640x640 images (and single size training models on 900x900 images had 0.982 [0.981–0.982]). Curriculum learning models on images up to 900x900 was 0.980 [0.980–0.981] and multi-scale augmentation models was 0.980 [0.979–0.980].

For all three training strategies, the highest resolution training images was 900x900. For the out-of-distribution high-resolution images, such as 1024x1024, the best average AUROC was 0.983 [0.982–0.983] with single size model training on 640x640 images. Multi-scale augmentation models achieved an AUROC of 0.979 [0.978–0.979] and the best curriculum learning models was trained on images up to 900x900 achieved 0.979 [0.979–0.979]. On 2048x2048 images, the models trained with single size training on 900x900 images had the highest AUROC of 0.959 [0.958–0.959]. Models trained with multi-scale augmentation had an AUROC of 0.893 [0.891–0.894] and curriculum learning models on images up to 900x900 had 0.931 [0.930–0.932].

On test images of resolution from 256x256 to 900x900, the sensitivity of the best models sets trained under curriculum learning, multi-scale augmentation, and single size strategies ranged between 85.4% to 90.1%, 82.8% to 87.7%, and 84.2% to 87.5%, respectively (Table 3). Their specificity was between 98.4% to 99.4%, 96.4% to 98.9%, and 98.7% to 99.3%, respectively. The sensitivity when evaluated on the out-of-distribution high-resolution 1024x1024 and 2048x2048 test images was 89.2 [89.1–89.4] and 78.2 [78.0–78.4] for curriculum learning models, 79.7 [79.4–80.1] and 49.4 [48.7–50.1] for multi-scale augmentation models, and 84.5 [84.3–84.6] and 72.8 [72.4–73.1] for single size trained models, respectively. The average of the ResNet models trained under the curriculum learning strategy consistently achieved the highest sensitivity across all image resolutions. The ResNet models trained with multi-scale augmentation had a drastic drop in sensitivity on the 2048x2048 images compared to the lower resolution images.

Generally, curriculum learning trained models performed the best on test images that matched the image resolution of the final training image resolution with the highest sensitivity continuously increasing up to 900x900 images. The best single size trained models performed well on image resolutions centered around the training image resolution and, out of all the single sized trained models, the best sensitivity was on 512x512 images. Models trained with multi-scale augmentation had the best performance on test images centered around 512x512 (Figure 2 and 3).

Curriculum learning models took significantly longer to train (11.8 [interquartile range: 11.1–16.4] hours) than multi-scale (6.7 [4.9–8.9] hours) and single size (8.0 [5.6–10.4] hours) training strategies ($P<.001$). The inference speed using the smallest CNN model, ResNet-18, was 4.59 times faster than the largest model, ResNet-101 (Figure 4). Inference speed was 1.66 times faster using 128x128 images than 1024x1024 images.

Discussion

Prior work on assessing the impact of radiograph image resolution on DL classification performance focused on chest radiographs and mammography (5–7,9–11). While DL has been studied in fracture identification on radiographs, to our knowledge, no previous studies have investigated the impact of image resolution on the performance of DL models for fracture identification on radiographs (12–15). Calcaneus fractures can be difficult to identify on radiographs with a reported sensitivity between 80-100% and specificity between 30-100% (16,17). On foot and ankle radiographs, diagnosis of fractures have a sensitivity between 66-100% and specificity between 40-100% (16–20). Our study aims to report the performance of calcaneus fracture identification on radiographs using DL models with different training strategies to overcome challenges with variable radiograph image resolutions. The ResNet models trained with multi-scale augmentation between 128x128 and 900x900 had a high sensitivity and specificity across the range of image resolutions between 256x256 and 1024x1024 with notable weakness with 128x128 and 2048x2048 images. With any particular image resolution, the ResNet models trained with curriculum learning had the highest sensitivity and specificity but required longer training times than the other training strategies and would need a different model trained up to the desired image resolution for each image resolution as the performance degrades quickly with large mismatch in the training and test image resolutions. Training using a single image size would circumvent the longer training times and would have sensitivity slightly lower than the curriculum learning trained models, even on high-resolution radiographs up to 2048x2048. It shares the same limitation with performance degradation on images resolutions out-of-distribution.

We demonstrated that the model training strategy addressing image resolution resulted in different performance on identifying calcaneus fractures. Bengio et al first described curriculum learning as a

strategy related to boosting algorithms that starts with easier examples and progressively shift to challenging examples rather than learning on a uniform distribution as is done with multi-scale augmentation (21). Similar applications of curriculum learning in radiographic fracture detection demonstrated improved performance (22,23). Curriculum learning attempts to mimic the human approach in graduated learning to optimize model training to incrementally represent complex radiographic image concepts. It can be thought as the model learning the residuals distribution with the higher resolution images.

