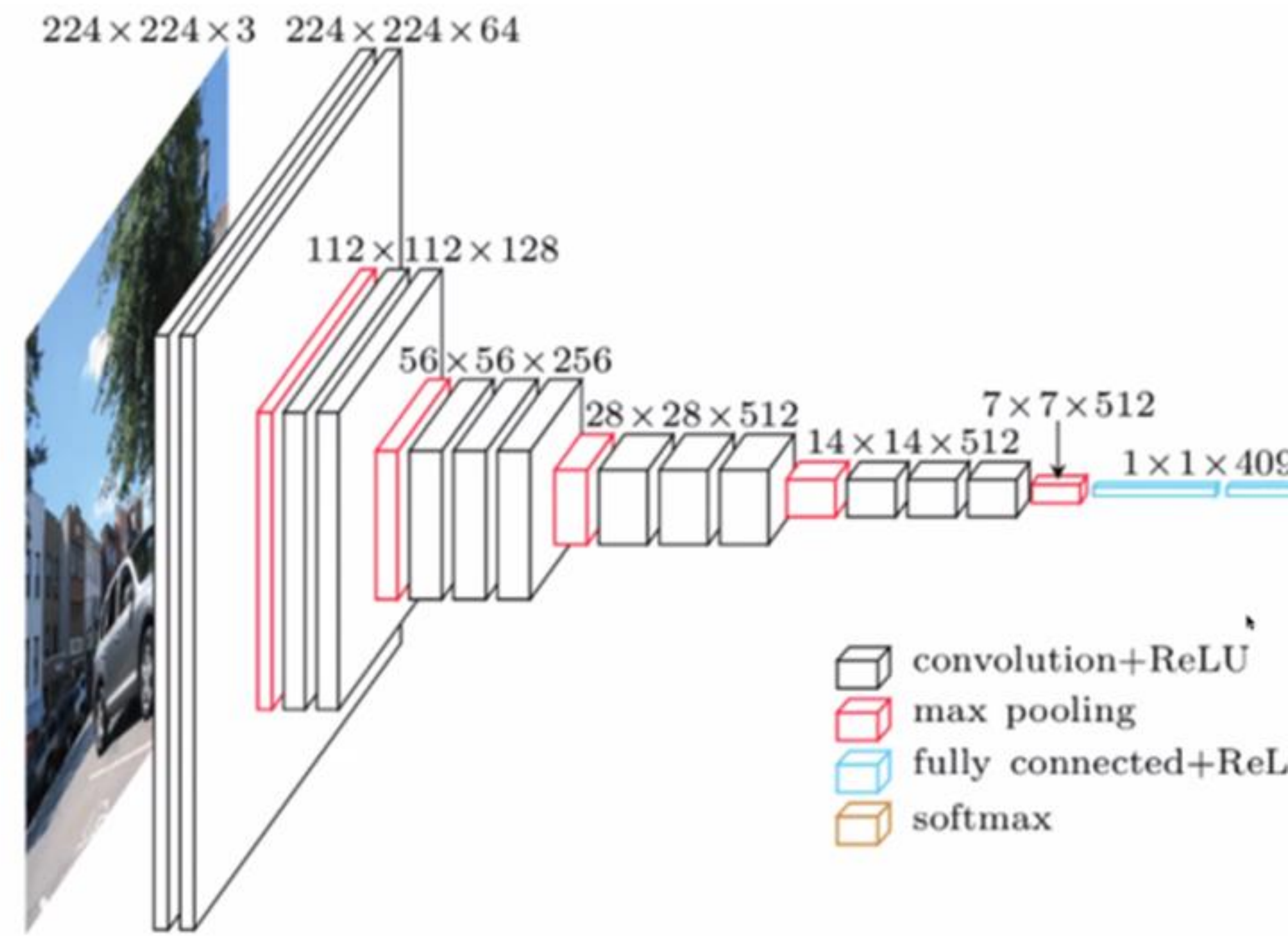
2.3 Review of AI for medical imaging and diagnosis
Prof Jong Chul Ye (KAIST)

Book mentioned earlier: https://link.springer.com/book/10.1007/978-981-16-6046-7

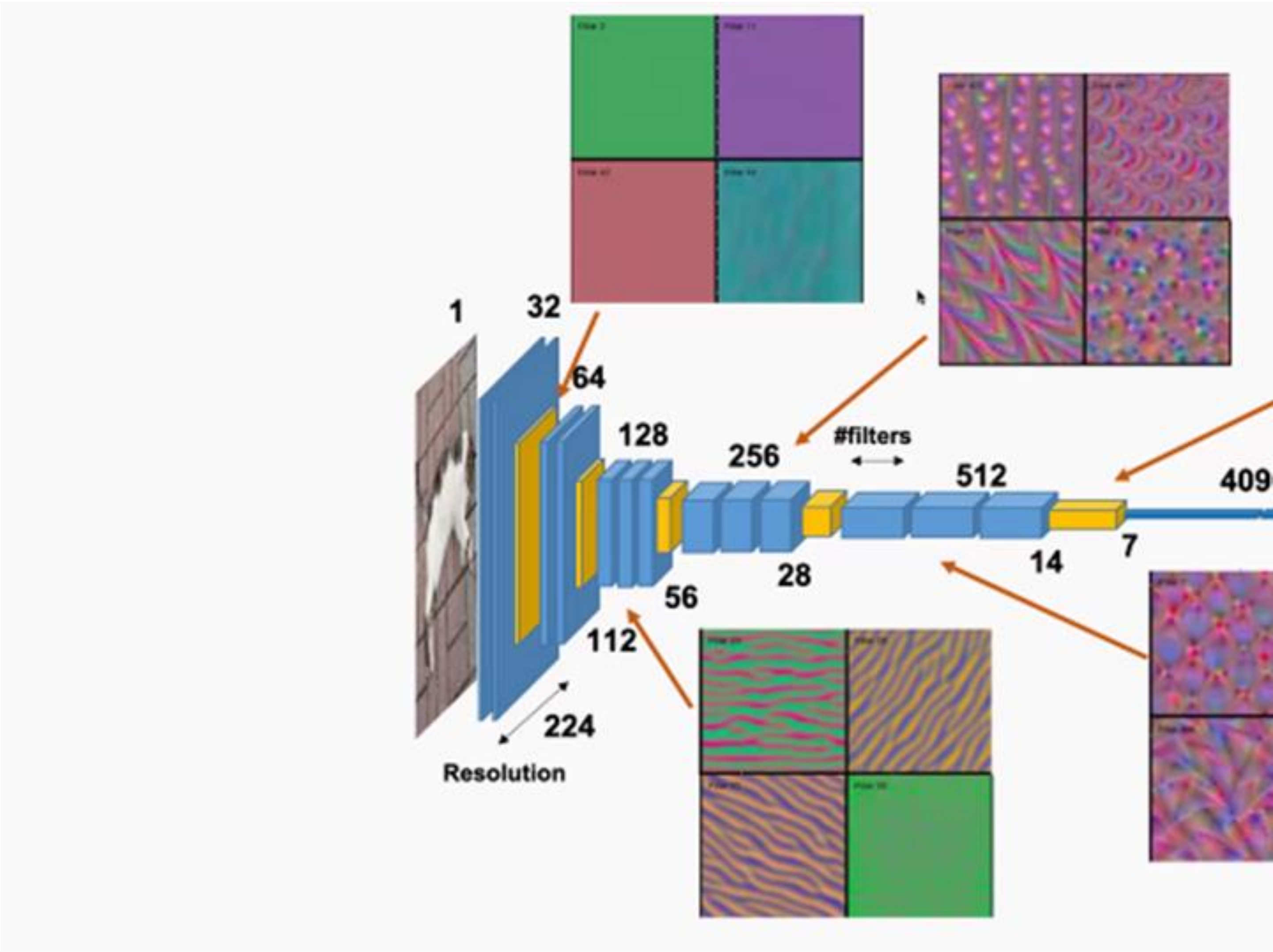
Technical Challenges in AI for Medical Imaging

- Limited data
  - Data privacy – **federated learning**
  - Cost of labelling – **self-supervised learning**
  - No paired reference – **generated models**
  - Overfitting – **vision transformer**
- Multimodal data – **vision language pretraining**

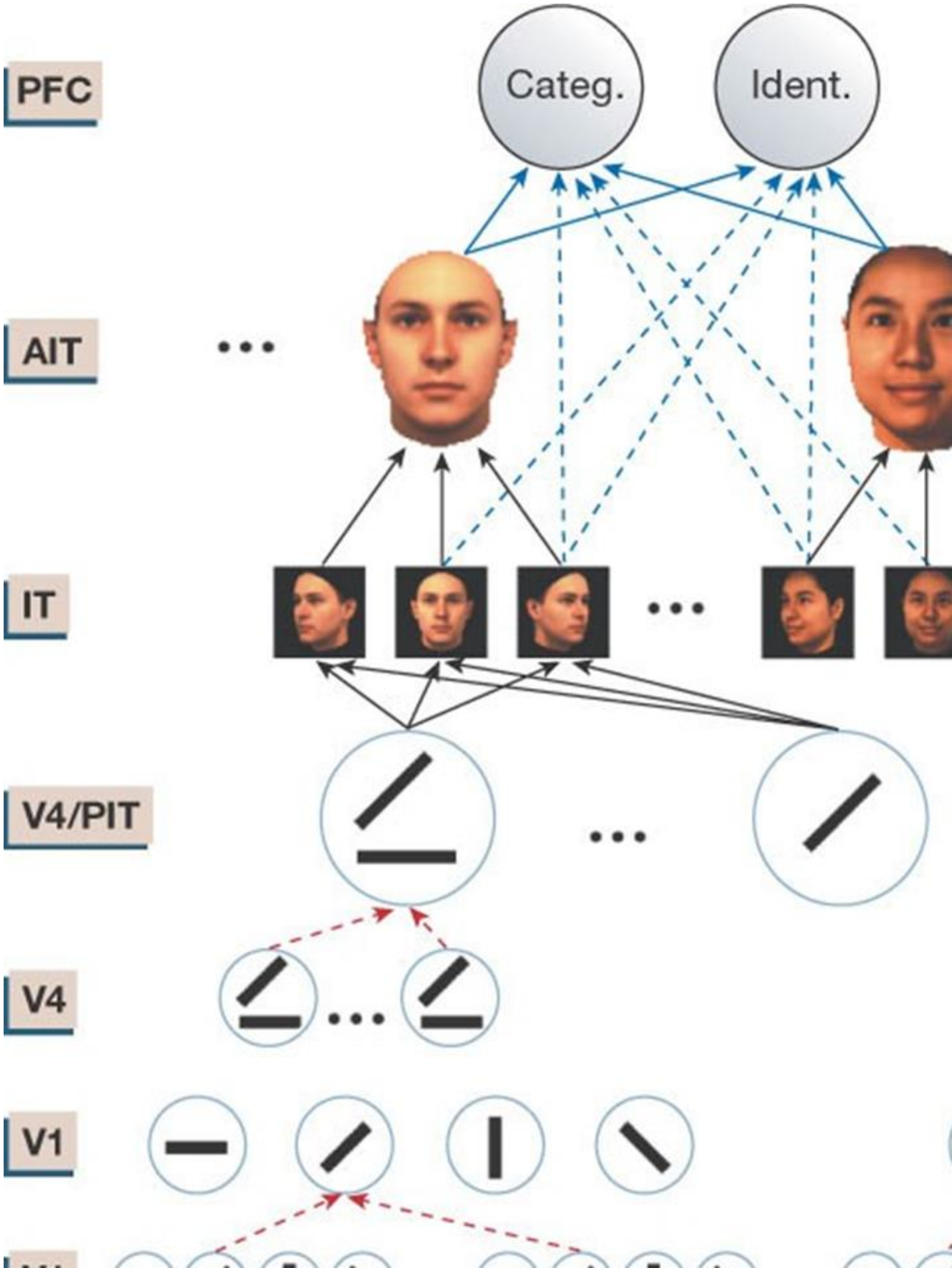
VGGNet: a CNN



Hierarchical Features in VGGNet



Information Processing in Brain



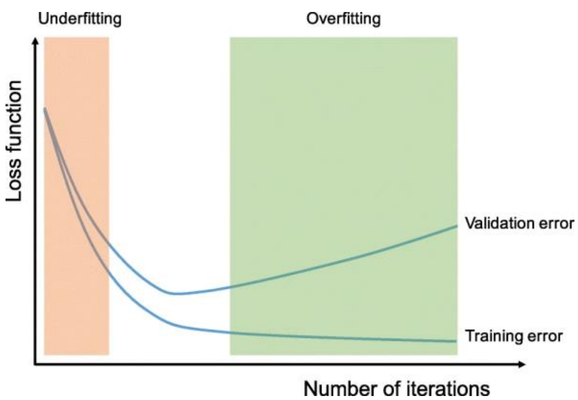
PFC
Categ.
Ident.
AIT
IT
V4/PIT
V4
V1

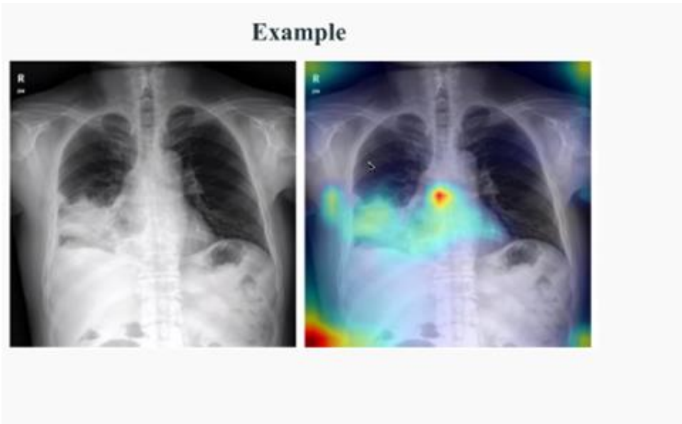
The model summarizes in quantitative terms other models and many data about visual recognition in the ventral stream pathway in cortex. The correspondence between the layers in the model and visual areas is an oversimplification. Circles represent neurons and arrows represent connections between them; the dots signify other neurons of the same type. Stages of neurons with bell-shaped tuning (with black arrow inputs), that provide example-based learning and generalization, are interleaved with stages that perform a max-like operation3 (denoted by red dashed arrows), which provides invariance to position and scale. An experimental example of the tuning postulated for the cells in the layer labelled inferotemporal in the model is shown in Fig. 1. The model accounts well for the quantitative data measured in view-tuned inferotemporal cortex cells10 (J. Pauls, personal communication) and for other experiments55. Superposition of gaussian-like units provides generalization to three-dimensional rotations and together with the soft-max stages some invariance to scale and position. IT, infratemporal cortex, AIT, anterior IT; PIT, posterior IT; PFC, prefrontal cortex. Adapted from M. Riesenhuber, personal communication.

