Large-Scale Radiograph Pre-training: Reducing Label Dependency in Medical Imaging

Niklas Bühler, Paul Hager

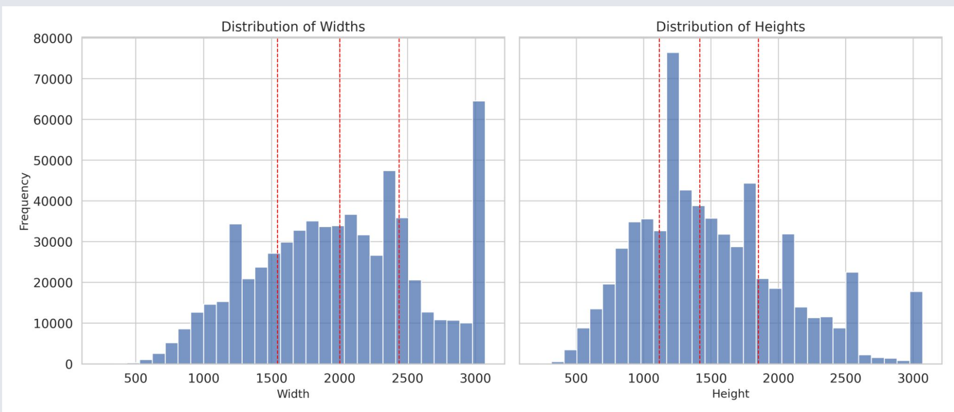
Background and Motivation

- **Large-scale medical datasets** are **underutilized** due to **costly manual labeling**
- **Self-Supervised Learning** (SSL) enables **label-efficient exploitation**
- **Masked Autoencoders** (MAEs) can **capture fine-grained details** essential for medical tasks
- Challenges: **High radiograph resolutions** and **extreme variability in resolutions**

We perform **scalable SSL pre-training** tailored to medical imaging, **reducing dependency on labeled data** and **optimizing training efficiency**.

Method

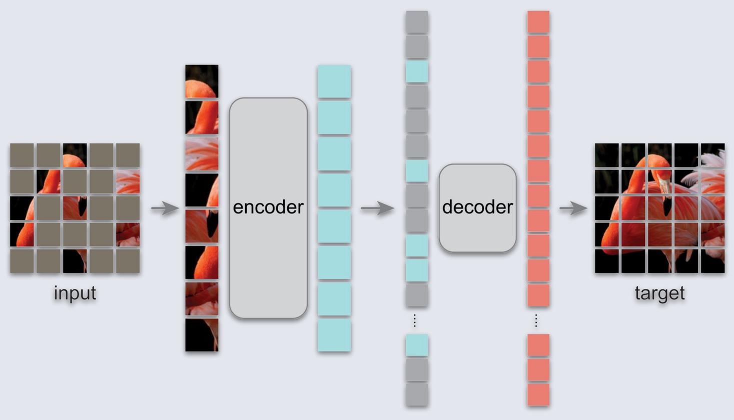
Real-world Clinical Dataset



- **639,877 radiographs** across 14 anatomical regions
- Extracted from MRI PACS
- Extreme variability in resolutions, with **391,013 unique sizes up to 3,072×3,072 pixels**

Models

- Pre-training: **ViT MAE Base** [2]
- Baselines: **ViT-Base** [1]
- Bilinear interpolation of positional encoding for variable resolutions



Adapted from [2].

Pre-training

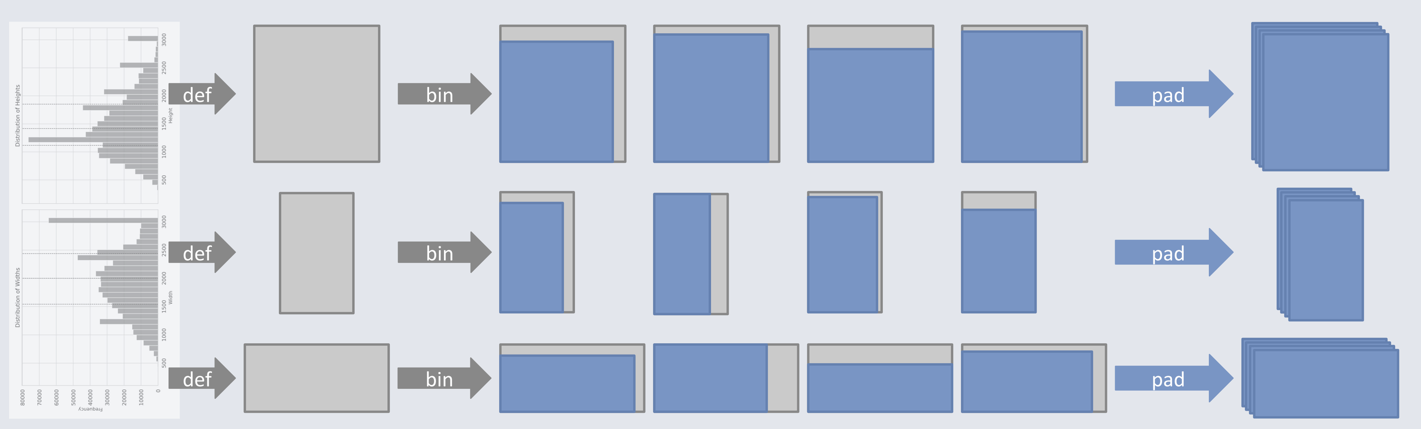
- **Masked 75% of input patches** and trained on self-supervised **masked image reconstruction task**