X-ray radiographs are typically the first diagnostic imaging used in assessing musculoskeletal injuries and its diagnostic utility depends primarily on a high sensitivity for a low false negative rate. Training image resolution ranged from 128x128 to 900x900. Using the single size training strategy, the sensitivity increased as the image resolution increased up to 512x512 but trended downwards as training image resolution approached 900x900. Other DL approaches to diagnostic imaging interpretation found similar plateauing effects (5–7). The CNN architecture is designed with hierarchical feature representations with deeper layers capturing a broader receptive field (i.e. a larger subset of the original image is represented in an element of the layer's feature map) and accumulation of semantic information (24–26). A smaller anatomical region of the original image is represented at the same depth in a higher resolution image such as 2048x2048 compared to a 512x512 image resulting in potentially important anatomical landmarks in identifying fractures being omitted. At low resolutions, the radiographic features of a fracture may be indistinguishable making for poor model image interpretation. The ResNet models trained under multi-scale augmentation appears to follow a similar trend to single size trained models of maximizing sensitivity around 512x512 images. It is conceivable that this may be attributable to the models learning the optimal scale at which relevant radiographic features become sufficiently resolved for discrimination, while minimizing noise from unnecessary image detail at higher resolutions. It is also

possible that the multi-scale trained models are regressing toward the mean of the training data's uniform distribution of image resolutions. The model may be implicitly representing the central range of resolutions, leading to the observed suboptimal performance relative to the curriculum learning and single size trained models at the distributional extremes.

In contrast to single size and multi-scale training strategies, the models trained with curriculum learning achieved the best sensitivity at each image resolution out of the 3 training strategies and the performance continued to improve up to 900x900 resolution. On out-of-distribution image resolutions, the performance expectedly decreased but still outperformed the models trained with multi-scale augmentation and single size training strategies. While it takes longer to train a model with curriculum learning, the inference time was equivalent as the underlying architecture was the same. The downside would be the additional memory requirements for the multiple model weights and the inference pipeline would need to route the data towards the appropriate model instance based on the image resolution. The multi-scale augmentation trained model would be a single model solution.

Our work focused on assessing training strategies to address the range of image resolution of foot radiograph concerning for calcaneus fractures to improve DL image classification performance. Fundamentally, our work described alternative pathways to explore the model hypothesis space for the optimal weights that minimizes our binary classification loss function. Among the countless methods to optimize this non-convex optimization problem, further work in improving robustness towards image resolutions and general fracture identification performance should investigate model pre-training strategies that adapt to challenges faced in the musculoskeletal domain. Our models had its weights initialized through pre-training on the ImageNet dataset containing 14,197,122 image of a variety of image resolutions (8). This process, known as transfer learning, leverages models trained on larger

datasets to improve the performance on tasks with limited data by adapting the learned image representations (i.e. image feature extractors) of general images sourced from the internet towards our target domain of fracture identification on radiographs (27–32). Further modeling improvements may benefit from domain-adaptive pre-training where model weights are fine-tuned on a larger dataset of musculoskeletal images (which may or may not contain the foot) prior to the final training on foot radiographs (33–35). FracAtlas and MURA are large public musculoskeletal fracture datasets that present an opportunity for domain-adaptive pre-training to improve weight initialization on downstream tasks (36,37).

As the musculoskeletal radiograph datasets develop in quality and quantity, algorithmic advancements will independently improve model performance. Our results on the model training strategies of curriculum learning and multi-scale augmentation is limited to the foot radiographs collected a single hospital system for the assessment of a single binary classification task of fracture identification. Future work should explore the generalizability of these training strategies to other pathologies and broader DL tasks including multiclass classification, object detection, segmentation, and keypoint detection. Hyperparameter tuning was performed within the limits of the available computational resources. As hardware advances, there will be opportunities to investigate the effects of larger models, larger batch sizes, and other hyperparameters on a broader range of image resolutions. Additionally, training and inference metrics were reported on the hardware available and training efficiency was optimized to the best of our knowledge and time.

This study demonstrated that DL could identify calcaneus fractures on radiograph of a variety of image resolutions. Furthermore, it compared different training strategies on the performance on in- and out-of-distribution image resolutions and examined the differences in required computational resources. Model

training strategies, such as curriculum learning and multi-scale augmentation, presents an opportunity to algorithmically improve model robustness towards different image resolutions without requiring additional annotated data.

Acknowledgments (Anonymized, if any)
Thank you to the team at the Foot & Ankle Research and Innovation Lab at Massachusetts General Hospital (Harvard Medical School), Orthopaedic Biomechanics Laboratory in the Holland Bone and Joint Program at the Sunnybrook Research Institute (University of Toronto), Institute of Biomedical Engineering (University of Toronto), and the Division of Orthopaedic Surgery (University of Toronto) for their support on this project. This work was possible through the scholarship support from the Canadian Institutes of Health Research (Canada Graduate Research Scholarship – Master's program and Michael Smith Foreign Study Supplements), the Royal College of Physicians and Surgeons of Canada/University of Toronto Surgeon Scientist Training Program (Clinician Investigator Program), and the Ontario Ministry of Health and Long-Term Care.

References

1. Fu T, Viswanathan V, Attia A, et al. Assessing the Potential of a Deep Learning Tool to Improve Fracture Detection by Radiologists and Emergency Physicians on Extremity Radiographs. Academic Radiology. 2024;31(5):1989–1999. doi: 10.1016/j.acra.2023.10.042.

2. Thian YL, Li Y, Jagmohan P, Sia D, Chan VEY, Tan RT. Convolutional Neural Networks for Automated Fracture Detection and Localization on Wrist Radiographs. Radiol Artif Intell. 2019;1(1):e180001. doi: 10.1148/ryai.2019180001.

3. Yu Q, Liu Y, Li H, et al. Multi-task learning for calcaneus fracture diagnosis of X-ray images. Biomedical Signal Processing and Control. 2025;99:106843. doi: 10.1016/j.bspc.2024.106843.

4. Huda W, Abrahams RB. X-Ray-Based Medical Imaging and Resolution. American Journal of Roentgenology. 2015;204(4):W393–W397. doi: 10.2214/AJR.14.13126.