Poggio, T., Bizzi, E. Generalization in vision and motor control. *Nature* **431**, 768–774 (2004). https://doi.org/10.1038/nature03014

Limitation of CNN

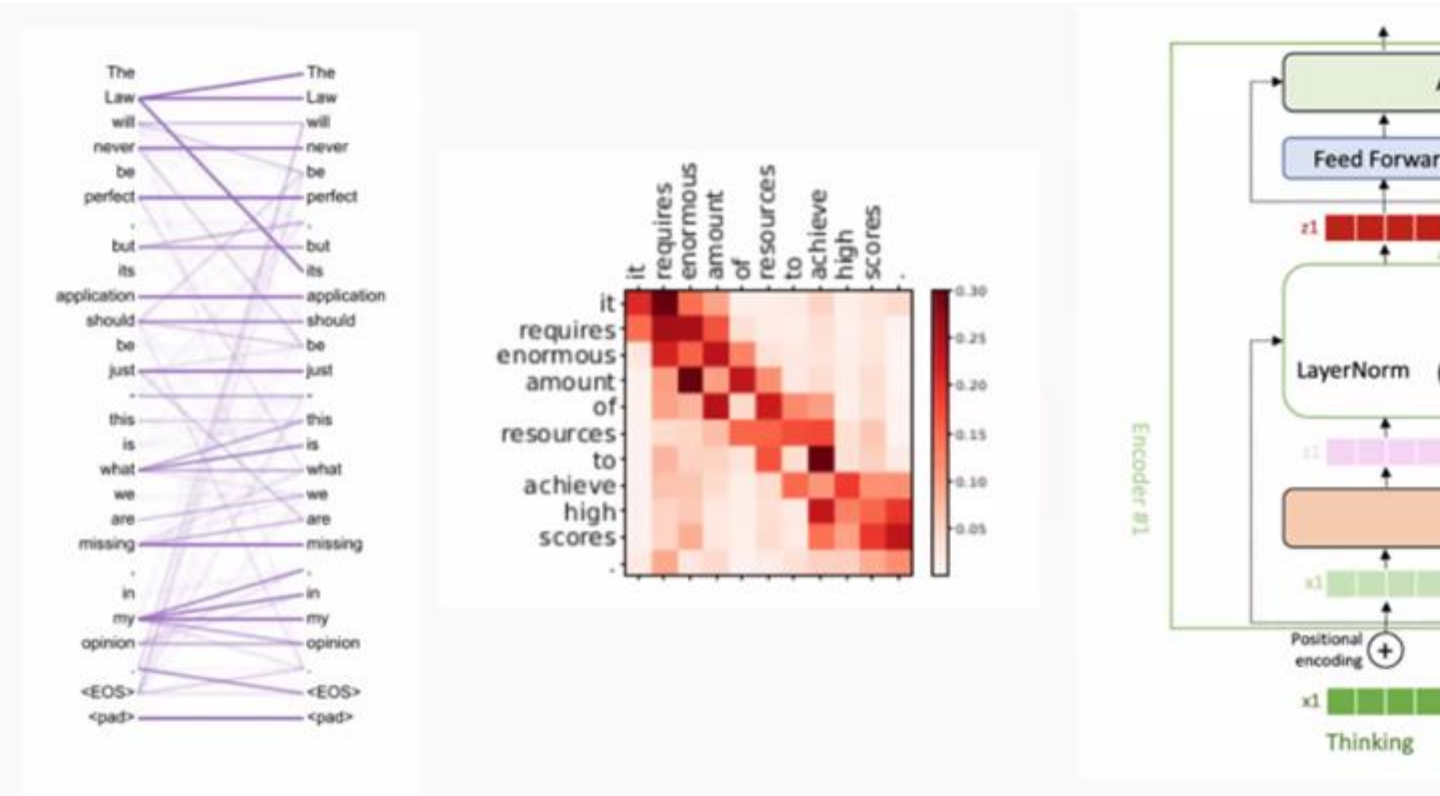
- Overfitting: Especially critical in medical imaging with limited data





A routine check for recognizing overfitting is to monitor the loss on the training and validation sets during the training iteration. If the model performs well on the training set compared to the validation set, then the model has been overfit to the training data. If the model performs poorly on both training and validation sets, then the model has been underfit to the data. Although the longer a network is trained, the better it performs on the training set, at some point, the network fits too well to the training data and loses its capability to generalize
Yamashita, R., Nishio, M., Do, R.K.G. et al. Convolutional neural networks: an overview and application in radiology. Insights Imaging 9, 611–629 (2018). https://doi.org/10.1007/s13244-018-0639-9

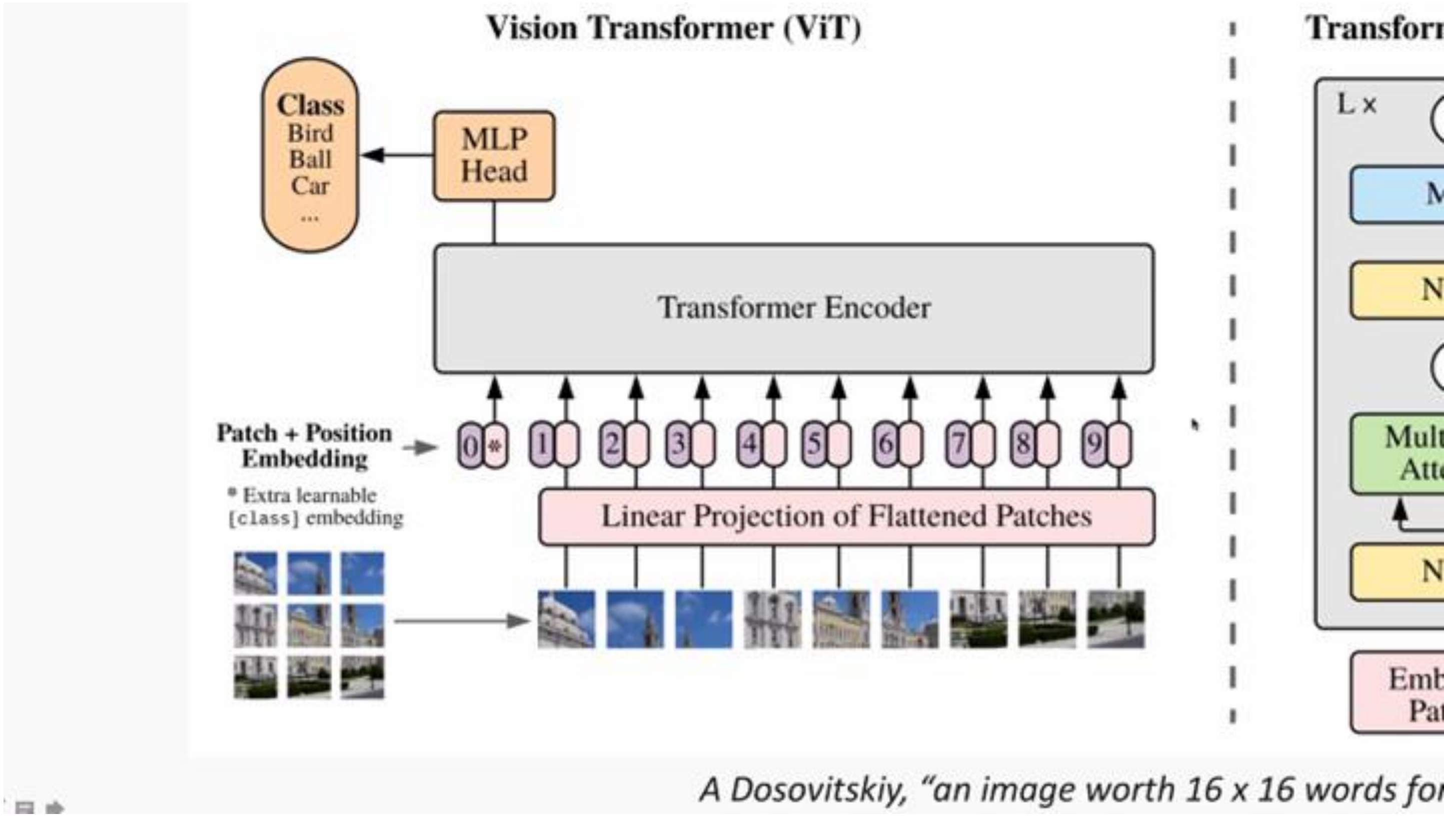
Transformer: Attention is All You Need



Vaswani, Ashish, et al. "Attention is all you need." *Advances in neur* *information processing systems* 30 (2017).

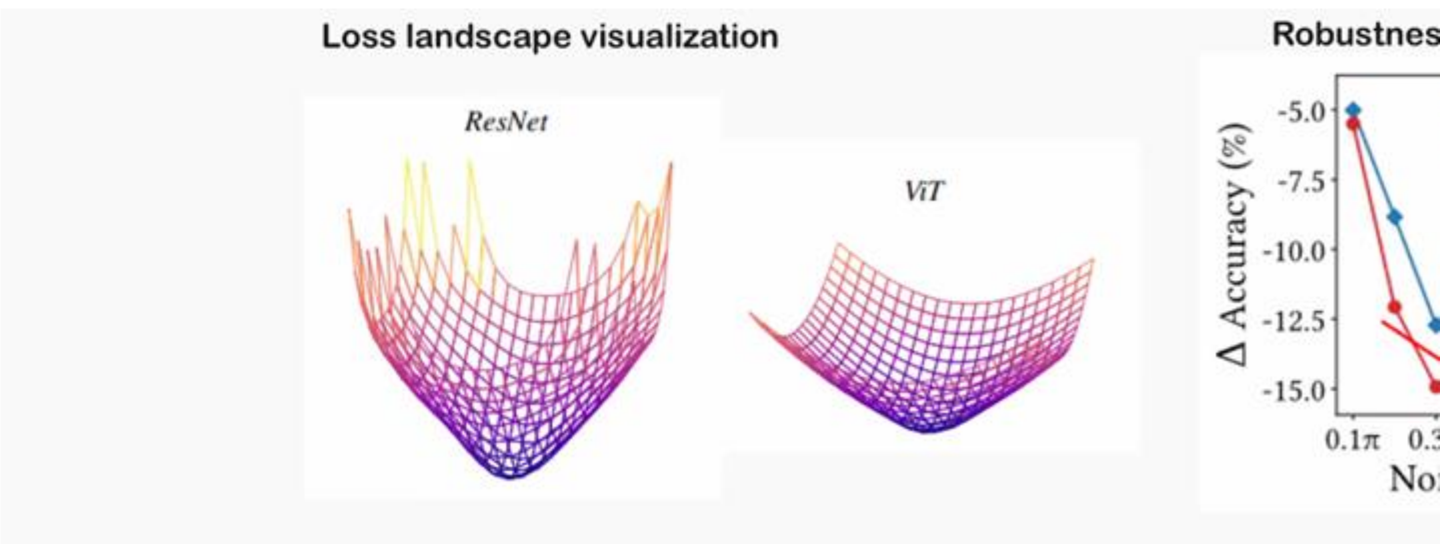
ViT: Vision Transformer: Farewell to convolution

- First successful approach to introduce **pure Transformer** to vision



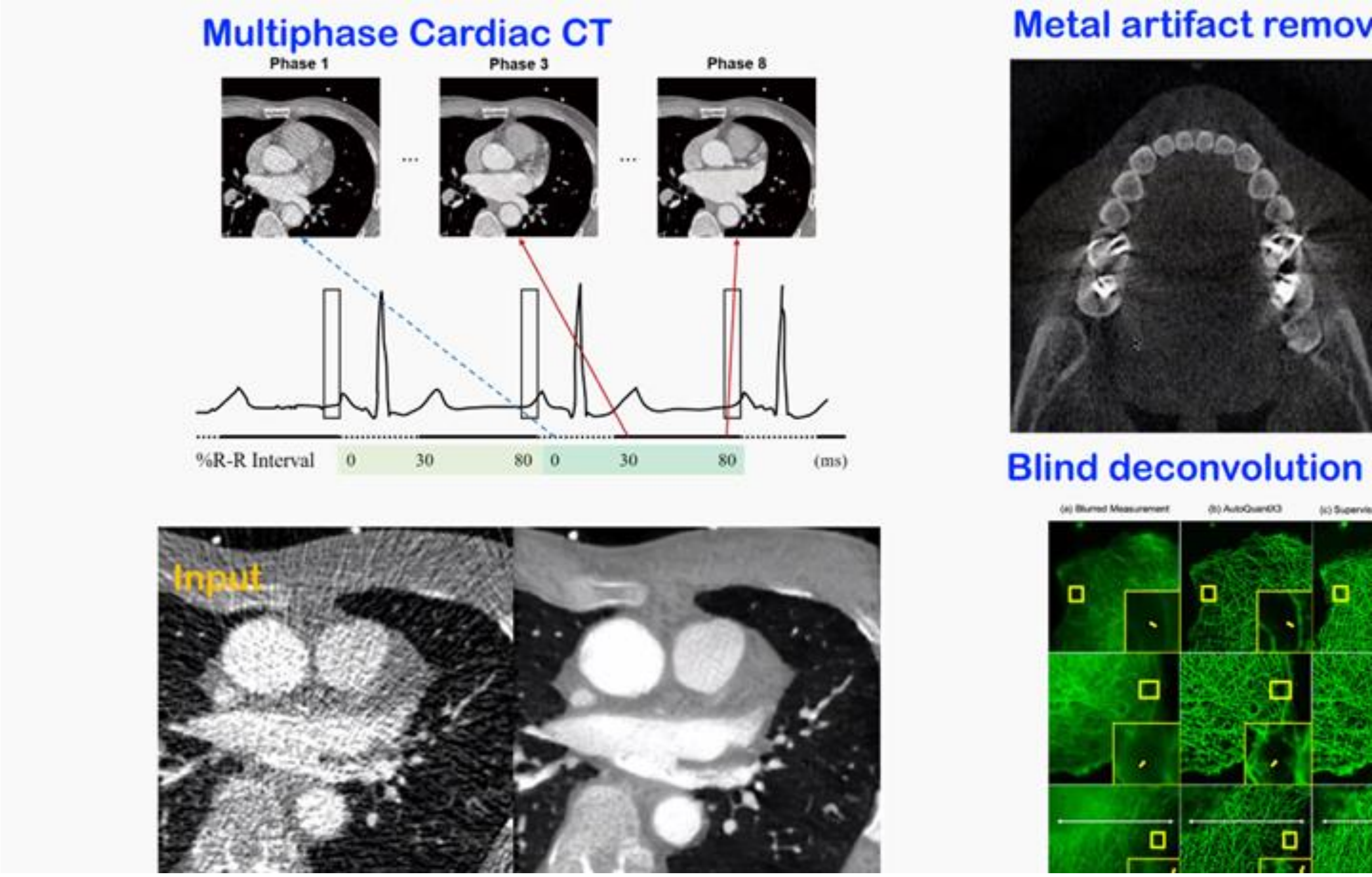
Why ViT works better than CNN?

