The Convergence of HPC and AI for Healthcare on Intel® Based Supercomputers

Dr. Valeriu Codreanu
Group Leader, High-Performance Machine Learning Group
SURFsara

SURFsara & Intel IPCC* Team

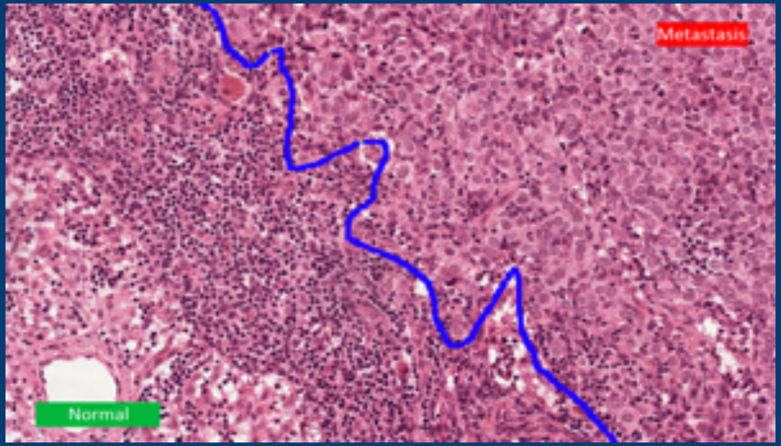
Valeriu Codreanu, PhD
Senior HPC/AI Consultant
IPCC PI, SURFsara

Damian Podareanu, MSc
HPC/AI Consultant
SURFsara

Vikram Saletore, PhD
Principal Engineer
IPCC Co-PI &
AI Products Group, Intel

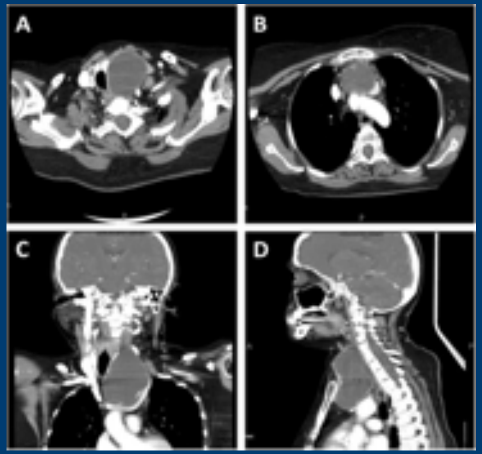
- SURFsara is a Dutch organization focused on computing, data management and storage at a national scale
- Support Dutch researchers in taking an early competitive advantage of upcoming technologies
- SURF Open Innovation Lab focuses on emerging technologies

*IPCC - Intel Parallel Computing Center



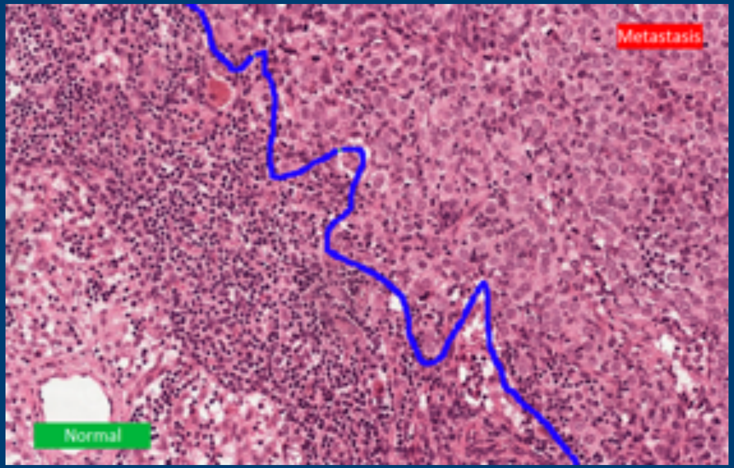
Radboud University Medical Center

Dell EMC



Netherlands Cancer Institute (NKI)