5. Haque MIU, Dubey AK, Danciu I, Justice AC, Ovchinnikova OS, Hinkle JD. Effect of image resolution on automated classification of chest X-rays. J Med Imaging (Bellingham). 2023;10(4):044503. doi: 10.1117/1.JMI.10.4.044503.

6. Rajaraman S, Yang F, Zamzmi G, Xue Z, Antani S. Does image resolution impact chest X-ray based fine-grained Tuberculosis-consistent lesion segmentation? arXiv; 2023. doi: 10.48550/arXiv.2301.04032.

7. Sabottke CF, Spieler BM. The Effect of Image Resolution on Deep Learning in Radiography. Radiology: Artificial Intelligence. Radiological Society of North America; 2020;2(1):e190015. doi: 10.1148/ryai.2019190015.

8. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. ImageNet: A Large-Scale Hierarchical Image Database. .

9. Budi Nugroho, Anny Yuniarti. The Effect of Image Size on the Deep Learning Approach for Classifying Covid-19 Pneumonia. Technium. 2023;16:164–168. doi: 10.47577/technium.v16i.9976.

10. GM H, Mori K, Verma S, Athanasiou L. The Influence of Image Cropping Sizes on Mammographic Breast Cancer Classification Using CNN. 2023 IEEE 23rd International Conference on Bioinformatics and Bioengineering (BIBE). 2023. p. 226–229. doi: 10.1109/BIBE60311.2023.00043.

11. Wollek A, Hyska S, Sabel B, Ingrisch M, Lasser T. Higher Chest X-ray Resolution Improves Classification Performance. arXiv; 2023. doi: 10.48550/arXiv.2306.06051.

12. Xie Y, Li X, Chen F, et al. Artificial intelligence diagnostic model for multi-site fracture X-ray images of extremities based on deep convolutional neural networks. Quant Imaging Med Surg. 2024;14(2):1930–1943. doi: 10.21037/qims-23-878.

13. Ashkani-Esfahani S, Mojahed Yazdi R, Bhimani R, et al. Detection of ankle fractures using deep learning algorithms. Foot and Ankle Surgery. 2022;28(8):1259–1265. doi: 10.1016/j.fas.2022.05.005.

14. Bae J, Yu S, Oh J, et al. External Validation of Deep Learning Algorithm for Detecting and Visualizing Femoral Neck Fracture Including Displaced and Non-displaced Fracture on Plain X-ray. J Digit Imaging. 2021;34(5):1099–1109. doi: 10.1007/s10278-021-00499-2.

15. Blüthgen C, Becker AS, Vittoria De Martini I, Meier A, Martini K, Frauenfelder T. Detection and localization of distal radius fractures: Deep learning system versus radiologists. European Journal of Radiology. 2020;126:108925. doi: 10.1016/j.ejrad.2020.108925.

16. Utukuri MM, Knowles D, Smith KL, Barrie JL, Gavan D. The value of the axial view in assessing calcaneal fractures. Injury. 2000;31(5):325–326. doi: 10.1016/S0020-1383(99)00303-4.

17. Vannier MW, Hildebolt CF, Gilula LA, et al. Calcaneal and pelvic fractures: Diagnostic evaluation by three-dimensional computed tomography scans. J Digit Imaging. 1991;4(3):143–152. doi: 10.1007/BF03168159.

18. Ozturk P, Aksay E, Oray NC, Bayram B, Basci O, Tokgoz D. Emergency physician accuracy using ultrasonography to diagnose lateral malleolar fracture. The American Journal of Emergency Medicine. 2018;36(3):362–365. doi: 10.1016/j.ajem.2017.08.020.

19. York T, Franklin C, Reynolds K, et al. Reporting errors in plain radiographs for lower limb trauma—a systematic review and meta-analysis. Skeletal Radiol. 2022;51(1):171–182. doi: 10.1007/s00256-021-03821-9.

20. Singh AK, Malpass TS, Walker G. Ultrasonic assessment of injuries to the lateral complex of the ankle. Emergency Medicine Journal. 1990;7(2):90–94. doi: 10.1136/emj.7.2.90.

21. Bengio Y, Louradour J, Collobert R, Weston J. Curriculum learning. Proceedings of the 26th Annual International Conference on Machine Learning. Montreal Quebec Canada: ACM; 2009. p. 41–48. doi: 10.1145/1553374.1553380.

22. Luo J, Kitamura G, Doganay E, Arefan D, Wu S. Medical Knowledge-Guided Deep Curriculum Learning for Elbow Fracture Diagnosis from X-Ray Images. Medical Imaging 2021: Computer-Aided Diagnosis. 2021. p. 35. doi: 10.1117/12.2582184.

23. Jiménez-Sánchez A, Mateus D, Kirchhoff S, et al. Medical-based Deep Curriculum Learning for Improved Fracture Classification. 2019. p. 694–702. doi: 10.1007/978-3-030-32226-7_77.

24. O'Shea K, Nash R. An Introduction to Convolutional Neural Networks. arXiv; 2015. doi: 10.48550/arXiv.1511.08458.

25. Dharmaretnam D, Fyshe A. The Emergence of Semantics in Neural Network Representations of Visual Information. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). New Orleans, Louisiana: Association for Computational Linguistics; 2018. p. 776–780. doi: 10.18653/v1/N18-2122.

26. Luo W, Li Y, Urtasun R, Zemel R. Understanding the Effective Receptive Field in Deep Convolutional Neural Networks. arXiv; 2017. doi: 10.48550/arXiv.1701.04128.

