Medical Imaging with Deep Learning Overview

Popular image problems:

- Chest X-ray
- Histology

Multi-modality/view

Segmentation

Counting

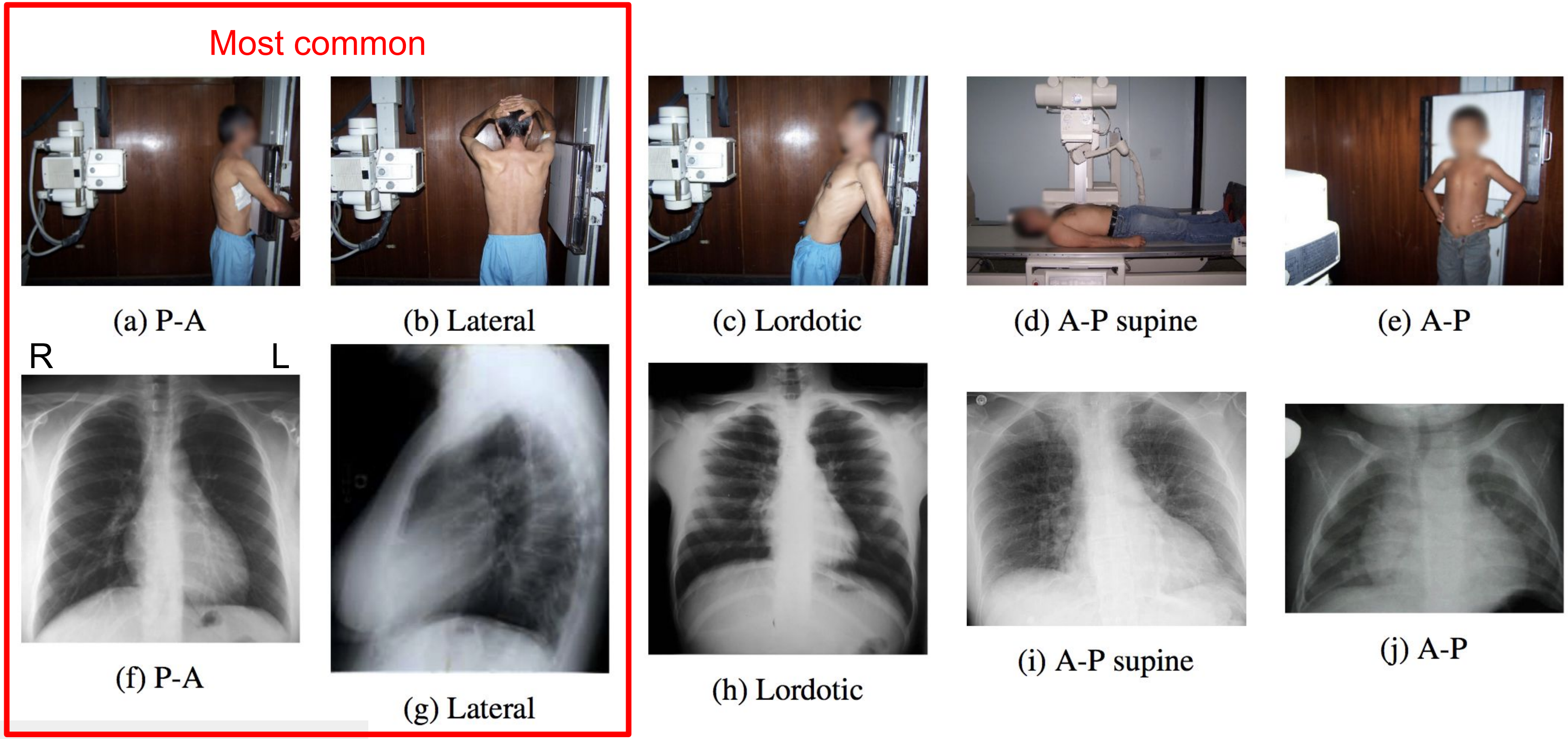
Incorrect feature attribution

Slides by Joseph Paul Cohen 2020

Chapter 1

Radiology and multi-view

Common X-ray projections



PA = PosteroAnterior

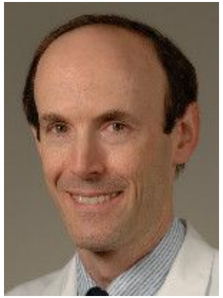
Image: [Bustos, “PadChest: A Large Chest x-Ray Image Dataset with Multi-Label Annotated Reports.” 2019]