- ViT can model long-range dependency between pixels.
- ViT has a more flat loss landscape than CNN (less overfitting).
- ViT is less vulnerable to high frequency noise than CNN.
- ViT is more shape-biased than CNN, like humans **(what we want!).**



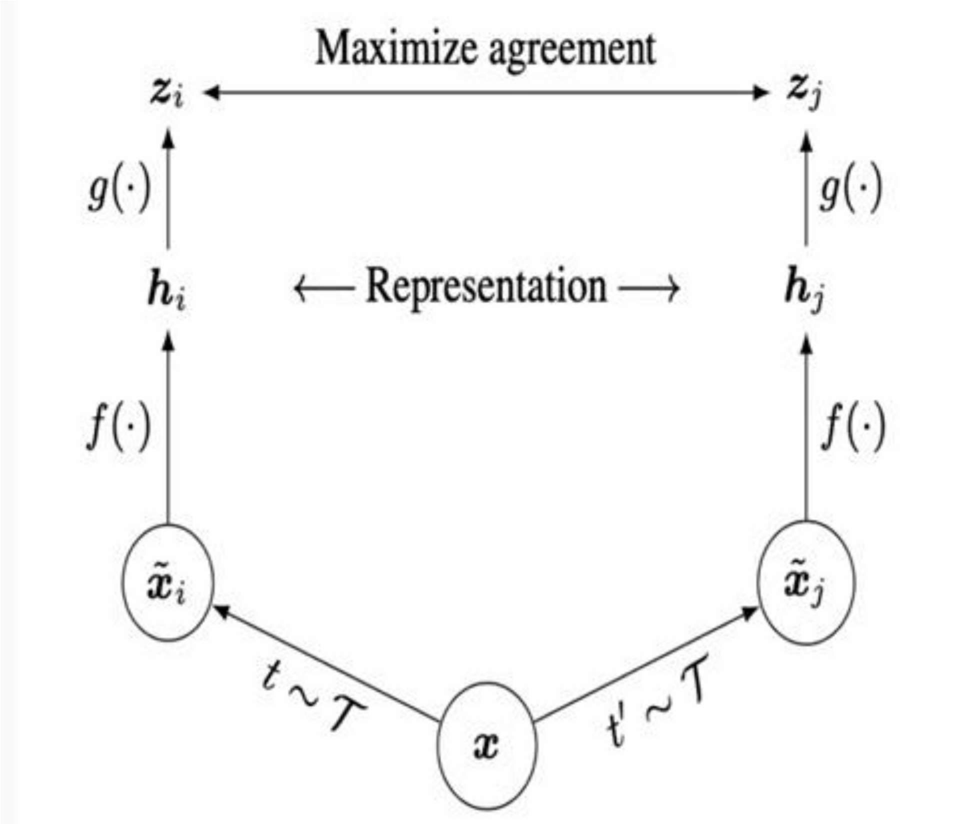
Self-Supervised Learning

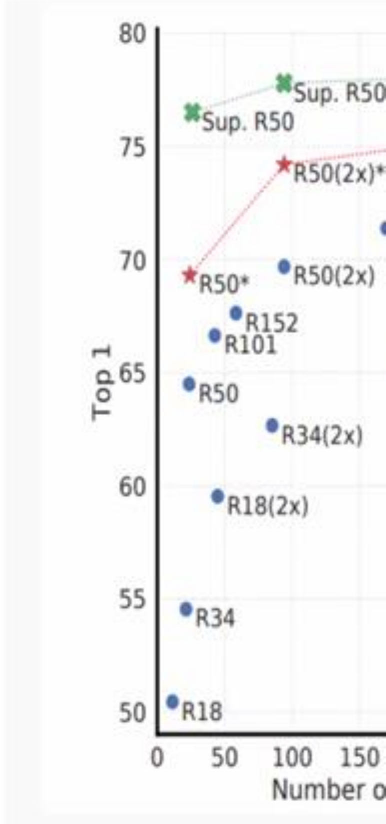
Limitation of Supervised Learning in Medical AI



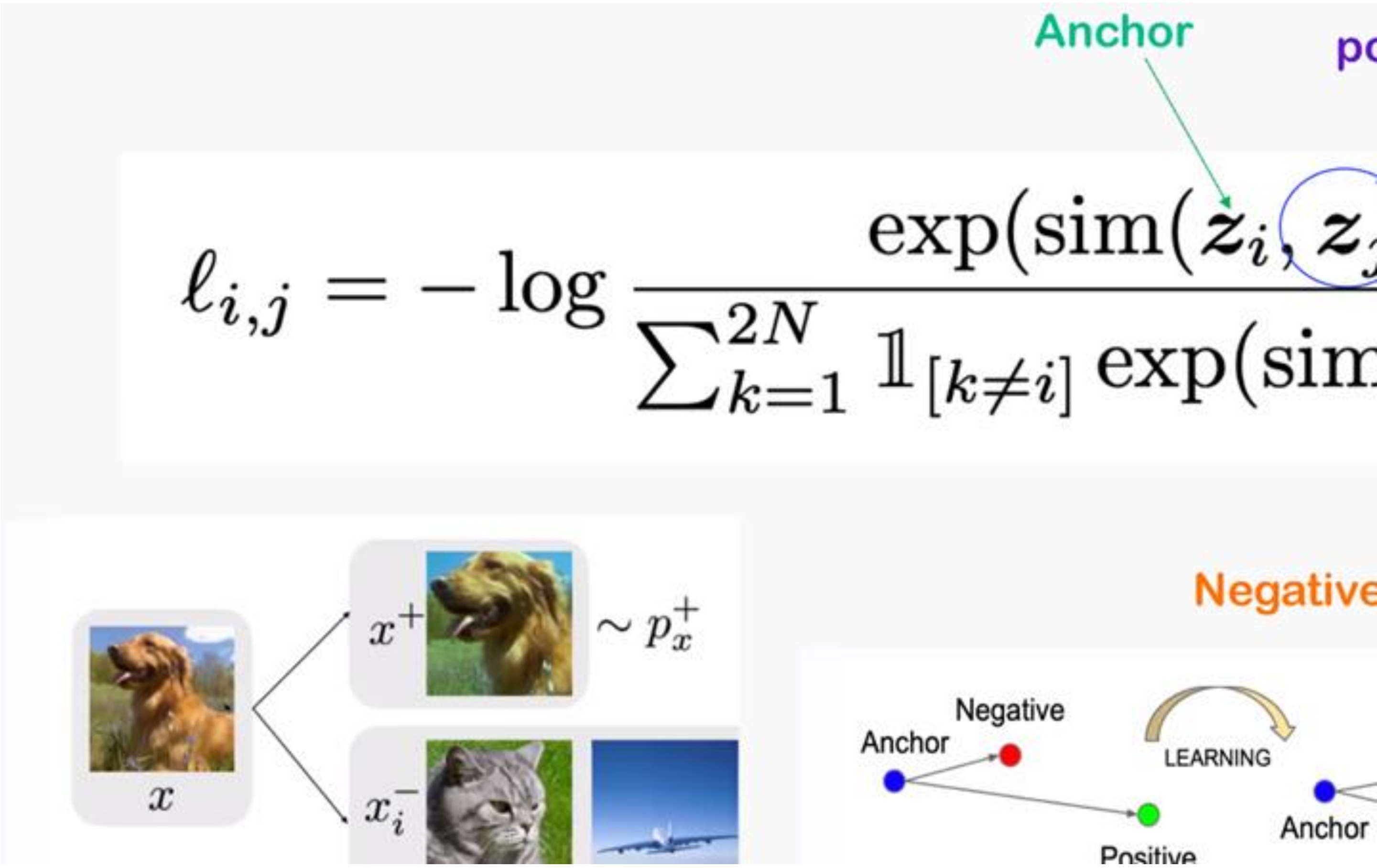
SimCLR: New Era of Contrastive Learning

(Chen et al, ICML 2020)





Contrastive Loss



Self-Supervised Learning with Distillation with No label (DINO)

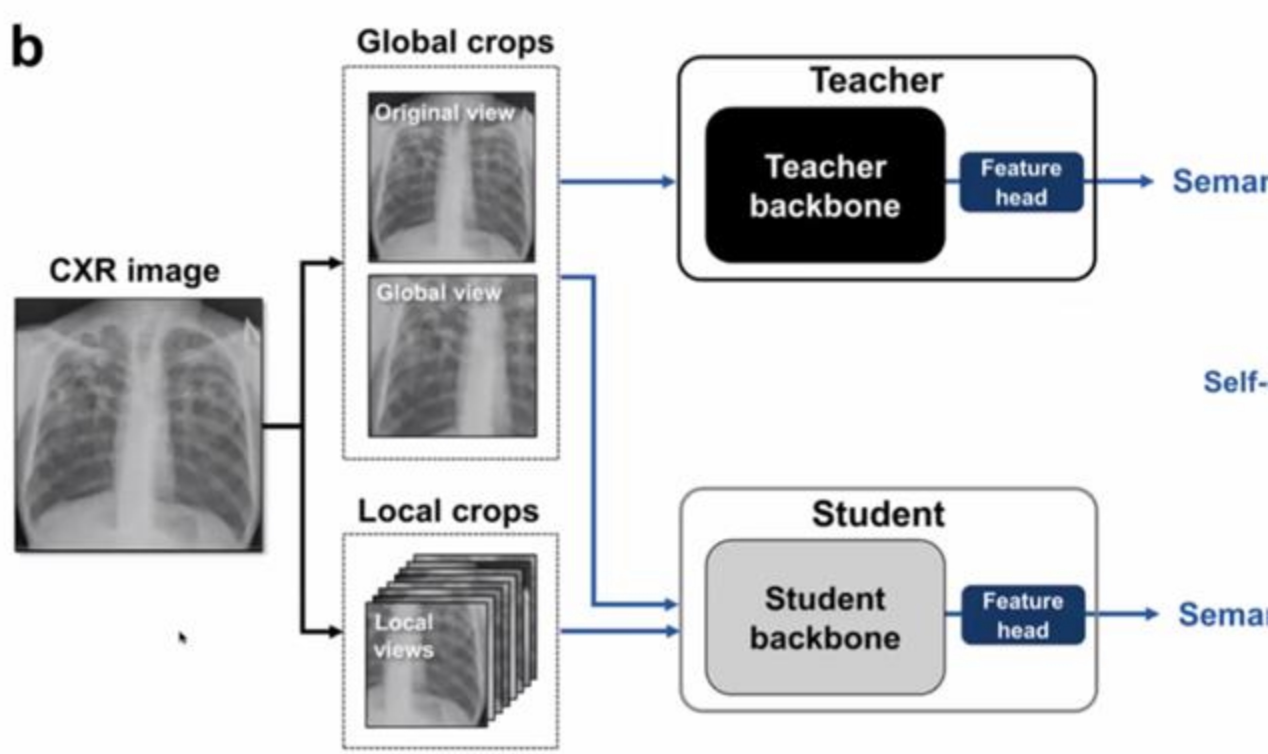
(CVPR 2021)

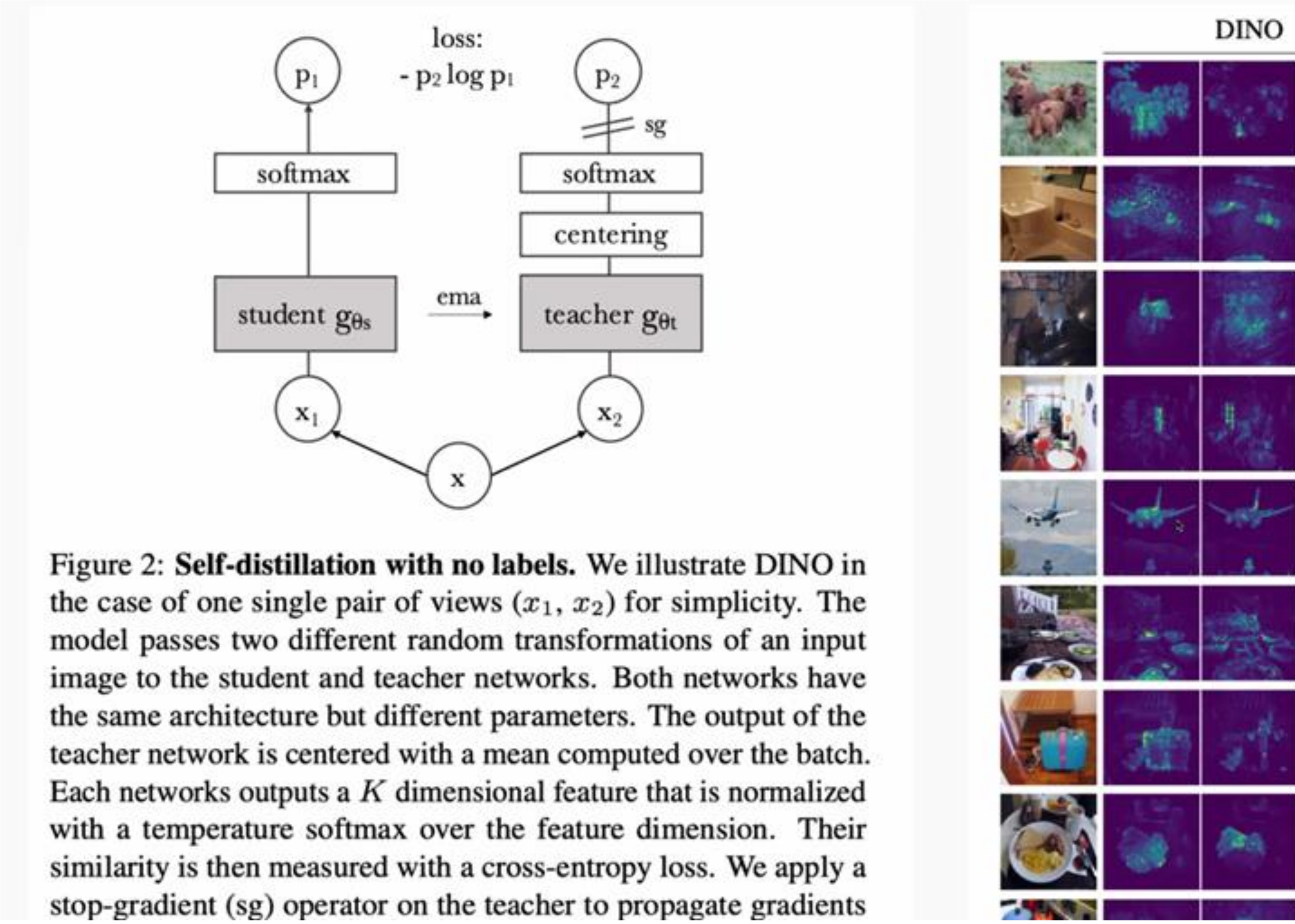
Emerging Properties in Self-Supervised Vision Transforn