27. Huh M, Agrawal P, Efros AA. What makes ImageNet good for transfer learning? arXiv; 2016. doi: 10.48550/arXiv.1608.08614.

28. Rahman R, Yagi N, Hayashi K, Maruo A, Muratsu H, Kobashi S. Enhancing fracture diagnosis in pelvic X-rays by deep convolutional neural network with synthesized images from 3D-CT. Sci Rep. Nature Publishing Group; 2024;14(1):8004. doi: 10.1038/s41598-024-58810-4.

29. Tahir A, Saadia A, Khan K, Gul A, Qahmash A, Akram RN. Enhancing diagnosis: ensemble deep-learning model for fracture detection using X-ray images. Clinical Radiology. 2024;79(11):e1394–e1402. doi: 10.1016/j.crad.2024.08.006.

30. Yoon AP, Lee Y-L, Kane RL, Kuo C-F, Lin C, Chung KC. Development and Validation of a Deep Learning Model Using Convolutional Neural Networks to Identify Scaphoid Fractures in Radiographs. JAMA Netw Open. 2021;4(5):e216096. doi: 10.1001/jamanetworkopen.2021.6096.

31. Thian YL, Li Y, Jagmohan P, Sia D, Chan VEY, Tan RT. Convolutional Neural Networks for Automated Fracture Detection and Localization on Wrist Radiographs. Radiology: Artificial Intelligence. 2019;1(1):e180001. doi: 10.1148/ryai.2019180001.

32. Tanzi L, Vezzetti E, Moreno R, Moos S. X-Ray Bone Fracture Classification Using Deep Learning: A Baseline for Designing a Reliable Approach. Applied Sciences. Multidisciplinary Digital Publishing Institute; 2020;10(4):1507. doi: 10.3390/app10041507.

33. Guan H, Liu M. Domain Adaptation for Medical Image Analysis: A Survey. IEEE Trans Biomed Eng. 2022;69(3):1173–1185. doi: 10.1109/TBME.2021.3117407.

34. Mei X, Liu Z, Robson PM, et al. RadImageNet: An Open Radiologic Deep Learning Research Dataset for Effective Transfer Learning. Radiology: Artificial Intelligence. 2022;4(5):e210315. doi: 10.1148/ryai.210315.

35. Kim D, Wang K, Sclaroff S, Saenko K. A Broad Study of Pre-training for Domain Generalization and Adaptation. In: Avidan S, Brostow G, Cissé M, Farinella GM, Hassner T, editors. Computer Vision – ECCV 2022. Cham: Springer Nature Switzerland; 2022. p. 621–638. doi: 10.1007/978-3-031-19827-4_36.

36. Rajpurkar P, Irvin J, Bagul A, et al. MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs. arXiv; 2018. doi: 10.48550/arXiv.1712.06957.

37. Abedeen I, Rahman MA, Prottyasha FZ, Ahmed T, Chowdhury TM, Shatabda S. FracAtlas: A Dataset for Fracture Classification, Localization and Segmentation of Musculoskeletal Radiographs. Sci Data. 2023;10(1):521. doi: 10.1038/s41597-023-02432-4.

Tables (embedded, one per page)

Table 1: Training batch sizes for each model architecture and input image resolutions.

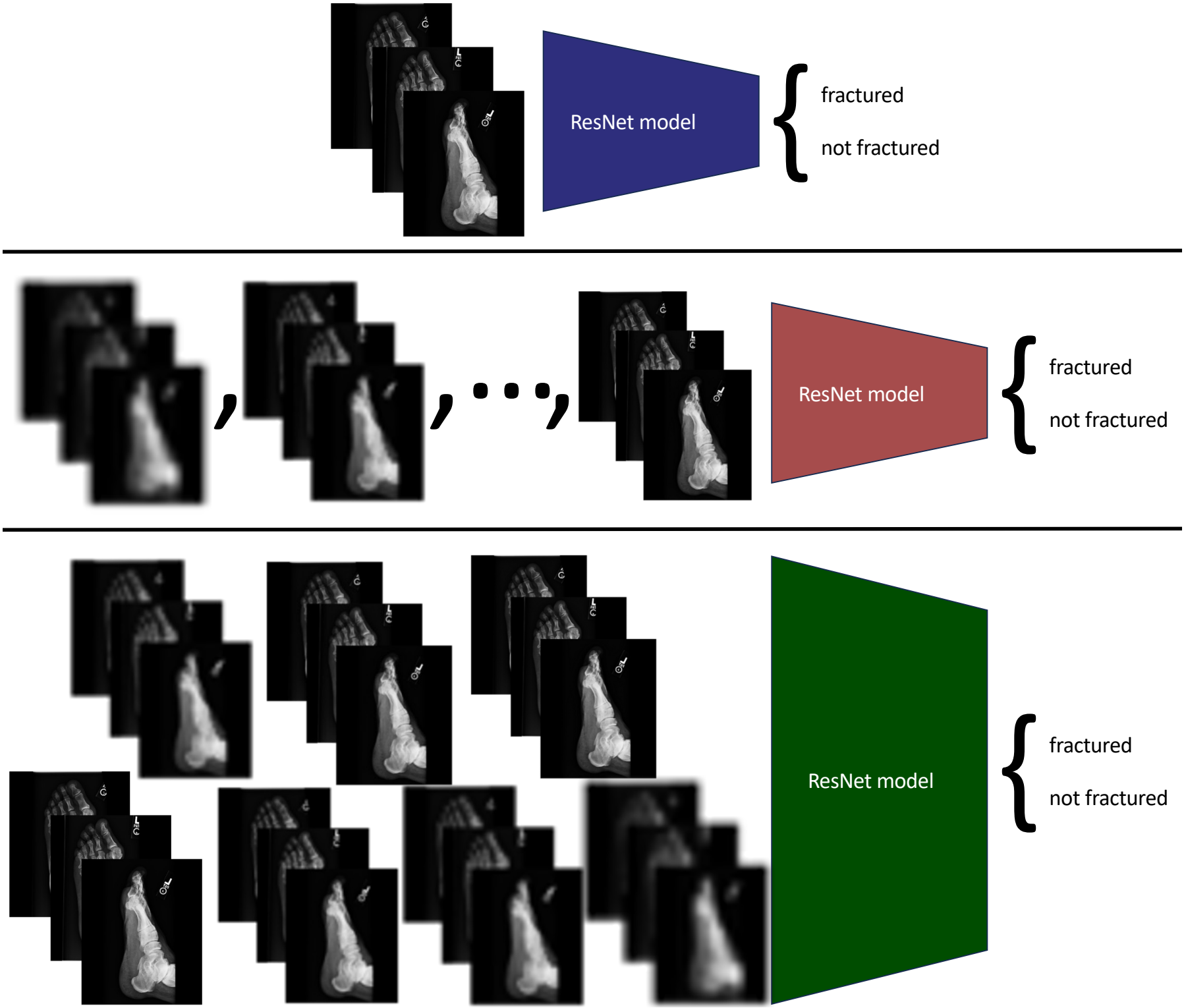
Model	Batch size				
	128x128	256x256	512x512	640x640	900x900
ResNet-18	256	128	64	64	16
ResNet-34	256	128	64	64	16
ResNet-50	256	128	64	32	16
ResNet-101	256	128	32	32	16