Chest X-ray14 Dataset

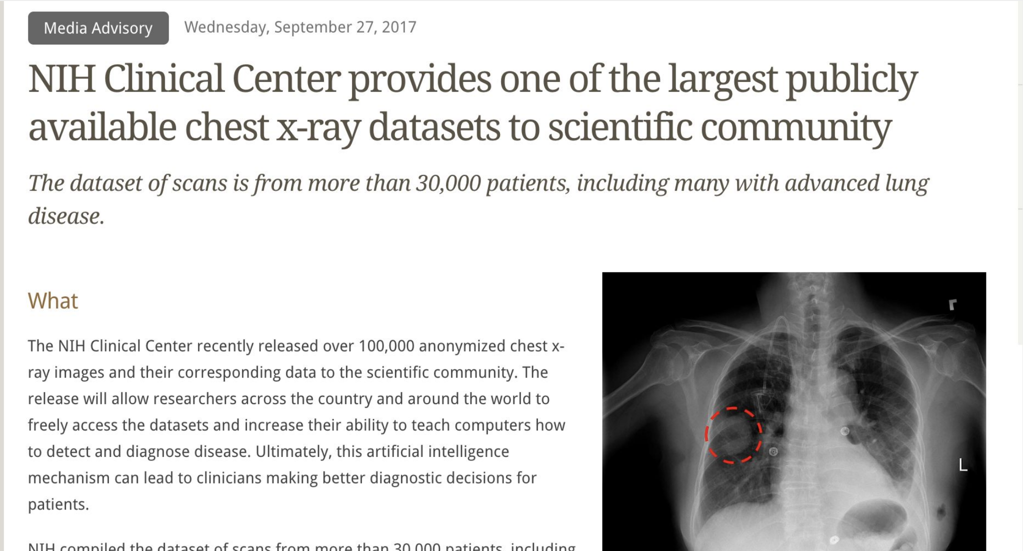
Released 2017, first large scale chest X-ray dataset

>100k PA images released without copyright.

Enabled the deep learning radiology revolution



Ronald Summers
NIH Clinical Center



Media Advisory Wednesday, September 27, 2017

NIH Clinical Center provides one of the largest publicly available chest x-ray datasets to scientific community

The dataset of scans is from more than 30,000 patients, including many with advanced lung disease.

What

The NIH Clinical Center recently released over 100,000 anonymized chest x-ray images and their corresponding data to the scientific community. The release will allow researchers across the country and around the world to freely access the datasets and increase their ability to teach computers how to detect and diagnose disease. Ultimately, this artificial intelligence mechanism can lead to clinicians making better diagnostic decisions for patients.

NIH compiled the dataset of scans from more than 30,000 patients, including

Stanford Pneumonia study

In 2017 Pranav Rajpurkar and Jeremy Irvin trained a DenseNet on NIH data scaled to 224x224 pixels

Set the benchmark performance which has not been significantly improved.

They evaluated pneumonia predictions against 4 radiologists.

"We find that the model exceeds the average radiologist performance on the pneumonia detection task."

	F1 Score (95% CI)
Radiologist 1	0.383 (0.309, 0.453)
Radiologist 2	0.356 (0.282, 0.428)
Radiologist 3	0.365 (0.291, 0.435)
Radiologist 4	0.442 (0.390, 0.492)
Radiologist Avg.	0.387 (0.330, 0.442)
CheXNet	0.435 (0.387, 0.481)

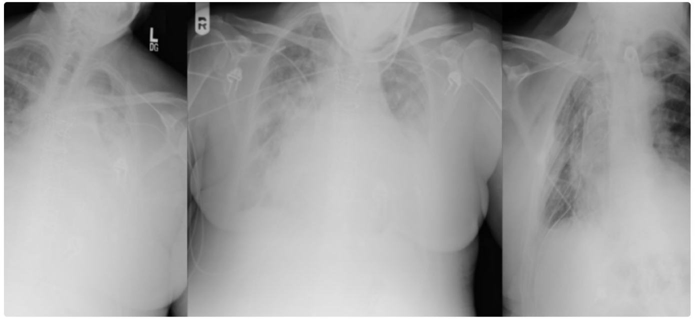
https://stanfordmlgroup.github.io/projects/chexnet/

Criticism of the Chest X-ray14 Dataset

In 2017 Luke Oakden-Rayner published a blog post discussing issues with the labels in the NIH data.