COLLABORATIONS UNDER IPCC



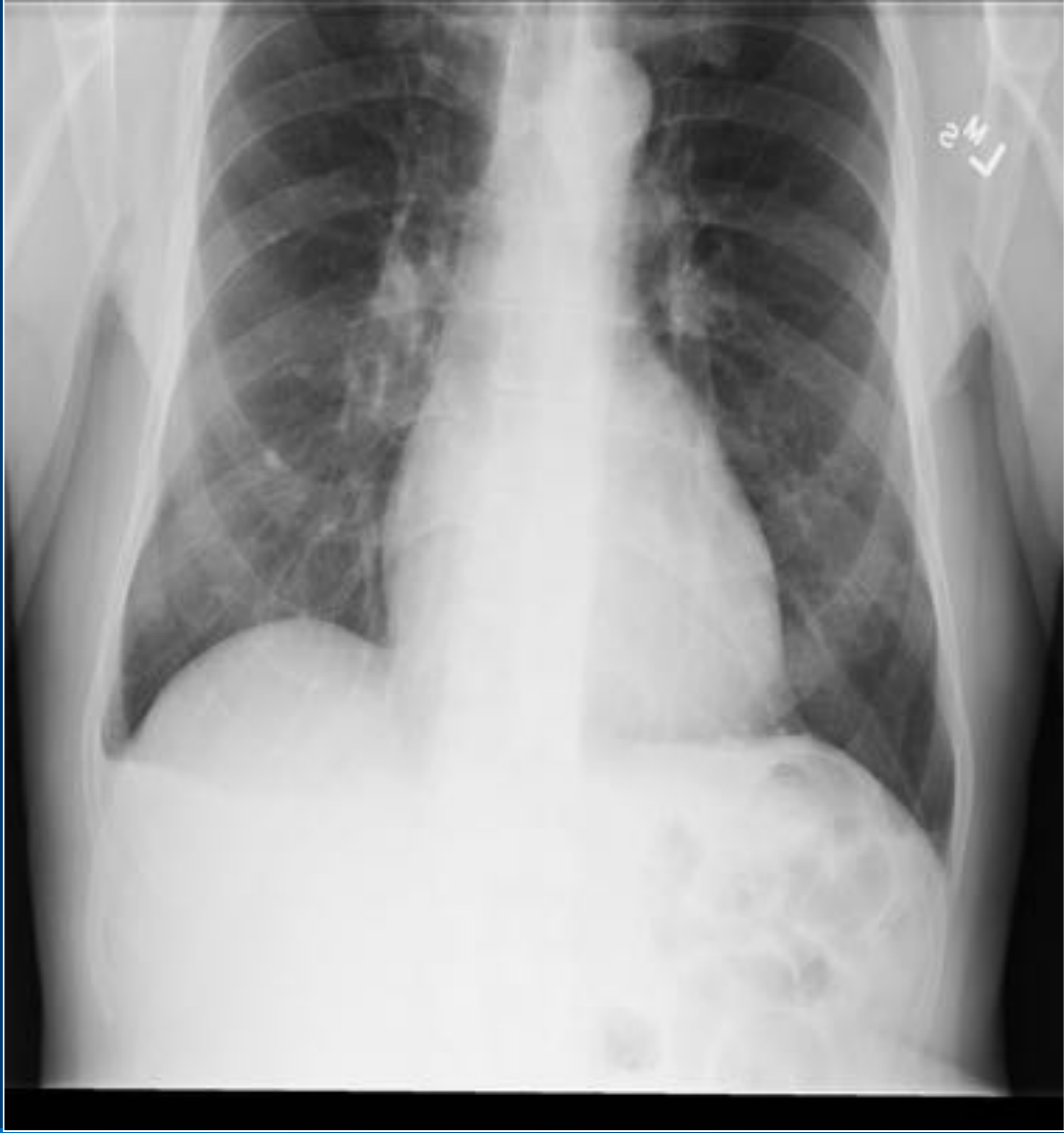
DELL EMC SURF SARA intel

Dell-SURFsara-Intel CheXNet

CheXNet (from Stanford) is a model for identifying thoracic pathologies from the NIH ChestXray14 dataset

- DenseNet121 topology
  - Pretrained on ImageNet
- Dataset contains 112K images
  - Multicategory / Multilabel
  - Unbalanced

https://stanfordmlgroup.github.io/projects/chexnet/



Transfer Learning Using Highly Accurate ResNet-50 for Real Use Case

Fine-tuned ResNet-50 that was pre-trained on ImageNet using the Zenith cluster.
Transfer learning using the TPU repository

To increase accuracy:

- When picking a pre-trained checkpoint do not pick the last one.
- Start with the learning rate at which the model was training when it was checkpointed.
- Perform gradual warmup of the learning rate, proportionally to the global batch size.

To compute efficiently:

- Pack data in TF Records and consume them efficiently (asynchronously with compute).
    - Preferably in Tensorflow

Comparative timings for 128-node transfer learning run

Global batch size	Framework	# nodes	Time/epoch
4096	Keras	128	**85 s**
4096	Tensorflow	128	**18 s**

Can we do better? Large-input ResNet

Extended ResNet-50 to ResNet-59

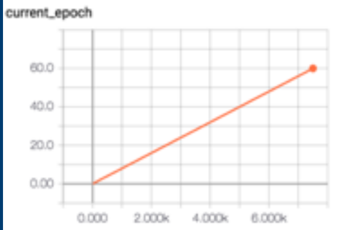
- Increased input size to 1024x1024
  - stride-2 convolution in the second residual block
- Higher memory and compute requirements:
  - ~4x heavier model
  - ~43GB memory use for a batch size of 64
- Trained on ImageNet for only 60 epochs:
  - Global batch size: 10240
  - top1/top5 accuracy of 78%/94%
  - Around 8h time-to-train on 256 nodes!

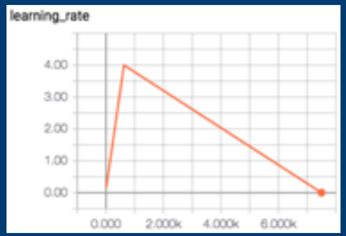
Layer	Output shape
Input	3x224x224
Conv 1, 7x7, stride 2	64x112x112
Max Pool, 3x3	64x56x56
ResBlock 2a	256x56x56
ResBlock 2b	256x56x56
ResBlock 2c	256x56x56
ResBlock 3a	512x28x28
ResBlock 3b	512x28x28
ResBlock 3c	512x28x28
ResBlock 3d	512x28x28
ResBlock 4a	1024x14x14
ResBlock 4b	1024x14x14
ResBlock 4c	1024x14x14
ResBlock 4d	1024x14x14
ResBlock 4e	1024x14x14
ResBlock 4f	1024x14x14
ResBlock 5a	2048x7x7
ResBlock 5b	2048x7x7
ResBlock 5c	2048x7x7
Average pool	2048x1x1
Flatten	2048
Output	14

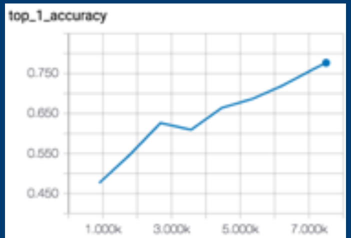
ResNet-50 architecture

Layer	Output shape
Input	3x896x896
Conv 1, 7x7, stride 2	64x448x448
Max Pool, 3x3	64x224x224
ResBlock 2a	256x112x112
ResBlock 2b	256x112x112
ResBlock 2c	256x112x112
ResBlock 3a	512x56x56
ResBlock 3b	512x56x56
ResBlock 3c	512x56x56
ResBlock 4a	1024x28x28
ResBlock 4b	1024x28x28
ResBlock 4c	1024x28x28
ResBlock 4d	1024x28x28
ResBlock 5a	2048x14x14
ResBlock 5b	2048x14x14
ResBlock 5c	2048x14x14
ResBlock 5d	2048x14x14
ResBlock 5e	2048x14x14
ResBlock 5f	2048x14x14
ResBlock 6a	4096x7x7
ResBlock 6b	4096x7x7
ResBlock 6c	4096x7x7
Average Pool	4096x1
Flatten	4096
Output	14

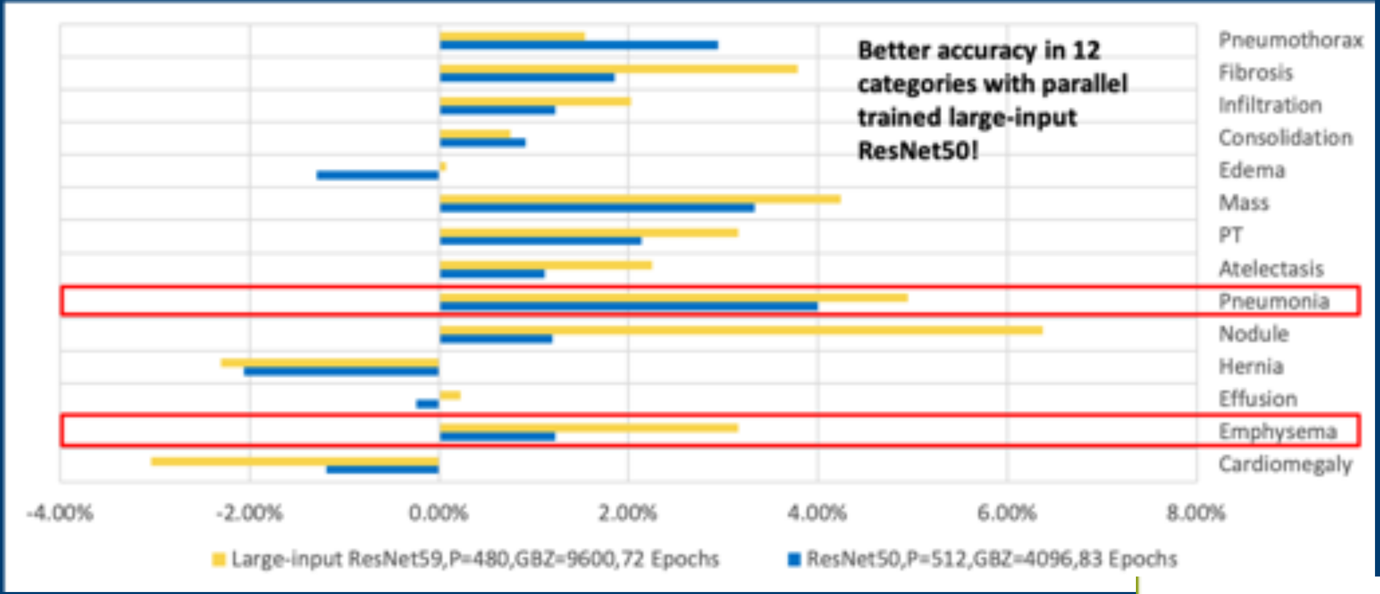
ResNet-59 architecture
(large input ResNet-50)