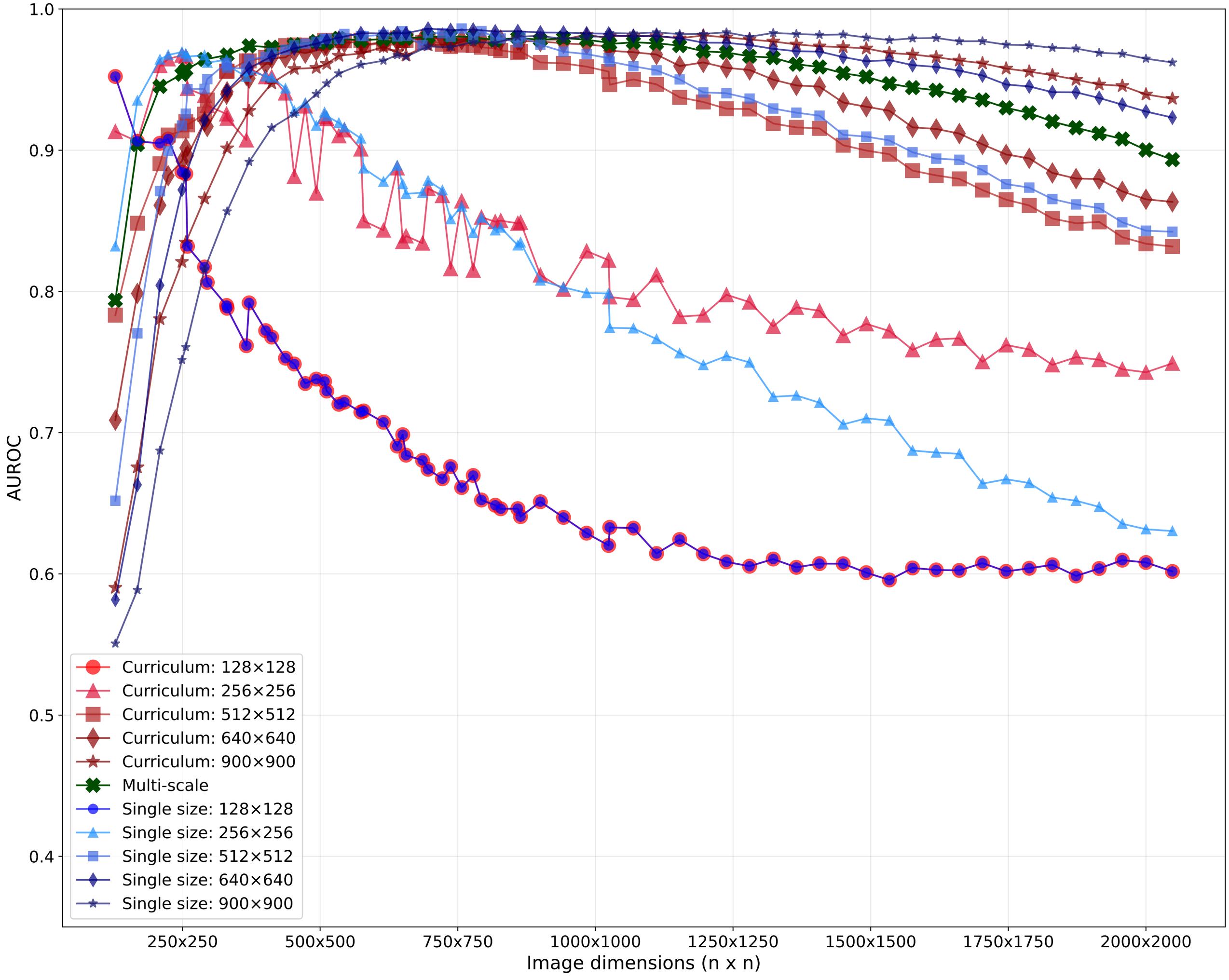
Table 2: AUROC of the training strategies trained on different image resolutions. Reports the average AUROC of the ResNet-18, -34, -50, and -101 models on a held-out test dataset. Low resolution includes square image resolutions between 128x128 to 900x900. High resolution includes image resolutions 1024x1024 and 2048x2048. Bolded values indicate best performance.