Fine-tuning

- Three **clinical downstream tasks** of varying difficulty: **Anatomical region classification** (ARC), **foreign material detection** (FMD), and **fracture detection** (FRAC)
- All with **minimal labels**
- Baseline initializations: Random and ImageNet-21k pre-trained
- Stratified by patients to prevent data leakage

Dynamic Batch Binning

- **Dynamic Batch Binning (DBB)**, inspired by VariViT batching [3]
- **Standard approaches** scaling, cropping, and padding are **suboptimal**
- **Clusters** images with **similar resolutions into bins**
- **Minimizes padding** to reduce computational overhead
- **Efficiently handles extreme variability in resolutions**



Define bins w.r.t. resolution distribution | Sort images into next larger bin | Pad images to bin resolution

Downstream Task Performance

	Labels	Supervised ViT-B	ImageNet-21k Pre-trained ViT-B	Radiograph Pre-trained ViT MAE B
ARC	1,000	56.89%	**92.21%**	73.57%
FMD	652	46.88%	**92.88%**	57.99%
FRAC	652	50.00%	56.86%	**58.46%**

Metrics are balanced top-1 accuracy.

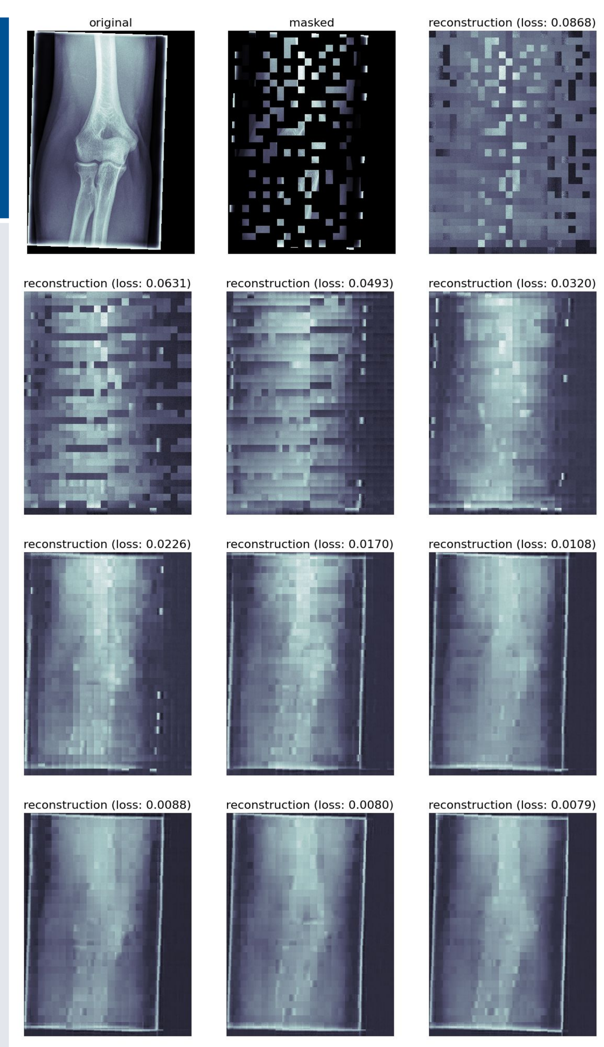
Results

- Final reconstruction loss of ViT MAE: 0.0523 MSE
- **Pre-training improved performance** across all downstream tasks with minimal labels
- **ImageNet-21k** pre-training excelled in **general vision tasks**, like ARC and FMD
- **Radiograph pre-training** improved performance on **complex medical tasks**, like FRAC

- **DBB** achieved **82% reduction in total compute** compared to fixed-size processing

	Total Input Tokens	Processed Tokens	Total Compute
Padded to patch size	$9.22 \cdot 10^8$	100%	100%
Fixed image size	$2.62 \cdot 10^9$	284%	807%
DBB	$1.10 \cdot 10^9$	119%	142%

Potential of Increasing Pre-training Volume



ImageNet-21k Pre-training

- 90 epochs on **14 million images**, resulting in **1.26b training iterations**
- Focus on **general-purpose visual features**

Radiograph Pre-training

- 10 epochs on **600,000 radiographs**, resulting in **6m training iterations**
- Constrained by data storage and compute
- Focus on **domain-specific medical details**
- **5% of samples** and **0.5% of iterations**

Conclusion

- **Dynamic Batch Binning reduced computational cost by 82%** compared to fixed-size processing, allowing us to efficiently train on **600,000+ radiographs** with **highly variable resolutions** up to **3,072×3,072 pixels**
- **Pre-trained MAEs outperformed random initialization** in all downstream tasks
- **Radiograph-specific pre-training surpassed ImageNet-21k pre-training** in fracture detection, demonstrating the value of **domain-specific embeddings**

References

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." In: International Conference on Learning Representations. ICLR. 2021.
[2] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. "Masked Autoencoders are Scalable Vision Learners." In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE. 2022, pp. 16000–16009.
[3] A. Varma, S. Shit, C. Prabhakar, D. Scholz, H. B. Li, D. Rueckert, B. Wiestler, et al. "VariViT: A Vision Transformer for Variable Image Sizes." In: Medical Imaging with Deep Learning. 2024.

Klinikum rechts der Isar
Technische Universität München