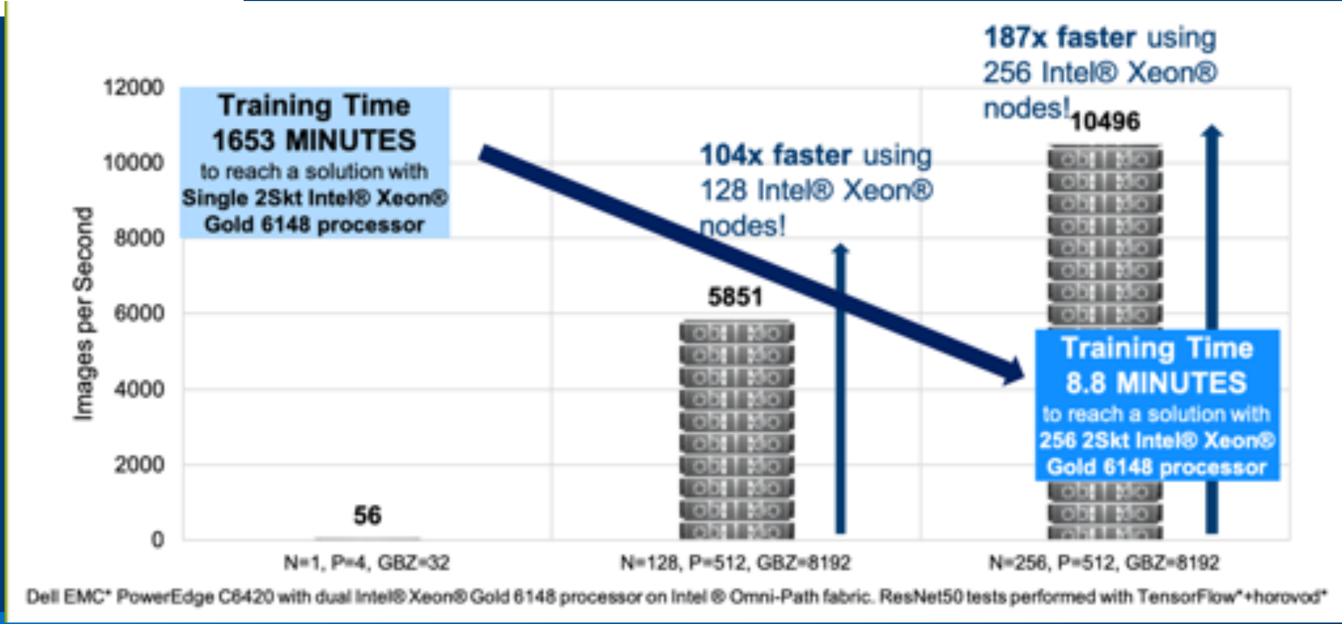
ResNet-50 and beyond



- ResNet-50 provides great scalability and improved results
- ResNet-59, with 896x896 input improves accuracy further
- It takes 50 minutes to train ResNet-59 on ChestXRay-14 using 240 nodes – for higher accuracy

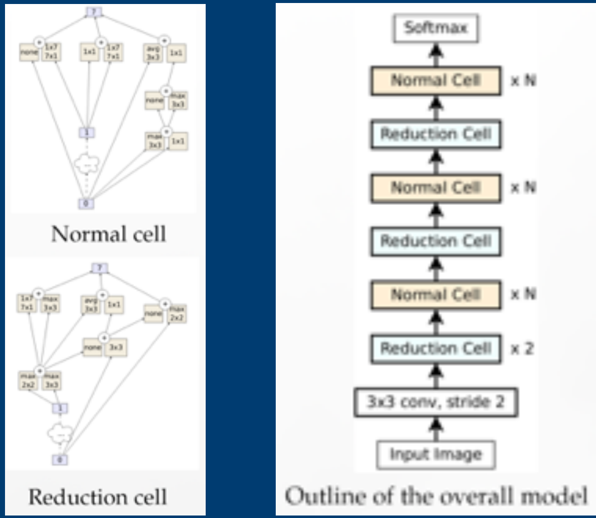
Challenge encountered:

- Large memory usage for large-input models

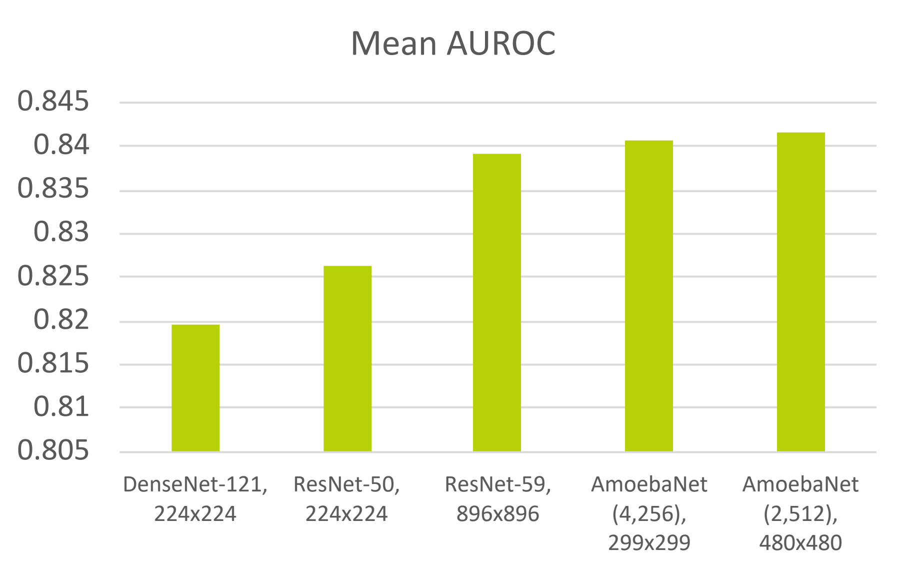


Exploring scalable, accurate AI radiology models

High computational and memory requirements!