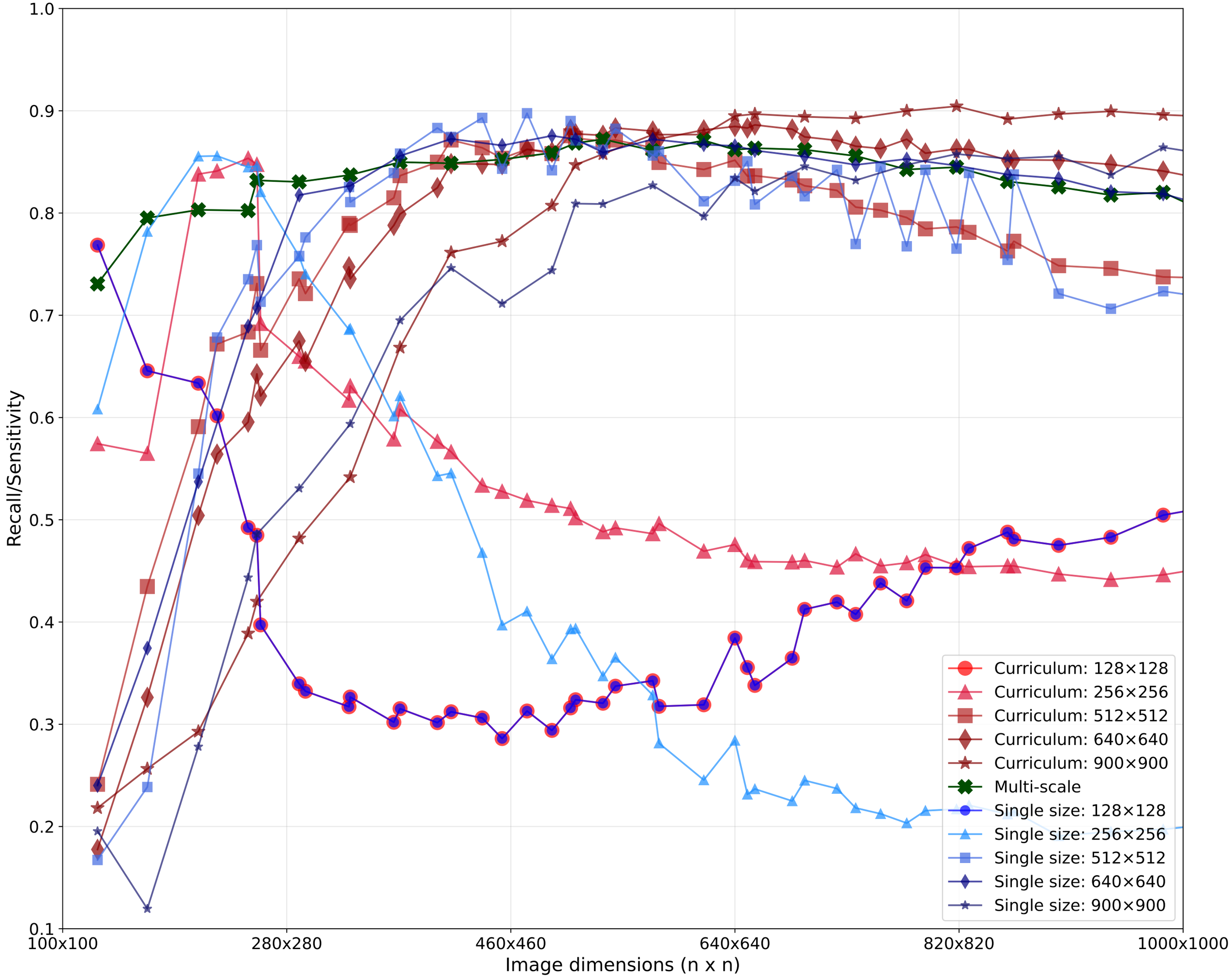
Test image dimensions:		128x128	256x256	512x512	640x640	900x900	1024x1024	2048x2048	Low resolution	High resolution
Training strategy	**Training image dimension**	**AUROC [95% CI]**	**AUROC [95% CI]**	**AUROC [95% CI]**	**AUROC [95% CI]**	**AUROC [95% CI]**	**AUROC [95% CI]**	**AUROC [95% CI]**	**Average AUROC [95% CI]**	**Average AUROC [95% CI]**
Curriculum learning	128x128	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Curriculum learning	Up to 256x256	0.907 [0.906–0.908]	0.967 [0.966–0.967]	0.926 [0.925–0.927]	0.887 [0.885–0.888]	0.808 [0.804–0.812]	0.826 [0.824–0.829]	0.745 [0.742–0.748]	0.899 [0.898–0.900]	0.786 [0.783–0.788]
Curriculum learning	Up to 512x512	0.775 [0.773–0.778]	0.919 [0.918–0.921]	0.976 [0.976–0.976]	0.981 [0.981–0.981]	0.963 [0.963–0.964]	0.955 [0.955–0.956]	0.828 [0.827–0.829]	0.923 [0.921–0.925]	0.892 [0.890–0.894]
Curriculum learning	Up to 640x640	0.713 [0.710–0.716]	0.914 [0.912–0.915]	0.975 [0.975–0.975]	0.976 [0.975–0.976]	0.976 [0.976–0.977]	0.972 [0.971–0.972]	0.861 [0.859–0.862]	0.911 [0.909–0.913]	0.916 [0.914–0.918]
Curriculum learning	Up to 900x900	0.589 [0.585–0.592]	0.834 [0.831–0.836]	0.960 [0.960–0.961]	0.972 [0.971–0.972]	0.980 [0.980–0.981]	0.979 [0.979–0.979]	0.931 [0.930–0.932]	0.867 [0.864–0.870]	0.955 [0.954–0.956]
Multi-scale	N/A	0.798 [0.795–0.800]	0.960 [0.959–0.960]	0.978 [0.977–0.978]	0.980 [0.980–0.980]	0.980 [0.979–0.980]	0.979 [0.978–0.979]	0.893 [0.891–0.894]	**0.939 [0.937–0.940]**	0.936 [0.934–0.937]
Single size	128x128	**0.954 [0.953–0.955]**	0.887 [0.886–0.889]	0.733 [0.729–0.737]	0.696 [0.694–0.699]	0.671 [0.670–0.673]	0.619 [0.617–0.621]	0.604 [0.599–0.609]	0.788 [0.786–0.791]	0.612 [0.609–0.614]
Single size	256x256	0.829 [0.825–0.833]	**0.970 [0.970–0.971]**	0.925 [0.924–0.926]	0.896 [0.895–0.898]	0.801 [0.797–0.806]	0.801 [0.798–0.805]	0.631 [0.626–0.635]	0.884 [0.883–0.886]	0.716 [0.712–0.720]
Single size	512x512	0.664 [0.661–0.667]	0.920 [0.919–0.922]	0.979 [0.979–0.979]	0.978 [0.978–0.979]	0.974 [0.973–0.974]	0.967 [0.966–0.967]	0.837 [0.834–0.839]	0.903 [0.901–0.906]	0.902 [0.899–0.904]
Single size	640x640	0.580 [0.576–0.584]	0.886 [0.885–0.888]	**0.981 [0.981–0.982]**	**0.985 [0.985–0.986]**	**0.983 [0.983–0.984]**	**0.983 [0.982–0.983]**	0.922 [0.921–0.923]	0.883 [0.880–0.886]	0.952 [0.951–0.953]
Single size	900x900	0.555 [0.553–0.557]	0.755 [0.752–0.758]	0.947 [0.946–0.947]	0.967 [0.966–0.967]	0.982 [0.981–0.982]	0.980 [0.979–0.980]	**0.959 [0.958–0.959]**	0.841 [0.838–0.844]	**0.969 [0.969–0.970]**