This led to more work on automatic label extraction.

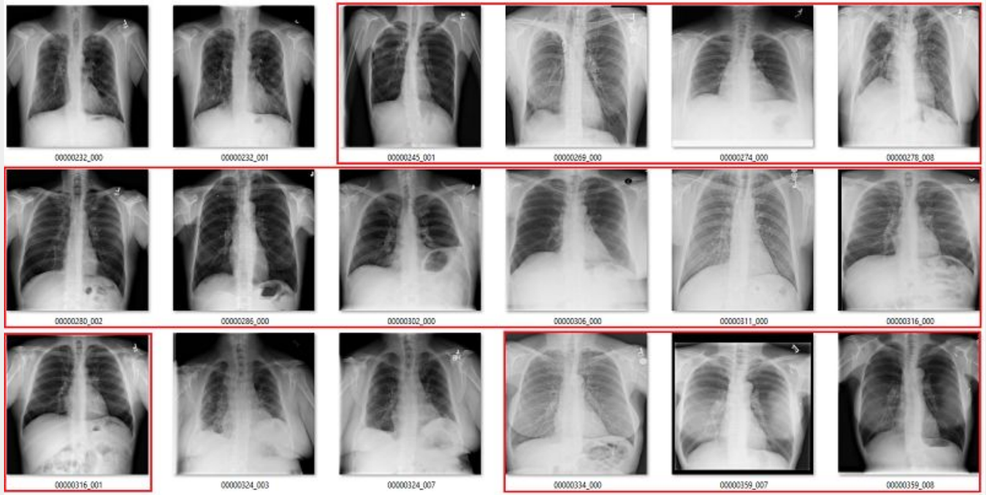
Exploring the ChestXray14 dataset: problems



DECEMBER 18, 2017 ~ LUKEOAKDENRAYNER

A couple of weeks ago, I mentioned I had some concerns about the ChestXray14 dataset. I said I would come back when I had more info, and since then I have been digging into the data. I've talked with Dr Summers via email a few times as well. Unfortunately, this exploration has only increased my concerns about

https://lukeoakdenrayner.wordpress.com/2017/12/18/the-chestxray14-dataset-problems/