AmoebaNet-D architecture



	ResNet50	ResNet59	AmoebaNet (4,256)	AmoebaNet (2,512)
No. parameters	22M	90M	80M	168M
Input size	224x224	896x896	299x299	480x480
Training throughput [img/s/node]	90	22.5	13.3	3.5
Memory consumption/ batch of 64 images [GB]	11	43	61	**125**
Approximate training time on 256 nodes [minutes]	8.8	36	53	198
Top-1 accuracy on ImageNet-1K	76.2%	78.1%	79.9%	80.9%

See also our research poster:
(RP22) Supercomputer-Scale Training of Large AI Radiology Models

Ongoing Investigations

Intel & SURFsara IPCC Research

Valeriu Codreanu
Senior HPC/AI Consultant
IPCC PI, SURFsara

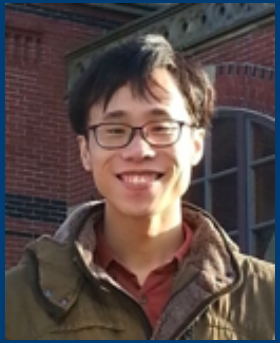
Damian Podareanu, MSc
HPC/AI Consultant
IPCC co-PI, SURFsara

Vikram Saletore
Principal Engineer &
Performance Architect
IPCC Co-PI, AI Products
Group, Intel

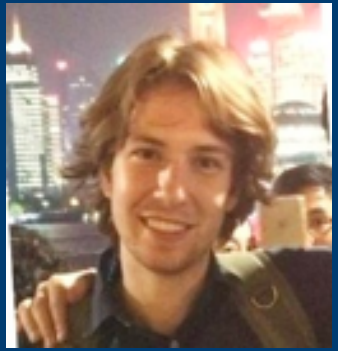
Kushal Datta
Machine Learning Engineer
AI Products Group, Intel