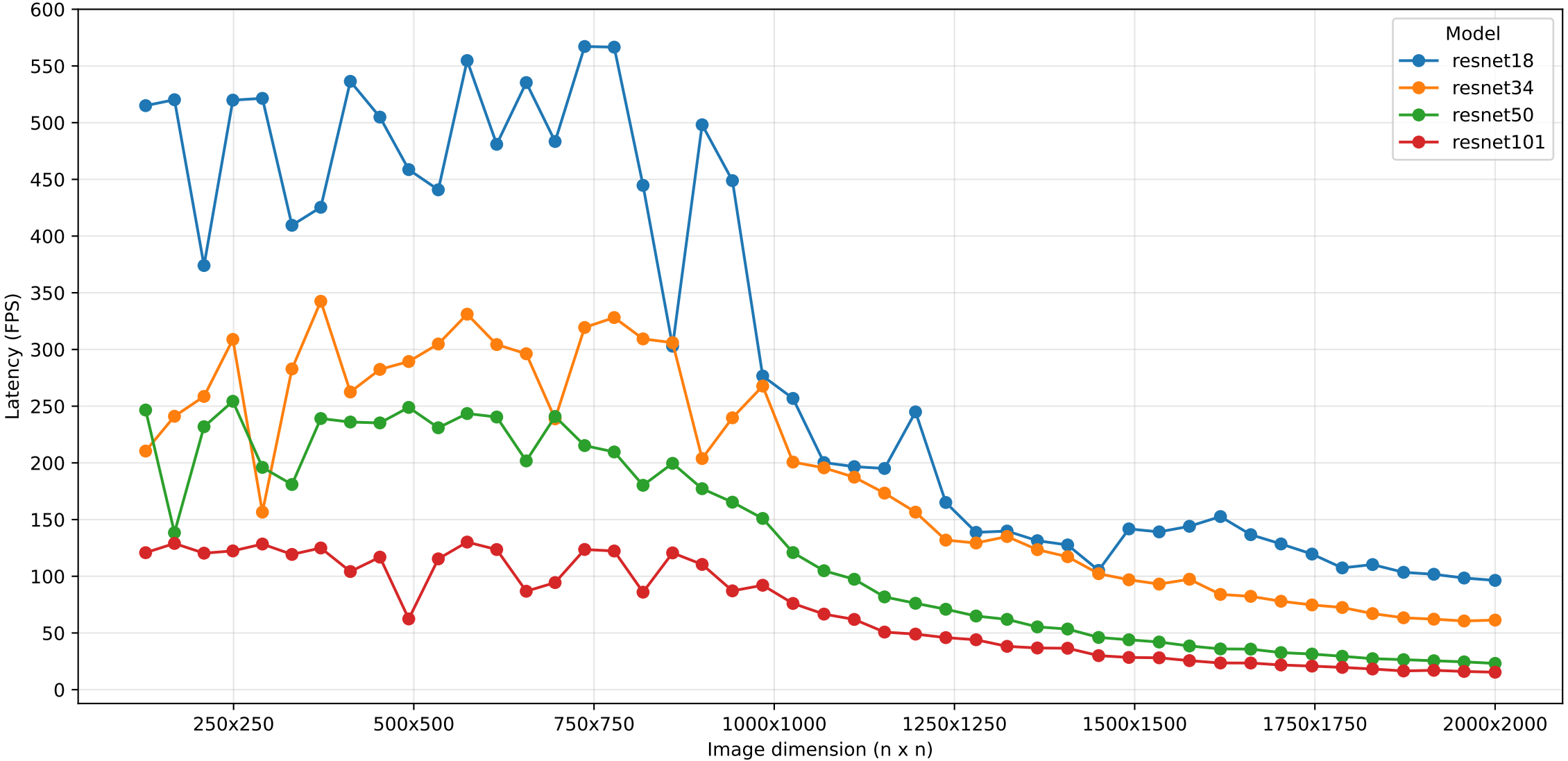
Table 3: Sensitivity and specificity with their 95% confidence interval on a held-out test dataset resized to different image resolutions. The bolded numbers indicates the best performance for that image dimension. The highest sensitivity was consistently achieved by ResNet models trained using curriculum learning, even in out-of-distribution image dimensions (i.e. 1024x1024 and 2048x2048). The best performance among the curriculum learning and single size trained models was reported. The same multi-scale trained models was used for all test image dimensions.