Mathilde Caron[1,2] Hugo Touvron[1,3] Ishan Misra[1] Hervé Jeg
Julien Mairal[2] Piotr Bojanowski[1] Armand Joulin[1]

[1] Facebook AI Research [2] Inria* [3] Sorbonne University

Figure 1: **Self-attention from a Vision Transformer with 8×8 patches trained with no supervision.** We look

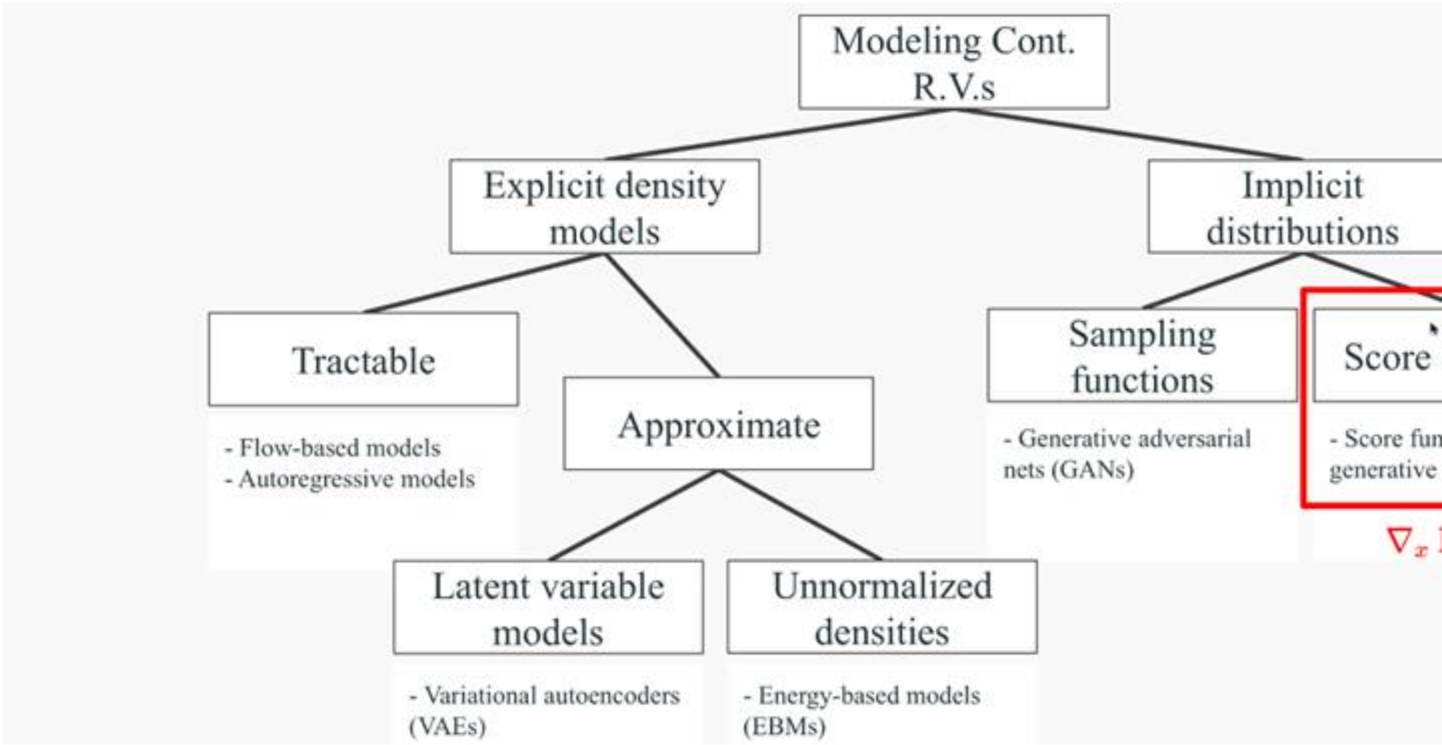




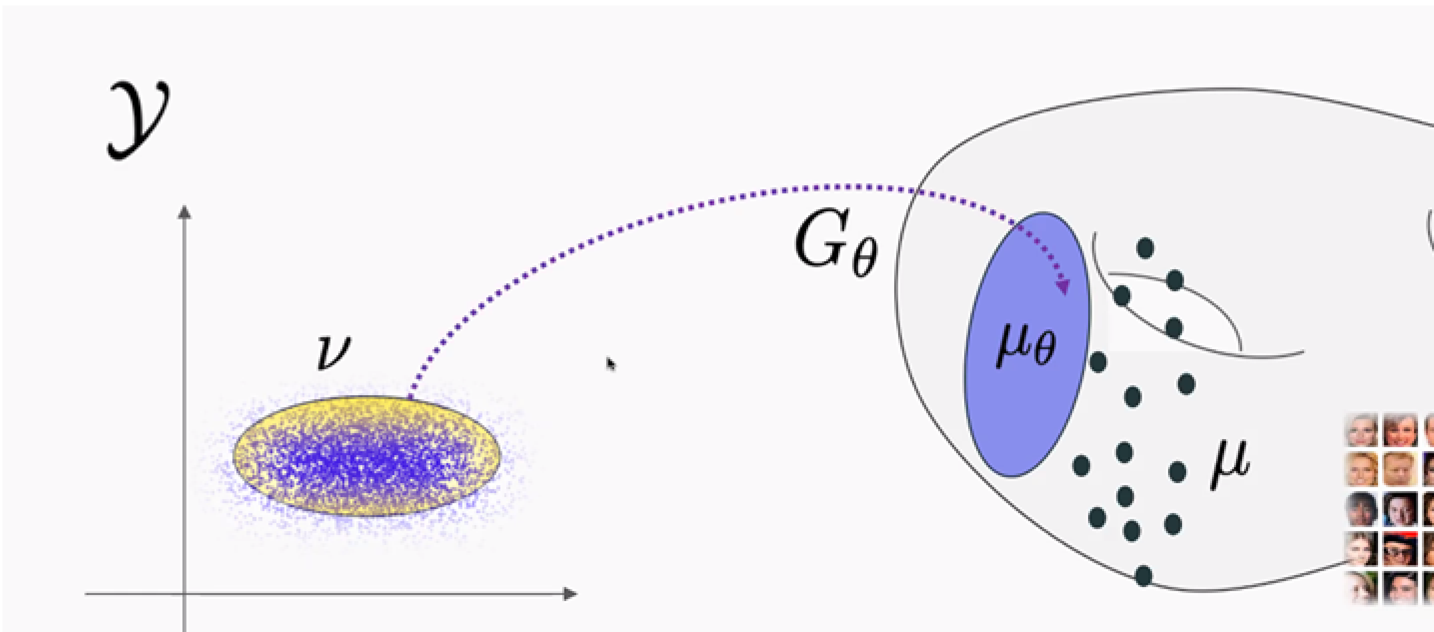
Generative Models

Various Ways of Modelling Continuous Variables

- Various ways to model continuous random variables
    - This taxonomic tree doesn't count on "training methods"

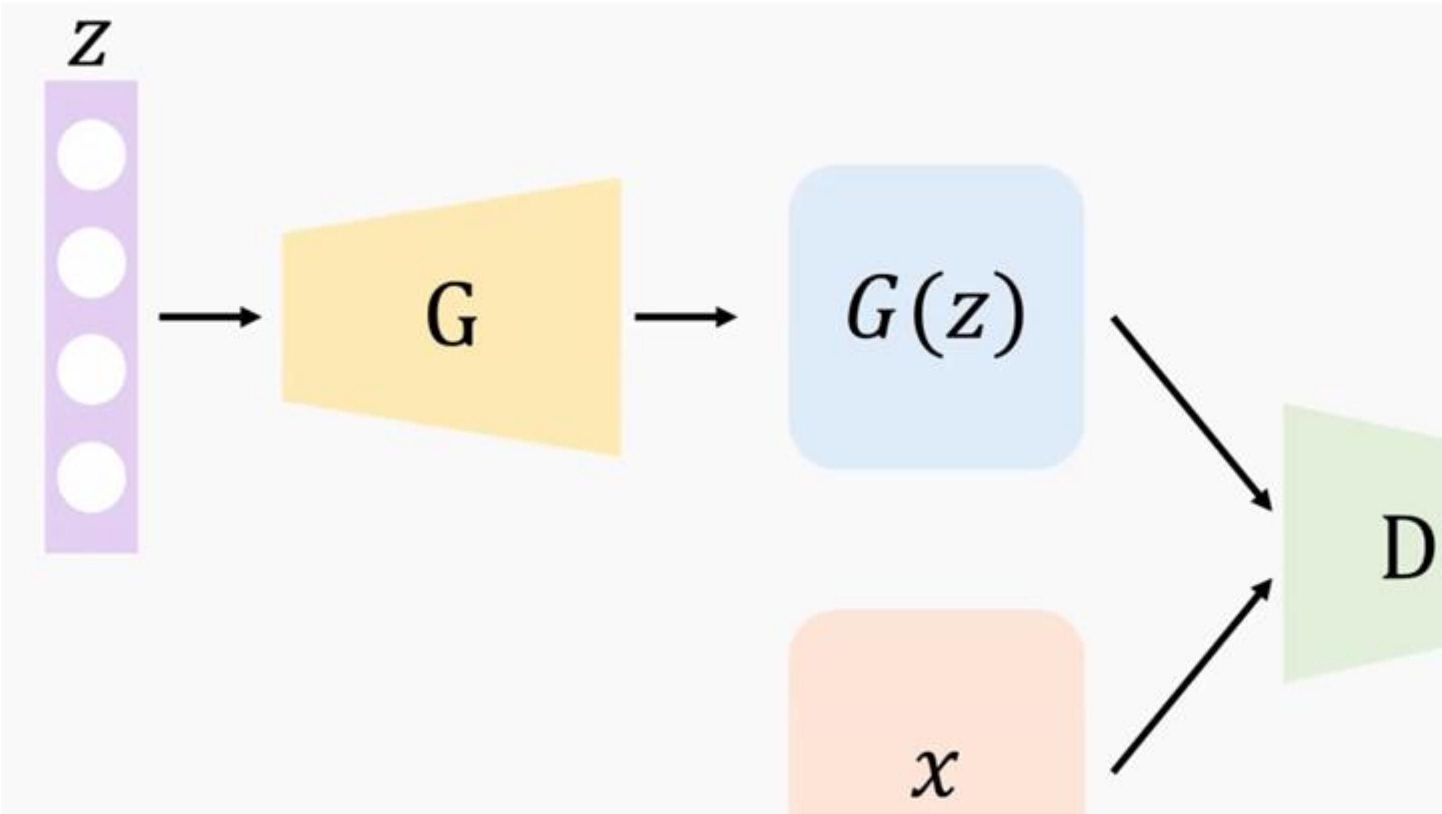


Generative Adversarial Nets (GAN)