Caspar van Leeuwen
HPC/AI Consultant,
SURFsara

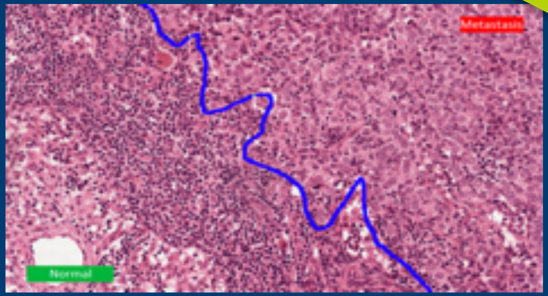
Ruben Hekster
AI Consultant,
SURFsara



Maxwell Cai
AI Consultant,
SURFsara



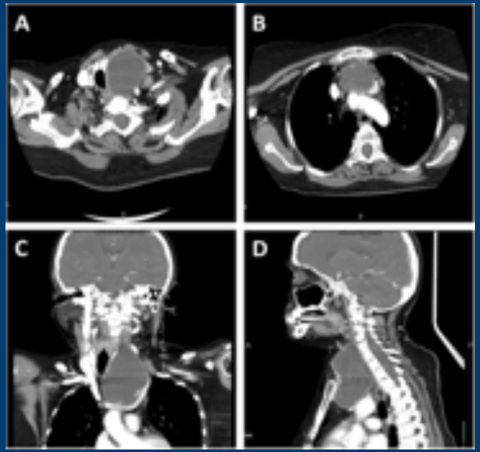
David Ruhe
AI Consultant,
SURFsara



Radboudumc
university medical center

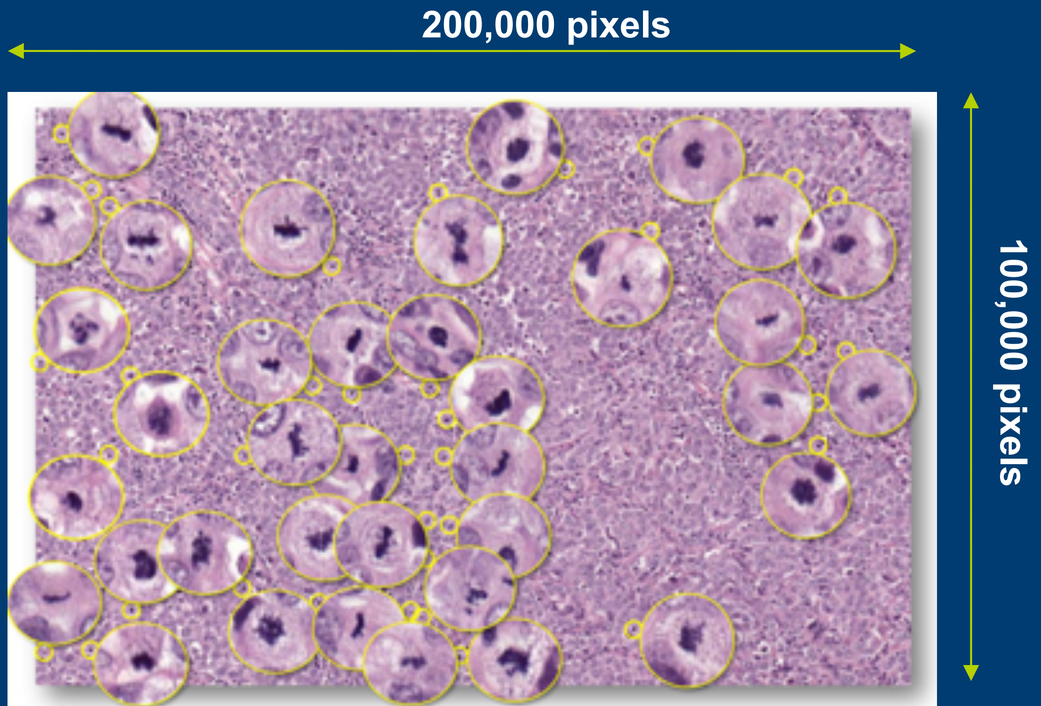
SURF SARA

intel

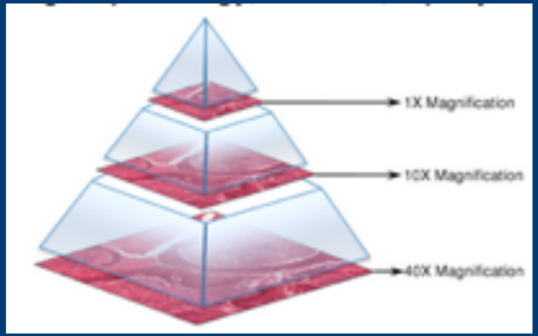


Histopathology image analysis

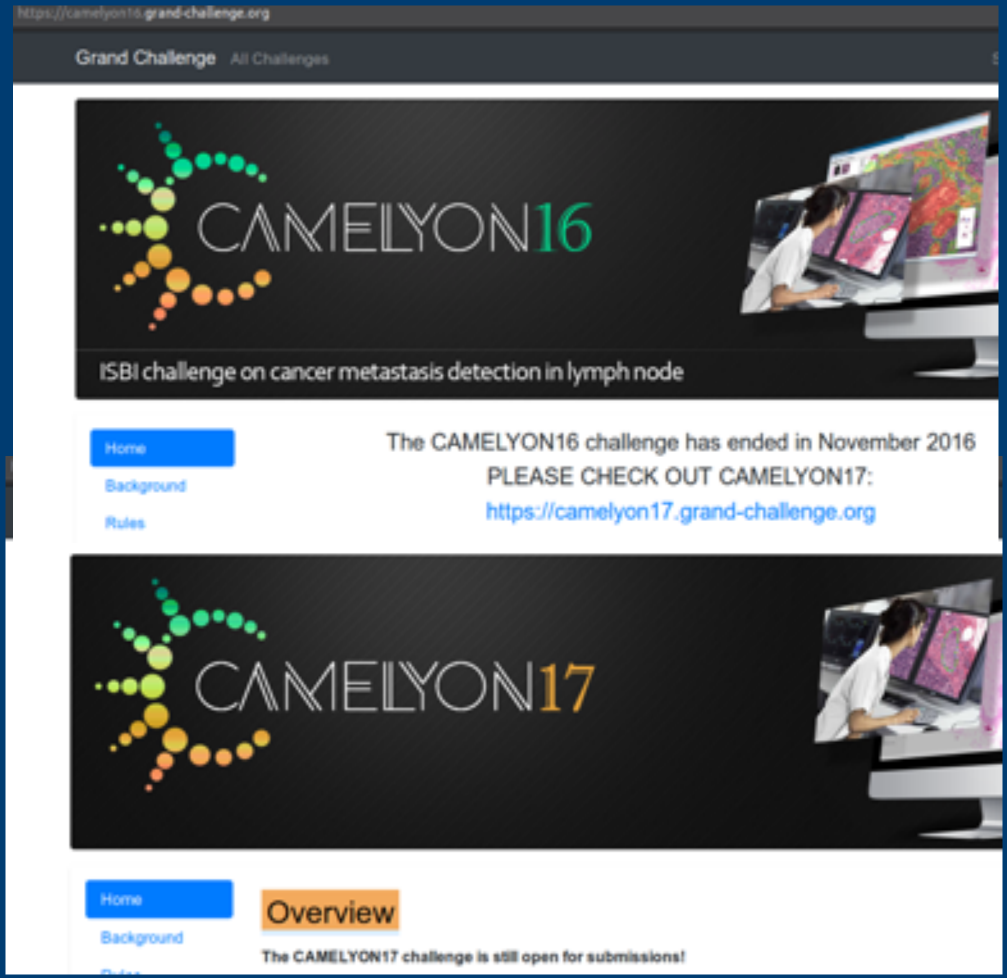
- A medium-sized digital pathology lab processes ~ 30.000 cases / year
- M.D.s need tools to optimize their daily workflow
- AI adoption is slow and few hospitals have such tools in production
- Samples are very large: **55.88GB** in 3 byte RGB pixel format
- Patches can be used, but for certain afflictions, absolute location is important
- Special acknowledgment: EXAMODE



Tellez-Martin D, et al. Whole-slide mitosis detection in H&E breast histology using pHH3 as a reference to train distilled stain-invariant convolutional networks. IEEE-Trans Med Im 2018.



Camelyon challenge



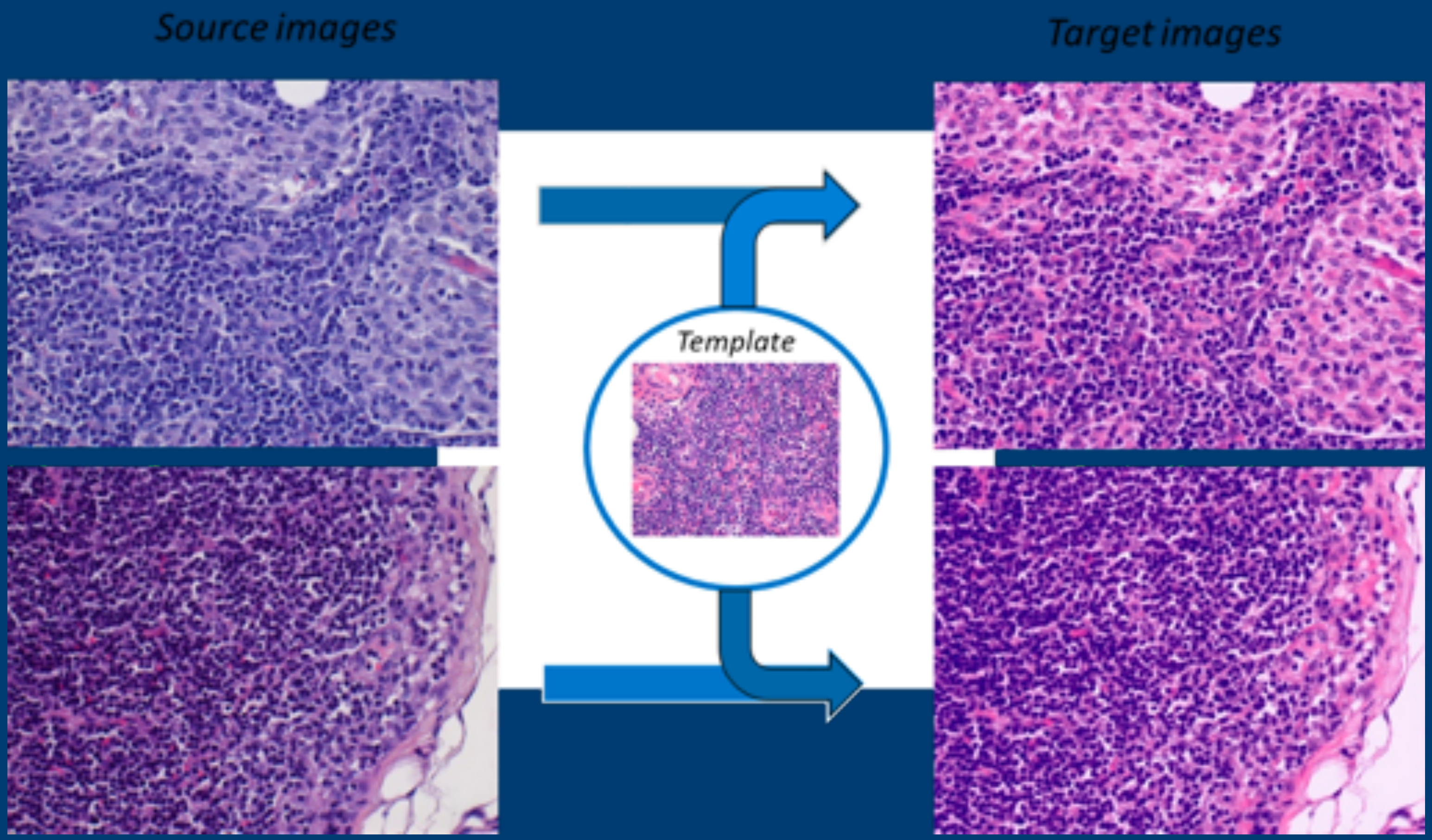
- […] automated detection and classification of breast cancer metastases in whole-slide images of histological lymph node sections.