In a sample of images red are said to be wrong

2019: the year of X-ray data

Universidad de Alicante

CheXpert

MIMIC

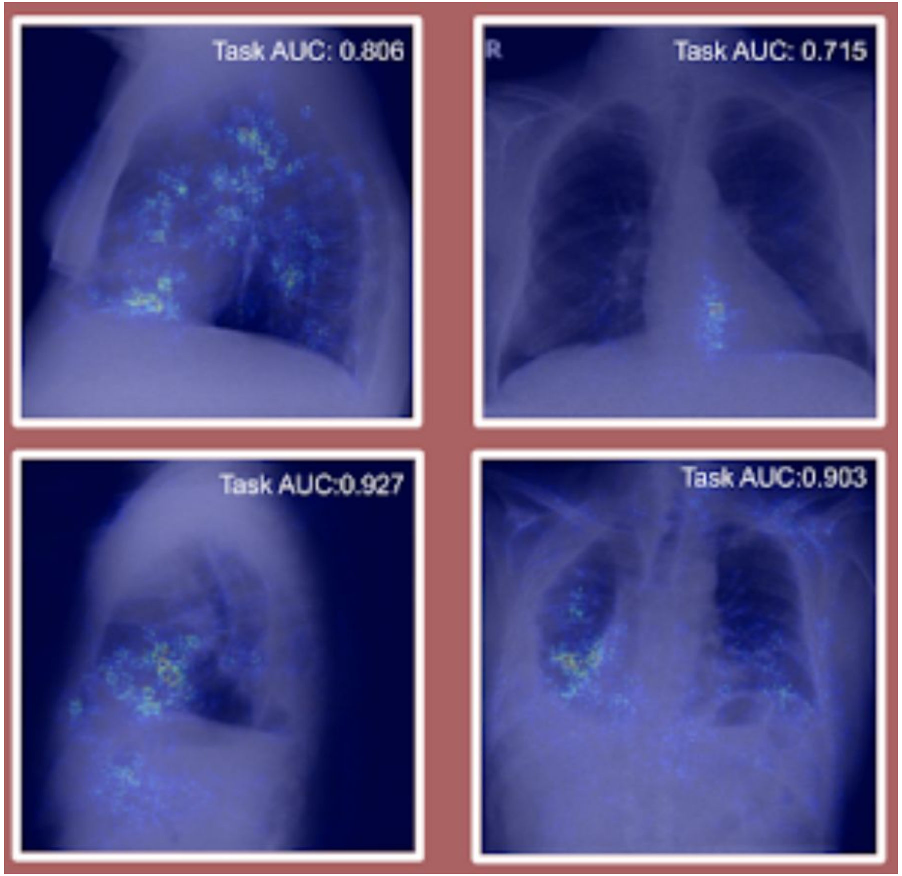
PADCHEST	CheXpert	MIMIC-CXR
160k images	224k images	377k images
Multiple views	PA and L views	PA and L views
Almost 200 labels	13 labels.	13 labels.
27% hand labelled, others using an RNN.	Automated rule-based labeler	Automated rule-based labeler. NIH (NegBio) and CheX labelers ran.
License:Creative Commons Attribution-ShareAlike	Non-commercial research purposes only	Non-commercial research purposes only. Confidentially training required.

Multi-modal/view inference (X-ray use case)

Lateral PA

Flattened diaphragm

Pleural effusion

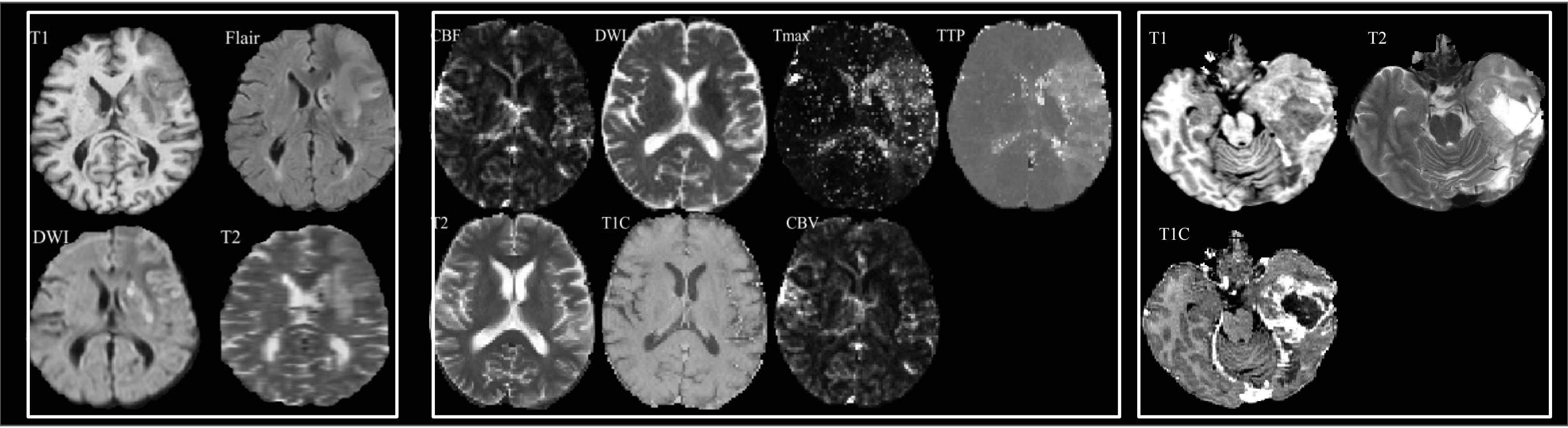


Here saliency maps are from models trained on single views.