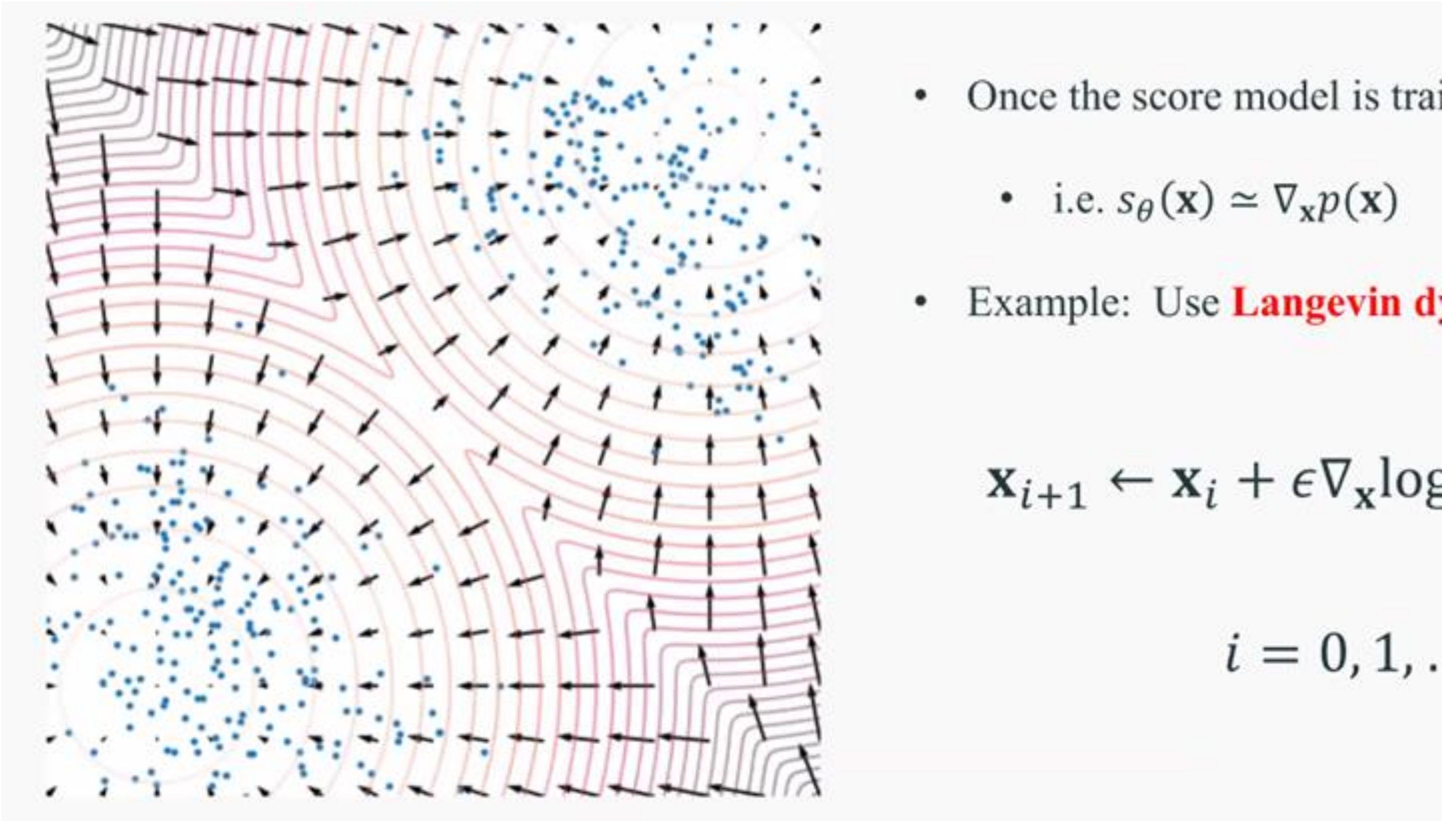


$$\min_{\theta} \operatorname{dist}(\mu_\theta, \mu)$$

subject to $\mu_\theta = G_{\theta\#}\nu$



Diffusion-based Generative Models



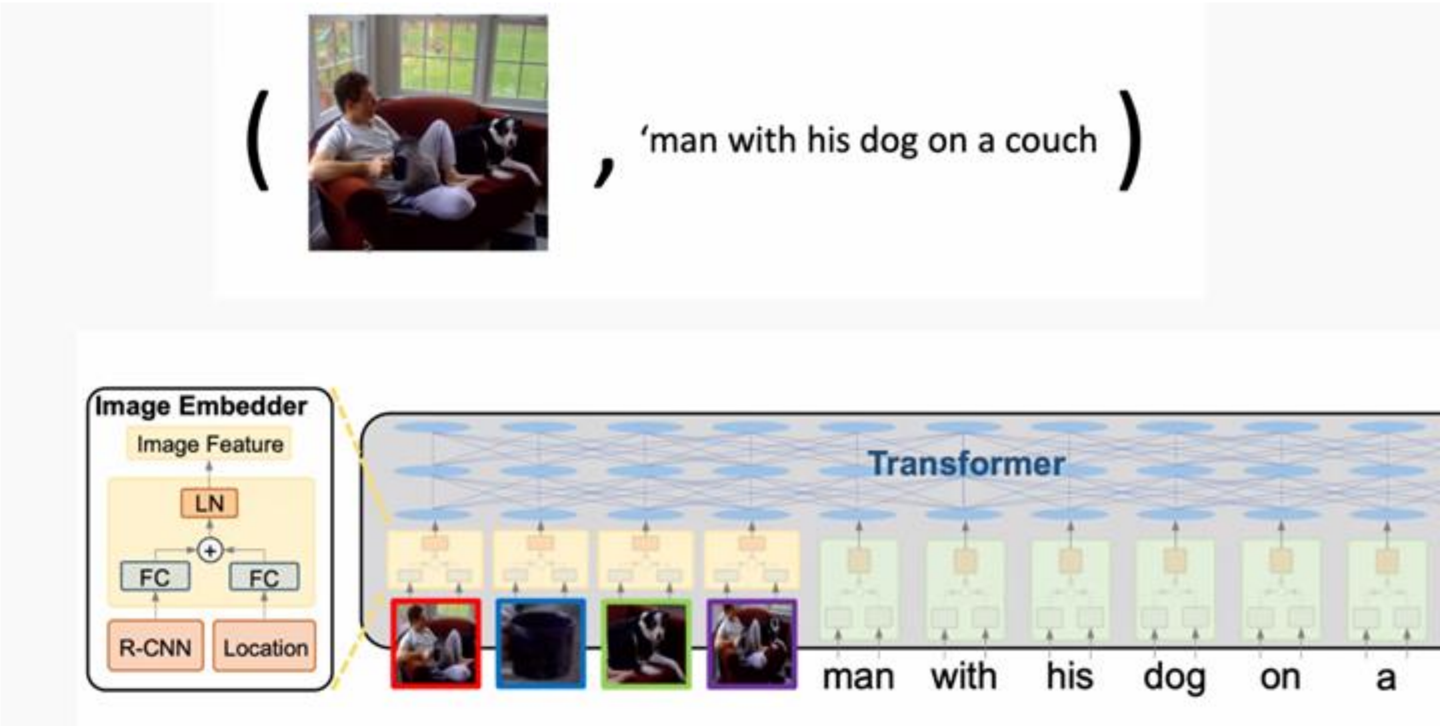
GAN vs Diffusion Model

Text-Guided Image Generation

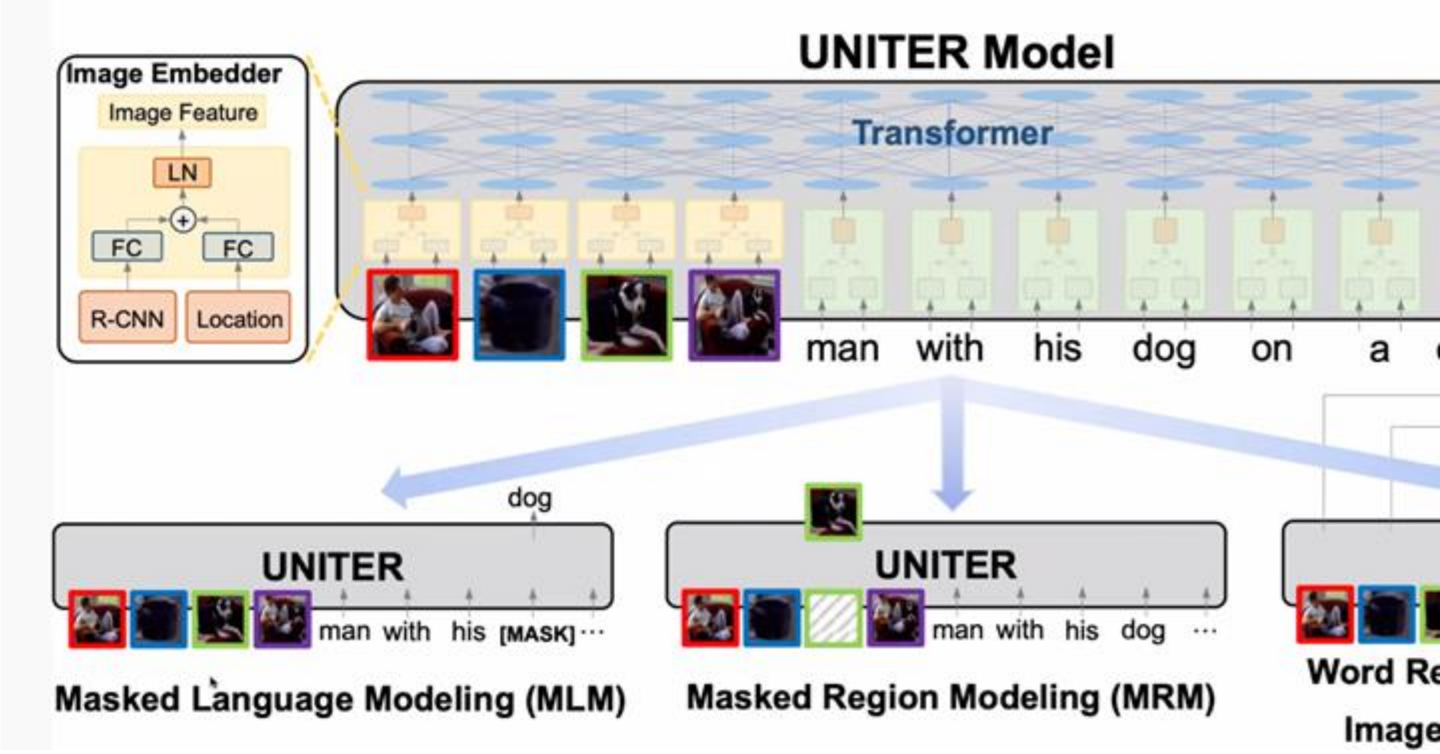
https://openai.com/dall-e-2/

Vision Language Pretraining

Single Stream Architecture: UNITER (Chen et al, 2019)