- This task has **high clinical relevance** and would normally require extensive microscopic assessment by pathologists. […]

- Semantic segmentation task

- 400 WSI from:
  - Radboud University Medical Center
  - University Medical Center Utrecht

WSI Stain normalization

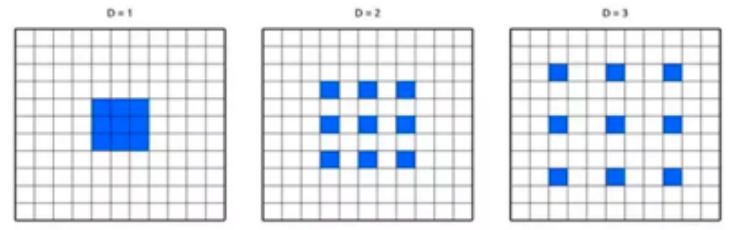


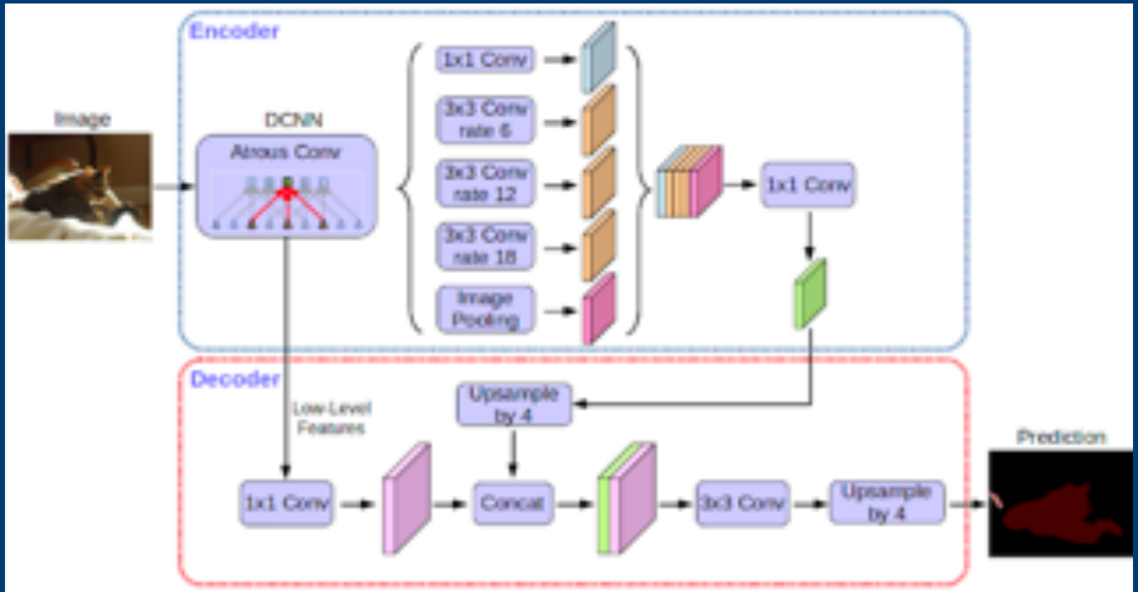
Zanjani, F. G., Zinger, S., Bejnordi, B. E., & van der Laak, J. A. (2018). Histopathology Stain-Color Normalization Using Deep Generative Models.

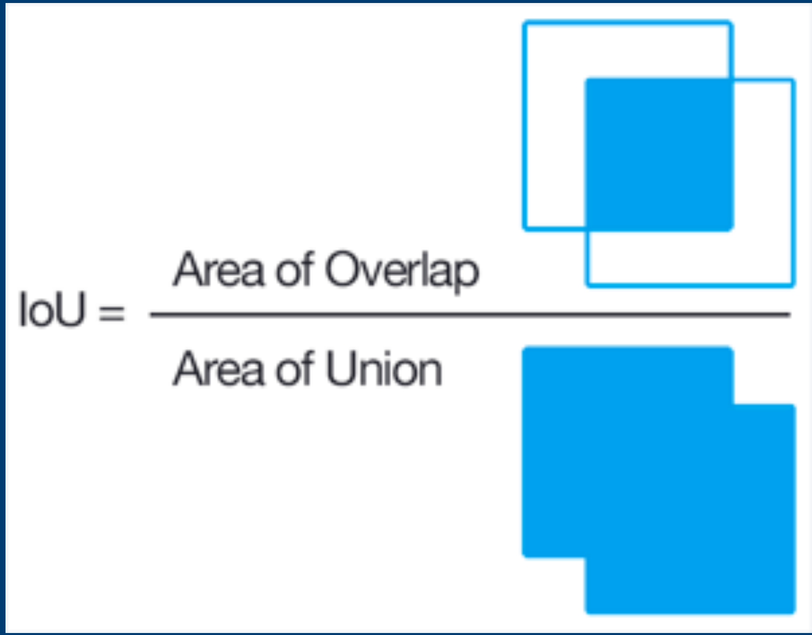
Atrous convolutions and mIoU

Main idea: Improve segmentation by learning spatial orientation of Regions of Interest in WSIs

Deeplabv3+ pretrained backbone of xception65 trained on ImageNet and COCO for the semantic segmentation task







Patch size dependent accuracy

Patch size	mIoU	Batch size	Memory consumption
704x704	0.78	8	11 GB
1024x1024	0.83	8	22 GB

- These results don't have stain normalization
- Batch size hinders performance and convergence
- Growing to larger patches seems very difficult without vast memory

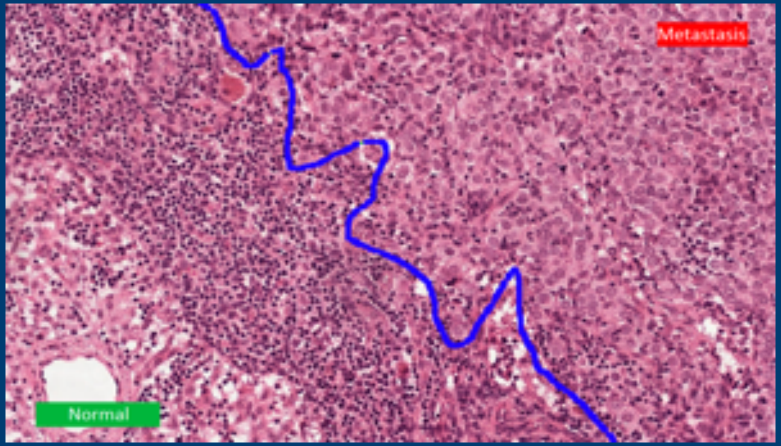
- For both patch sizes weak scaling up to 8 nodes (workers) :
  - ~98% efficiency
  - Lead to a better mIoU score (+0.04) for the larger patch size. We believe this is due to the small batch size used for 1 worker.