Test image dimension:	128x128			256x256			512x512			640x640			900x900			1024x1024			2048x2048		
	Best model	Sensitivity (%)	Specificity (%)	Best model	Sensitivity (%)	Specificity (%)	Best model	Sensitivity (%)	Specificity (%)	Best model	Sensitivity (%)	Specificity (%)	Best model	Sensitivity (%)	Specificity (%)	Best model	Sensitivity (%)	Specificity (%)	Best model	Sensitivity (%)	Specificity (%)
Curriculum learning	N/A	N/A	N/A	Up to 256x256	**85.4 [85.2–85.5]**	98.4 [98.4–98.4]	Up to 512x512	**87.7 [87.6–87.8]**	**99.4 [99.4–99.4]**	Up to 640x640	**87.7 [87.6–87.8]**	98.9 [98.9–98.9]	Up to 900x900	**90.1 [89.9–90.2]**	99.1 [99.0–99.1]	Up to 900x900	**89.2 [89.1–89.4]**	99.1 [99.1–99.2]	Up to 900x900	**78.2 [78.0–78.4]**	95.5 [95.5–95.6]
Multi-scale	N/A	73.2 [72.6–73.8]	68.2 [67.1–69.3]	N/A	82.8 [82.6–83.0]	96.4 [96.3–96.4]	N/A	**87.7 [87.6–87.8]**	98.9 [98.9–98.9]	N/A	**87.6 [87.5–87.8]**	98.9 [98.8–98.9]	N/A	82.3 [82.0–82.6]	98.9 [98.9–99.0]	N/A	79.7 [79.4–80.1]	98.7 [98.7–98.8]	N/A	49.4 [48.7–50.1]	98.3 [98.3–98.4]
Single size	128x128	**77.1 [77.0–77.3]**	**98.0 [98.0–98.0]**	256x256	84.2 [84.1–84.3]	**98.7 [98.7–98.8]**	512x512	87.5 [87.4–87.6]	99.0 [99.0–99.0]	640x640	87.4 [87.3–87.5]	**99.0 [99.0–99.0]**	900x900	85.2 [85.0–85.3]	**99.3 [99.3–99.3]**	900x900	84.5 [84.3–84.6]	**99.3 [99.3–99.4]**	900x900	72.8 [72.4–73.1]	**99.2 [99.1–99.2]**


ResNet model
fractured
not fractured
ResNet model
fractured
not fractured
ResNet model
fractured
not fractured


1.0
0.9
0.8
0.7
0.6
0.5
0.4
AUROC
250x250
500x500
750x750
1000x1000
1250x1250
1500x1500
1750x1750
2000x2000
Image dimensions (n x n)
Curriculum: 128×128
Curriculum: 256×256
Curriculum: 512×512
Curriculum: 640×640
Curriculum: 900×900
Multi-scale
Single size: 128×128
Single size: 256×256
Single size: 512×512
Single size: 640×640
Single size: 900×900


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Recall/Sensitivity
100x100
280x280
460x460
640x640
820x820
1000x1000
Image dimensions (n x n)
Curriculum: 128×128
Curriculum: 256×256
Curriculum: 512×512
Curriculum: 640×640
Curriculum: 900×900
Multi-scale
Single size: 128×128
Single size: 256×256
Single size: 512×512
Single size: 640×640
Single size: 900×900


Model
resnet18
resnet34
resnet50
resnet101
Latency (FPS)
0
50
100
150
200
250
300
350
400
450
500
550
600
250x250
500x500
750x750
1000x1000
1250x1250
1500x1500
1750x1750
2000x2000
Image dimension (n x n)