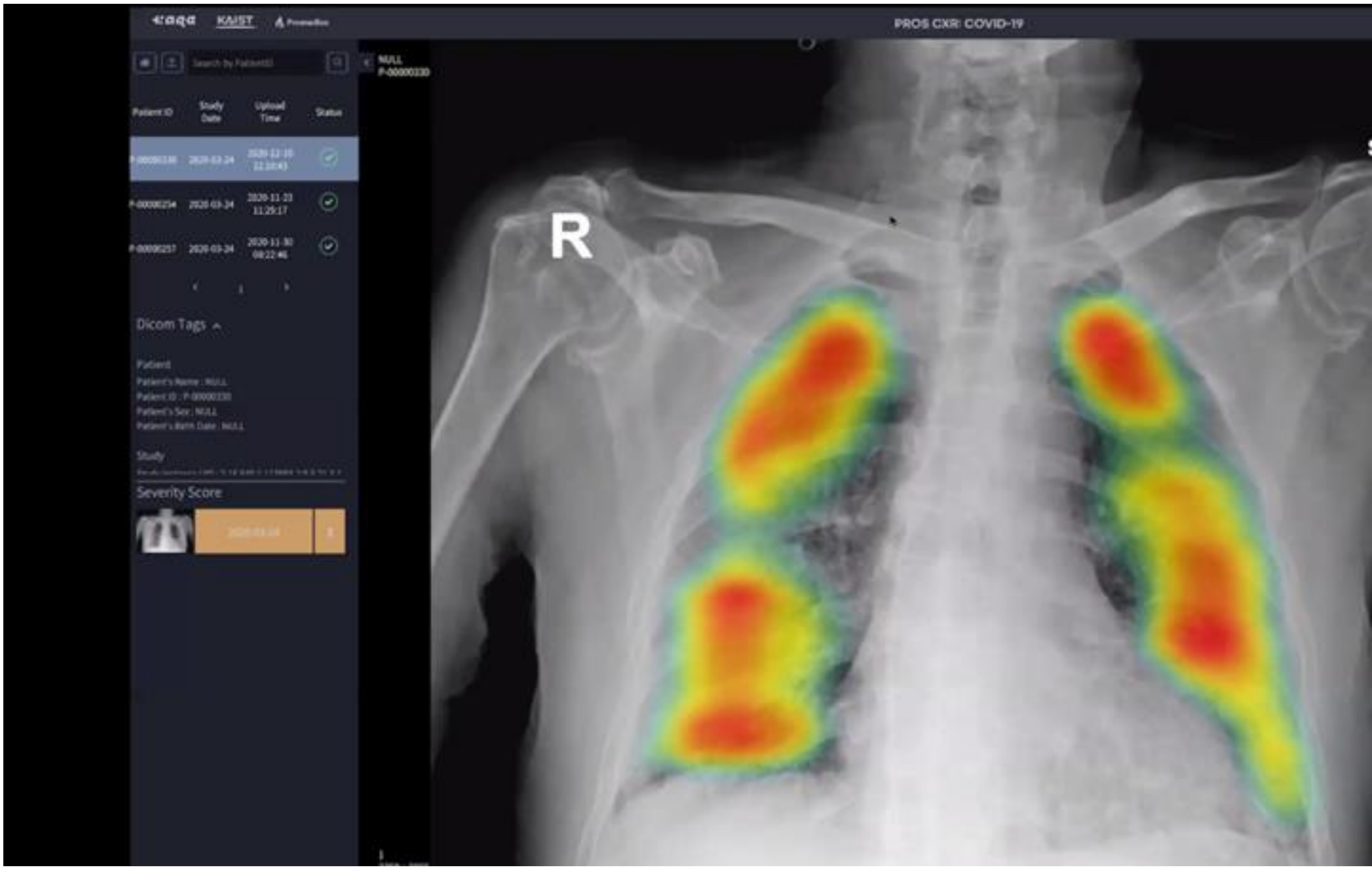
Pretraining

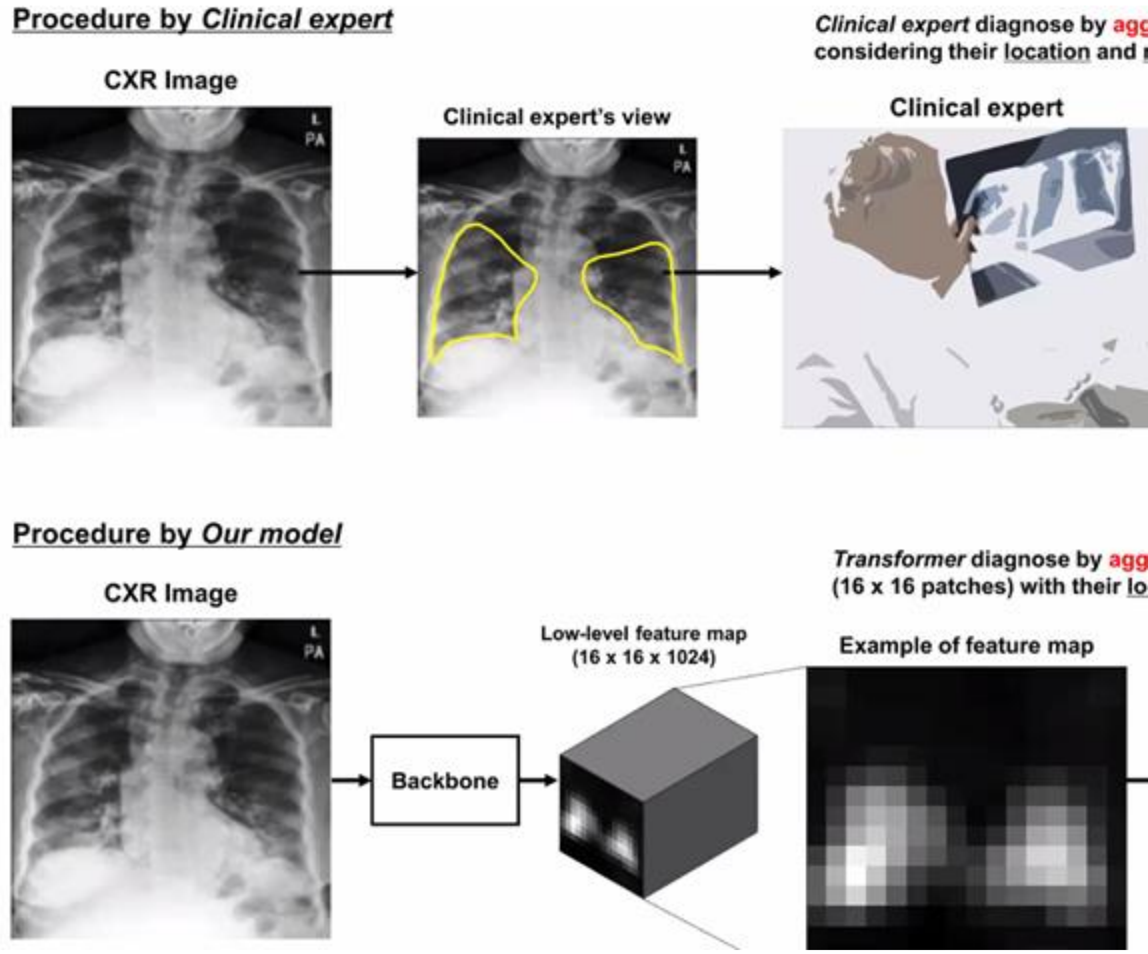


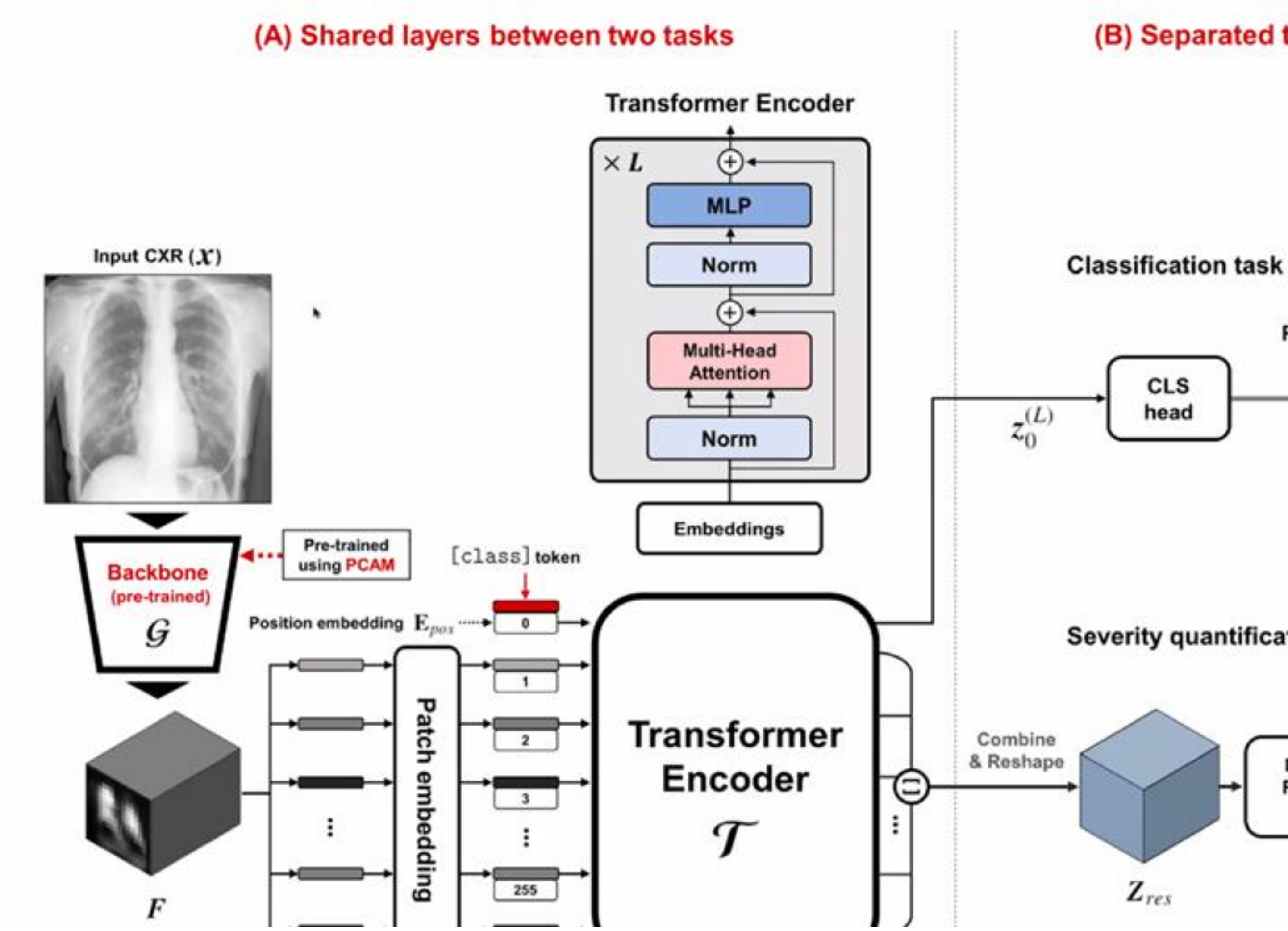
Downstream Tasks

AI-driven Diagnosis

COVID-19 Detection by CXR (Park et al, MEDIA, 2021)

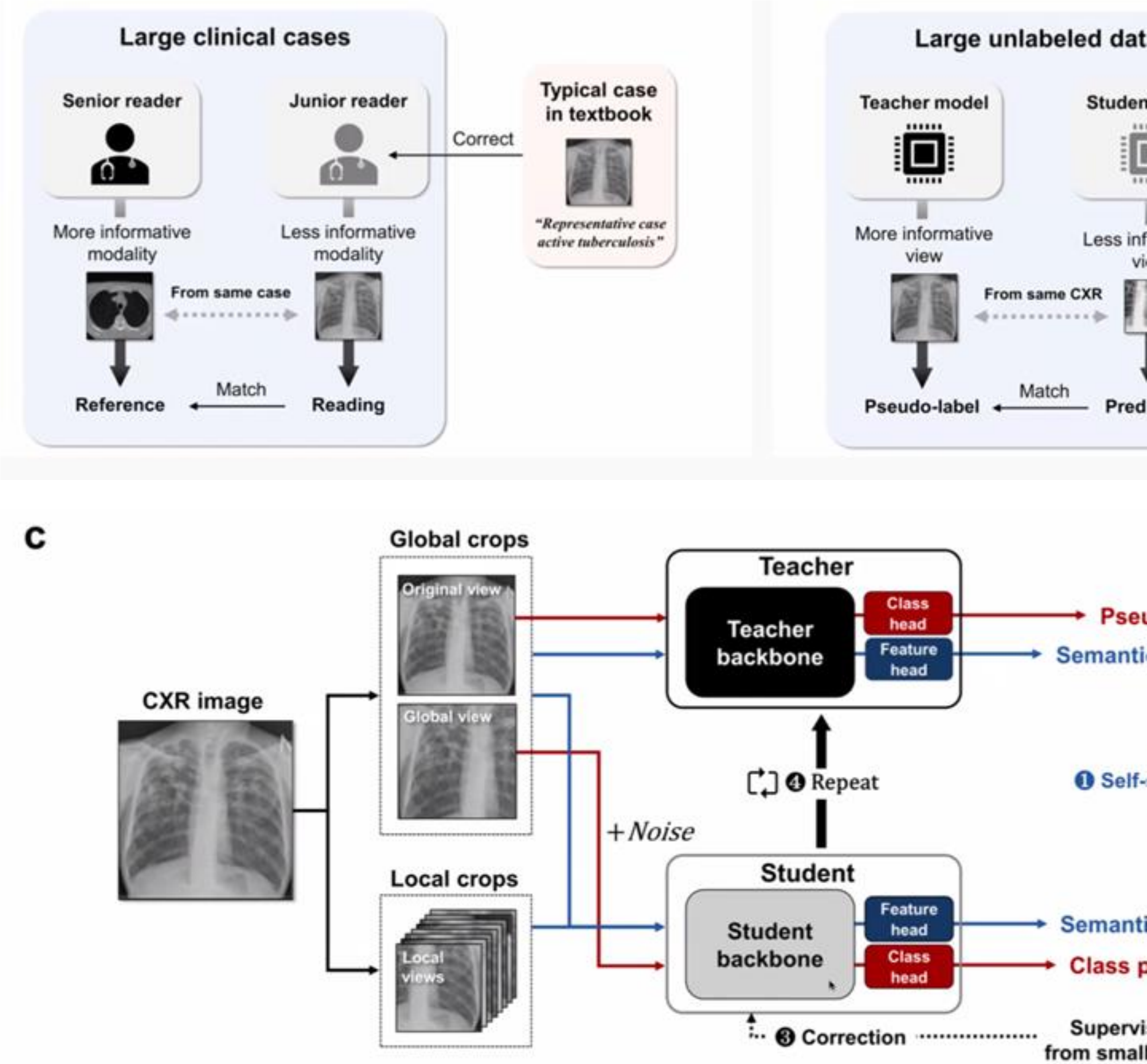

KAIST
PROS CXR: COVID-19
Patient ID
Study Date
Upload Time
Status
Dicom Tags
Severity Score
R


Procedure by Clinical expert
CXR Image
Clinical expert's view
Clinical expert diagnose by agg
considering their location and
Clinical expert
Procedure by Our model
CXR Image
Backbone
Low-level feature map
(16 x 16 x 1024)
Transformer diagnose by agg
(16 x 16 patches) with their lo
Example of feature map

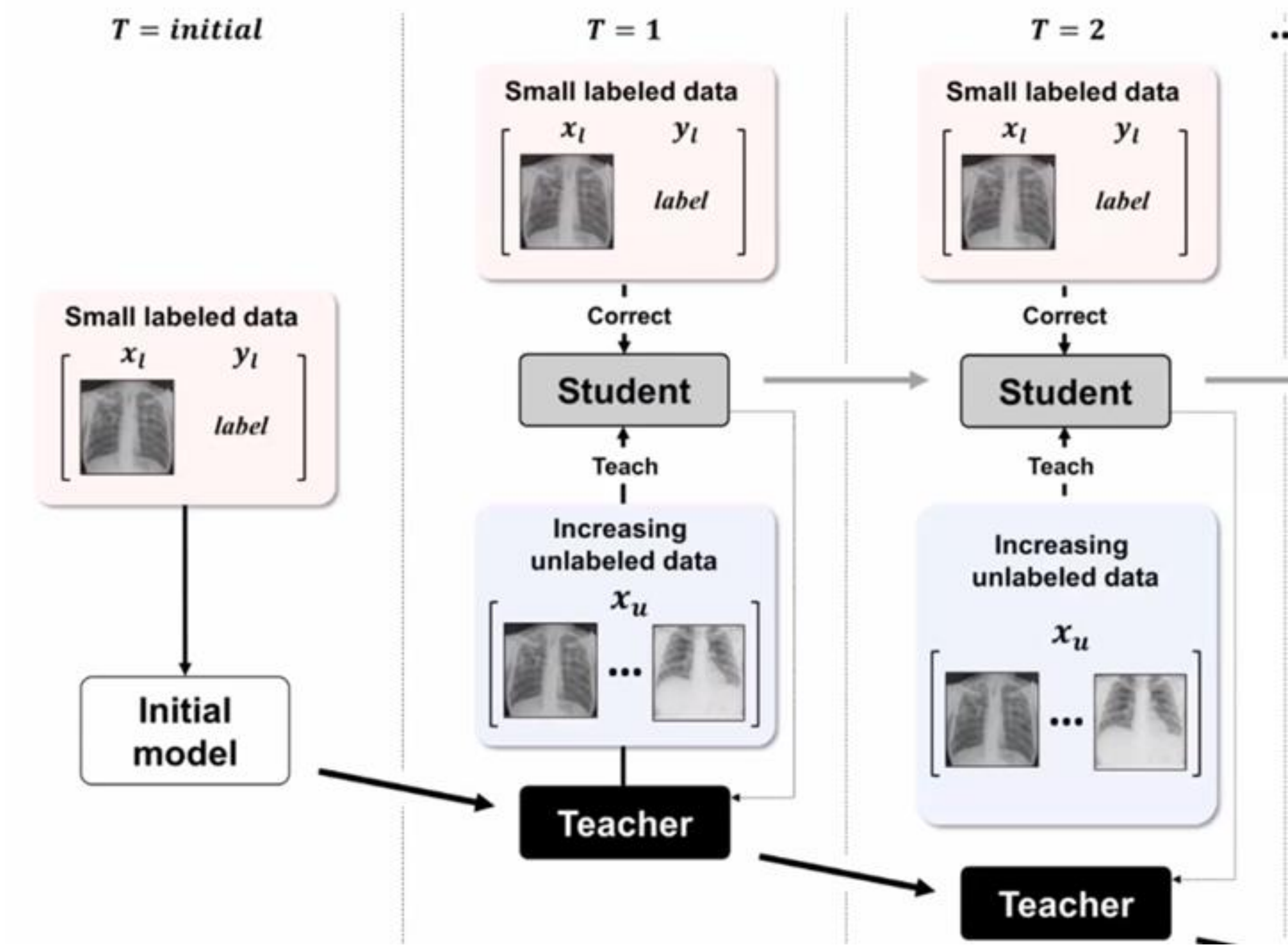


Distillation for Self-Supervised & Self-Train Learning (DISTL)

(Park et al, *Nature Comm*, 2022)