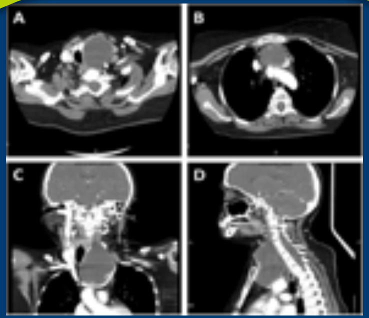
Future steps

- Analyze further large patch impact on accuracy
- Scale-out experiments to large batch sizes

Preliminary conclusion

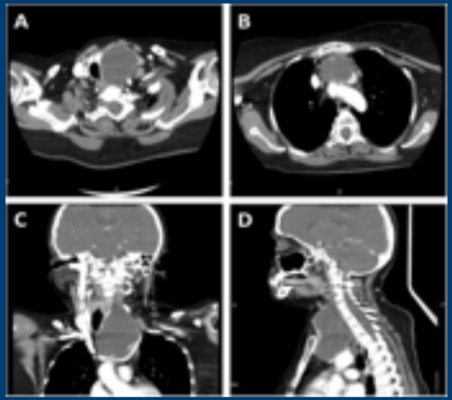
- Memory utilization can be traded to accuracy for the semantic segmentation task on the histology data from the Camelyon challenge





NETHERLANDS CANCER INSTITUTE ANTONI VAN LEEUWENHOEK
SURF SARA
intel

Netherlands Cancer Institute (NKI) use-case
Head & Neck CT synthesis with large-scale GANs



- Goal: Generate realistic 3D CT scans of large dimensionality
  - > 512^3 voxels (**500MB-1GB per example**)
- Can help medical centers:
  - Improve their discriminative models
  - Remove some privacy concerns – Easier data sharing
  - Predict treatment plan on radiotherapy – Conditional on patient

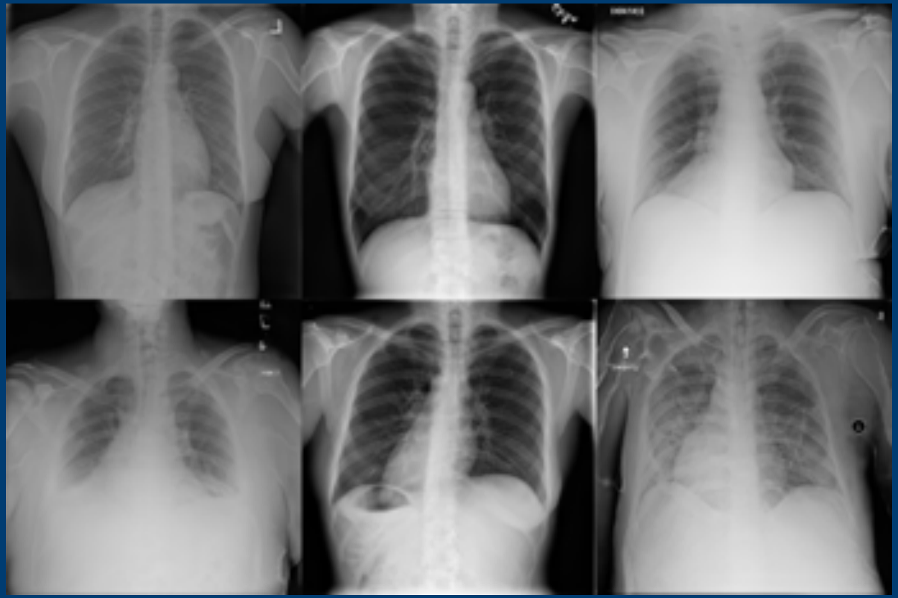
Training configuration	CELEBA Sliced Wasserstein distance $\times 10^3$ 128	64	32	16	Avg	MS-SSIM
(a) Gulrajani et al. (2017)	12.99	7.79	7.62	8.73	9.28	0.2854
(b) + Progressive growing	4.62	**2.64**	3.78	6.06	4.28	**0.2838**
(c) + Small minibatch	75.42	41.33	41.62	26.57	46.23	0.4065
(d) + Revised training parameters	9.20	6.53	4.71	11.84	8.07	0.3027
(e*) + Minibatch discrimination	10.76	6.28	6.04	16.29	9.84	0.3057
(e) Minibatch stddev	13.94	5.67	2.82	5.71	7.04	0.2950
(f) + Equalized learning rate	4.42	3.28	2.32	7.52	4.39	0.2902
(g) + Pixelwise normalization	**4.06**	3.04	**2.02**	**5.13**	**3.56**	0.2845
(h) Converged	2.42	2.17	2.24	4.99	2.96	0.2828

Maintaining a relatively large batch size tends to be very important for GANs (e.g. BigGAN, ProGAN):

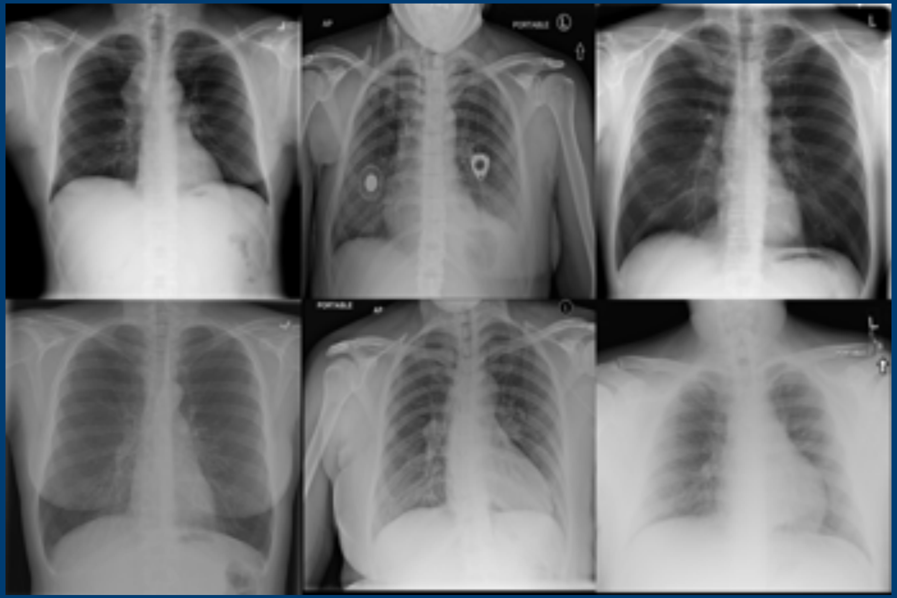
- Large memory systems will certainly help here
- Working towards 1st 3D large-resolution generation
- ProGAN 2D model – 46M parameters

Medical image synthesis: 2D large image generation

Which ones are fakes?