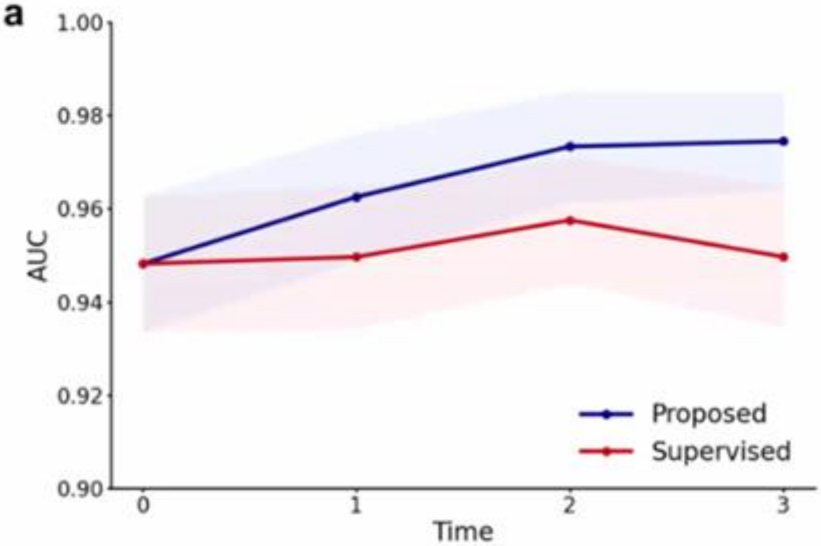


Large clinical cases
Senior reader
Junior reader
Correct
Typical case in textbook
"Representative case active tuberculosis"
More informative modality
Less informative modality
From same case
Reference
Match
Reading
Large unlabeled dat
Teacher model
More informative view
From same CXR
Pseudo-label
Match
c
Global crops
Original view
Global view
CXR image
Local crops
Local views
+Noise
Teacher
Teacher backbone
Class head
Feature head
Repeat
Student
Student backbone
Feature head
Class head
Correction



$T = initial$
$T = 1$
$T = 2$
Small labeled data
x_l
y_l
label
Correct
Student
Teach
Increasing unlabeled data
x_u
Initial model
Teacher

a



b

Methods	$T = initial$	$T = 1$	$T = 2$	$T = 3$
Proposed	0.948	0.962	0.973	0.974
Supervised	0.948	0.950*	0.958*	0.950*

c

Metrics	Pooled	CNUH
AUC (95% CI)	0.974 (0.964-0.985)	0.965 (0.926-1.0
Sensitivity (95% CI)	92.7 (89.3-95.3)	92.9 (76.5-99.
Specificity (95% CI)	92.0 (90.2-93.5)	90.3 (86.9-93.
Accuracy (95% CI)	92.2 (90.6-93.5)	90.4 (87.2-93.
PPV (95% CI)	0.776 (0.738-0.809)	0.400 (0.327-0.4
NPV (95% CI)	0.977 (0.966-0.984)	0.995 (0.979-0.9

d

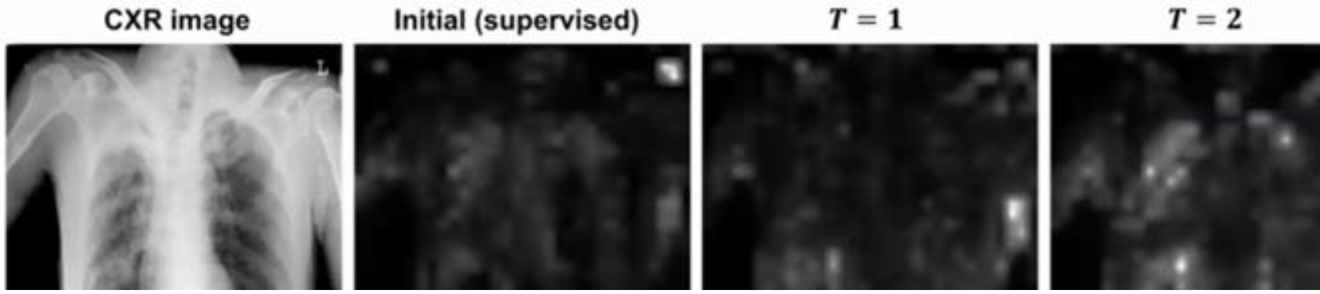
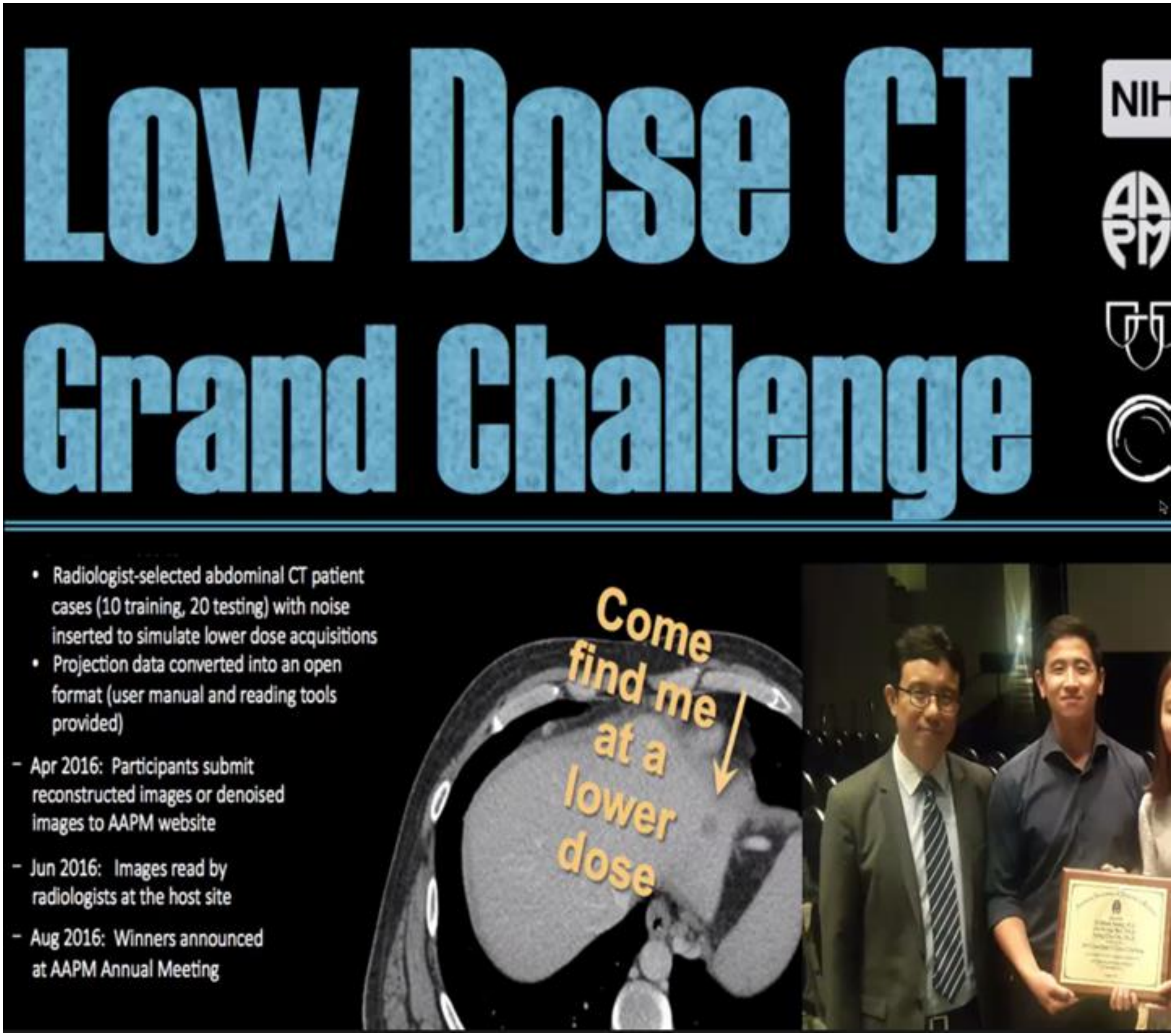
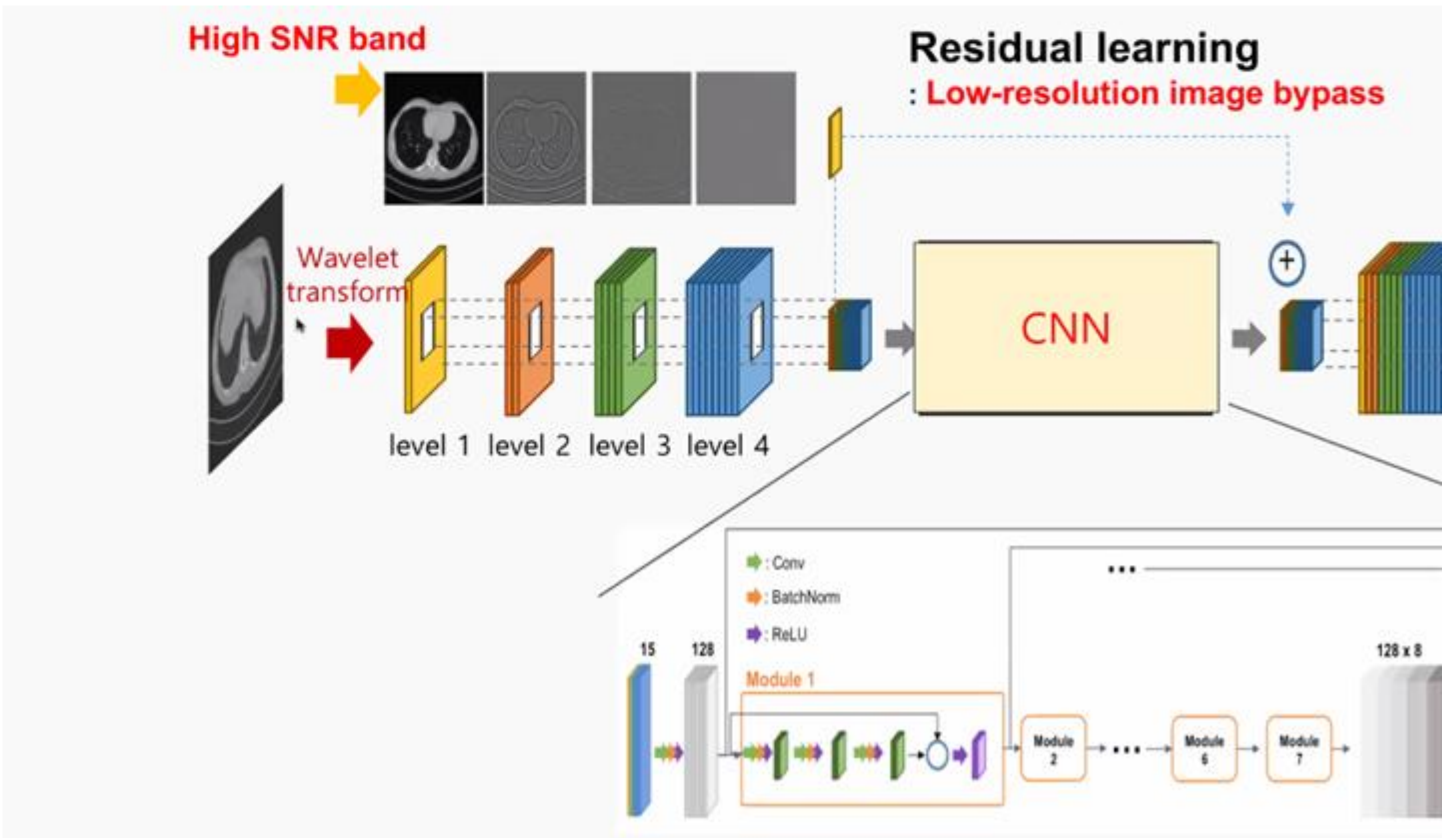
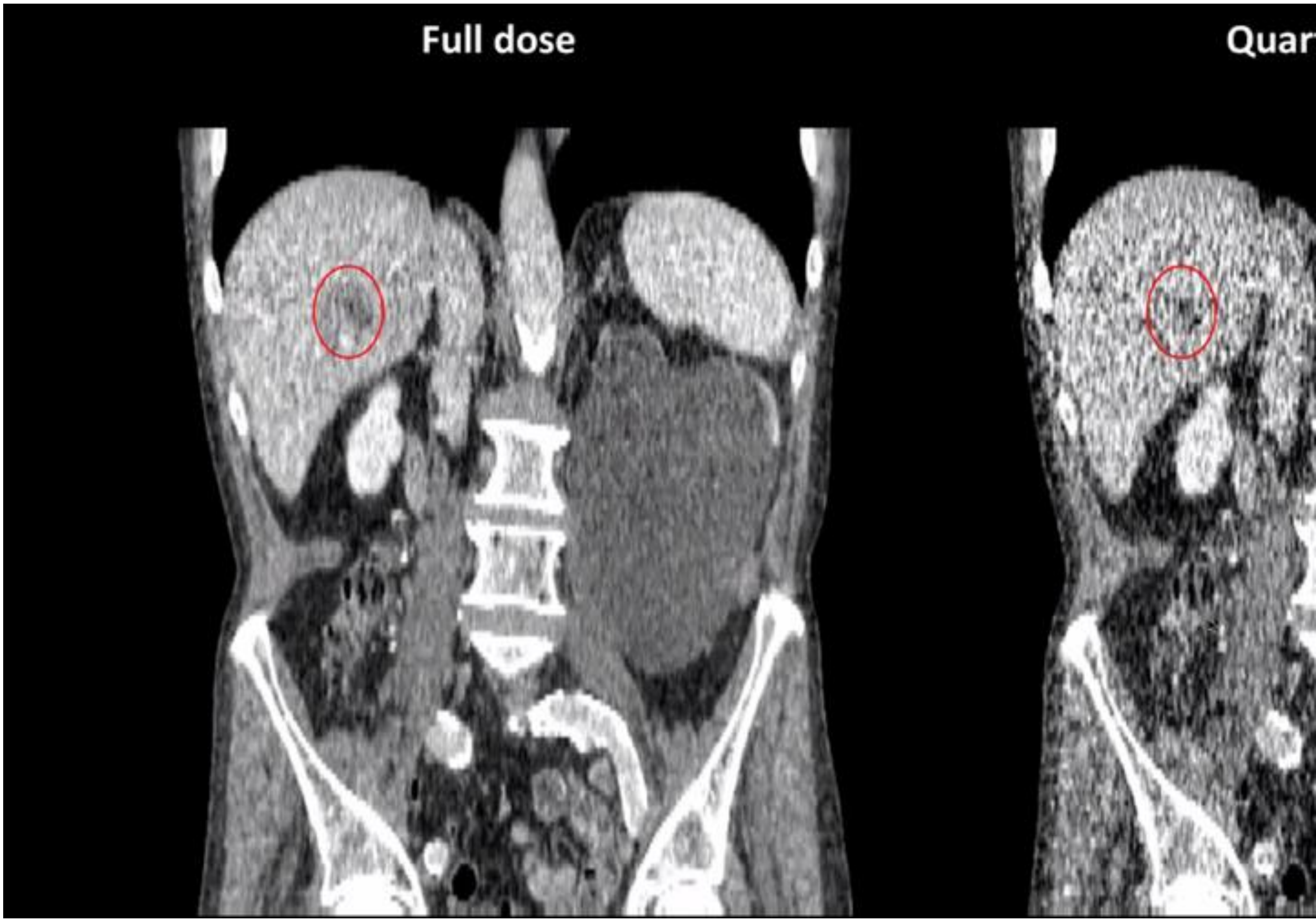


Image Enhancement



AAPM-Net: First deep learning for low-dose CT


High SNR band
Residual learning
: Low-resolution image bypass
Wavelet transform
CNN
level 1 level 2 level 3 level 4
: Conv
: BatchNorm
: ReLU
15
128
Module 1
Module 2
Module 6
Module 7
128 x 8

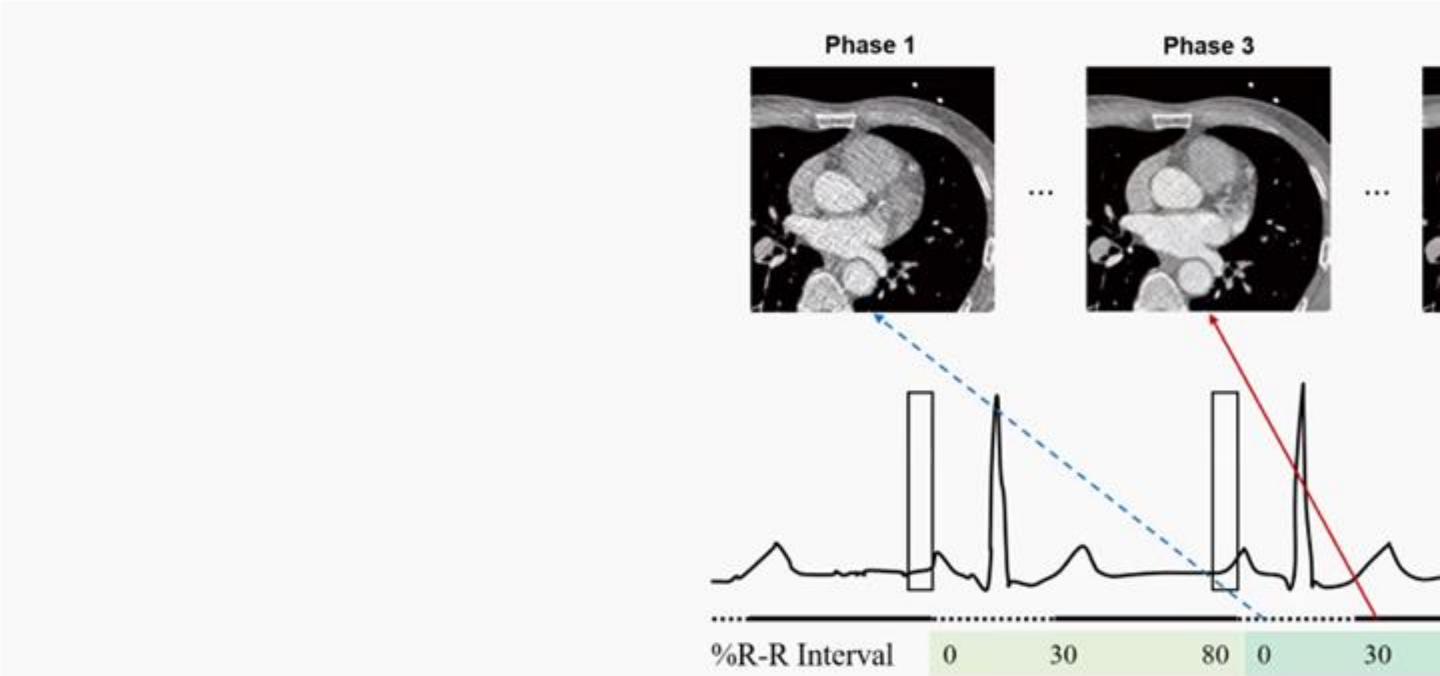


(Kang et al, 2017)

25% of the x-ray dose was used on the right so there is a lot of noise. The network was trained so the metastasis could still be detected.

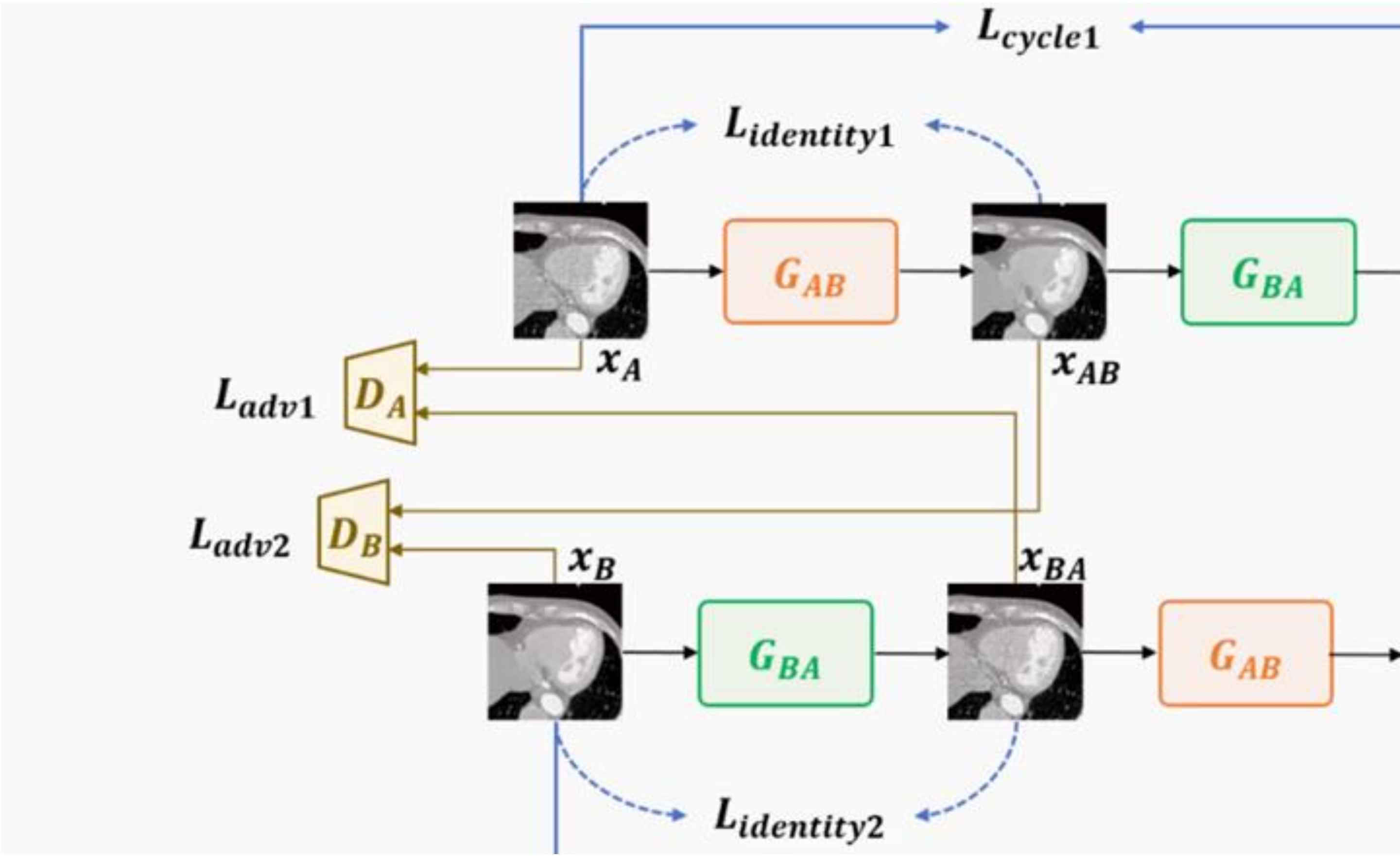
Low-dose CT Denoising without Reference

- Multiphase Cardiac CT denoising
    - Phase 1, 2: low-dose, Phase 3 ~ 10: normal dose
    - Goal: dynamic changes of heart structure
    - **No reference available**



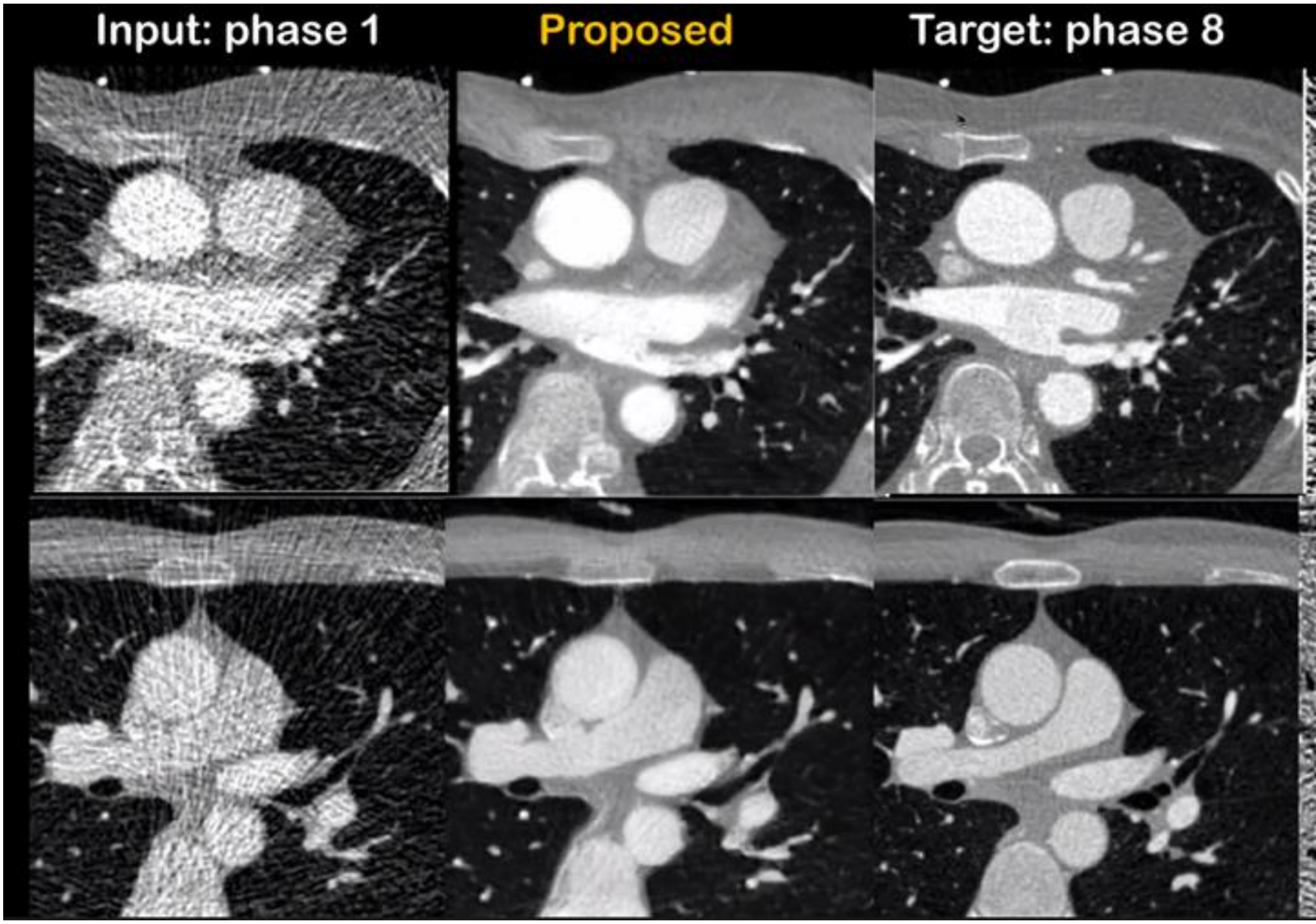
CycleGAN Denoising for Low-Dose CT

(Kang et al, Medical Physics, 2018)



Low dose (5%) -> high dose

Note the noise subtraction & successful image enhancement.



Unmet Needs in MRI

- MR is an essential tool for diagnosis
- MR exam protocol: 30 – 60 min/patient
    - **Should increase the throughput of MR scanning**
- Cardiac imaging, fMRI – **should improve temporal resolution**
- Multiple contrast acquisition in a short time

Accelerated MRI

Deep Learning for Accelerated MRI

Diffusion Models for Accelerated MRI

(H. Chung et al, MEDIA, 2022)

- Imposing **data consistency step** for each iteration
- **Agnostic** to **sampling patterns**
- High frequency **details preserved**
- **Agnostic** to contrast
- **Agnostic** to anatomy

Medical X-VL: Dual Stream VLP

(Park et al., *arXiv preprint arXiv:2208.05140,* 2022)

Future of AI in Medical Imaging: Foundation Models?

Examples of Foundation Models

Multimodal Embedding

(Bommasani, Rishi, et al. *arXiv:2108.07258*, 2021)

Foundation Models for Genomics

(Chen et al, doi: https://doi.org/10.1101/2022.08.06.503062)

Q&A

Are there any rules of thumb for selecting informative positive and negative pairs in contrastive learning? or are we reliant on domain knowledge.

Many people are more interested in non-contrastive learning even though this is just one class.

Can you also comment on progress in generating positive and negative examples for contrastive learning? Is there a work around to the curse of dimensionality?

To make model more robust and generalized, it is an effective way to apply data augmentation to the training set. In scenario of medical images, how can we determine the data augmentation methods and the augmented proportion?

Data augmentation has always been a problem in neural network training. Data augmentation involves adding noise, etc to the training set. The vision transformer method needs to be trained but is less prone to data overfitting. Sample supervised learning is very important in training.

Could this technique be used to recover signal drop-out in MRI?

There is always signal drop-out in an MRI and there are two ways to address this. There may be a signal with dropout and another signal without dropout, so you can do supervised learning or sequence training. This is a more correct way because you are using sequences designed to avoid the signal dropout. The main problem with this is its very time-consuming to utilise signals with sequences to avoid signal dropout. You can also combine those two approaches.

A similar question, can ML be used to detect false positives, for example, in prenatal ultrasound?

Would it be possible to have the list of papers mentioned in the presentation, please?

Do you think ViTs will fully surpass CNNs for image tasks? Or will there be some tasks where CNNs are better?