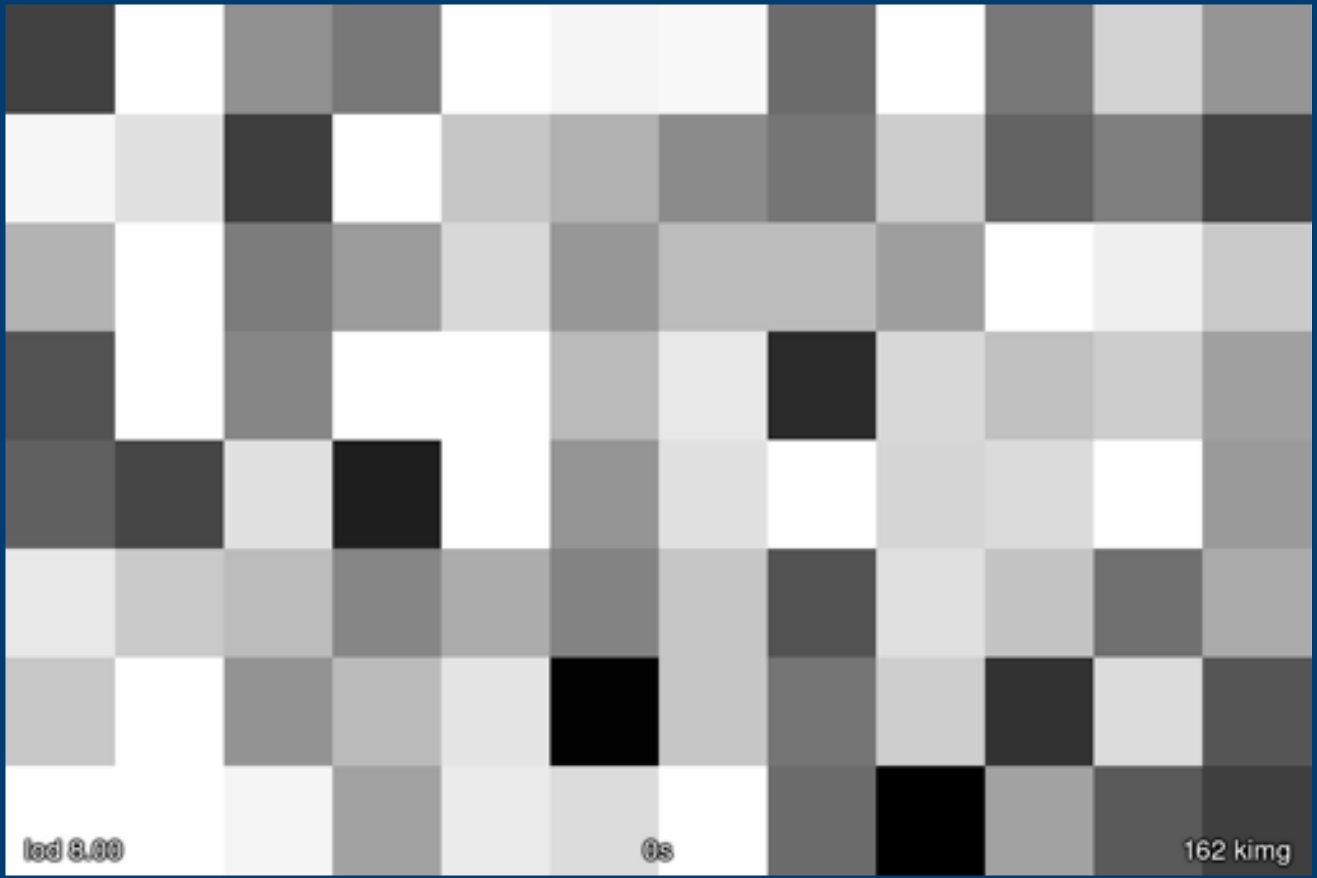


Fake Chest XRays



Real Chest XRays

Medical image synthesis: 2D large image generation using Progressive growing of GANs

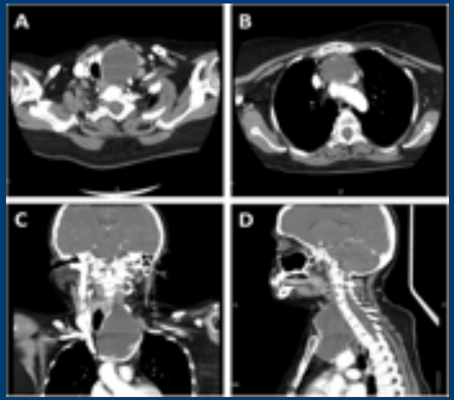


StyleGAN/ProGAN – memory/compute requirements

Table 1: Training requirements

Resolution	2D Batch size	2D Memory usage (GB)	3D Batch size	3D Memory usage (GB)
8^2 / 8^3	32	2.7	64	2.7
16^2 / 16^3	32	3.1	32	4.4
32^2 / 32^3	32	4.3	16	8.6
64^2 / 64^3	32	6.4	8	16.8
128^2 / 128^3	16	6.7	4	33.4
256^2 / 256^3	8	6.6	2	66.2
512^2 / 512^3	8	9.7	1	131.2
1024^2 / 1024^3	8	17.5	-	-
2048^2 / 2048^3	4	23.1	-	-
4096^2 / 4096^3	4	31.5	-	-
8192^2 / 8192^3	2	109.6	-	-

Batch size, memory consumption and throughput at different resolutions for the 2D and 3D progressive GAN. All results are using a single compute node.



Conclusions

- Medical imaging typically deal with large-scale data (either 2D or 3D)
  - This puts large pressure on memory subsystems
- Models become more accurate if enlarged
  - Our large-scale parallel-trained AI radiologist achieves superior performance
- Our histopathology models are more accurate if we use larger patches

- Intel® Optane Data Center Persistent Memory can enable large-scale models (using large-scale data):
    - 2D histopathology models
    - 3D generative models

Thank you! Questions?