These two tasks perform better when using lateral views.

[Bertrand, 2019]

Also: Multi-modal/view inference (MRI use case)

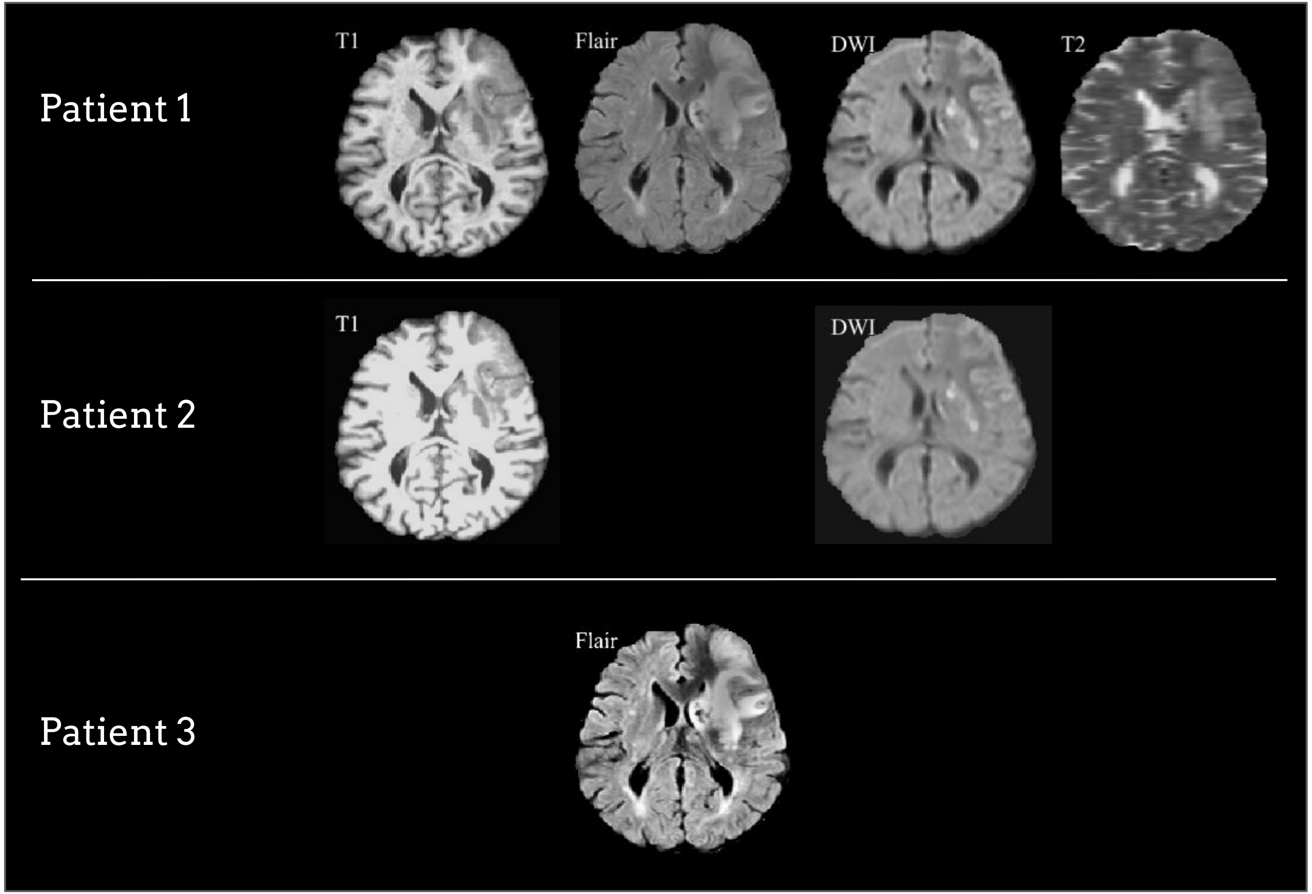


Ischemic stroke lesion segmentation

Stroke Perfusion Estimation

Brain tumor segmentation

Challenge: missing modalities/views



Incomplete Input!

Expected:

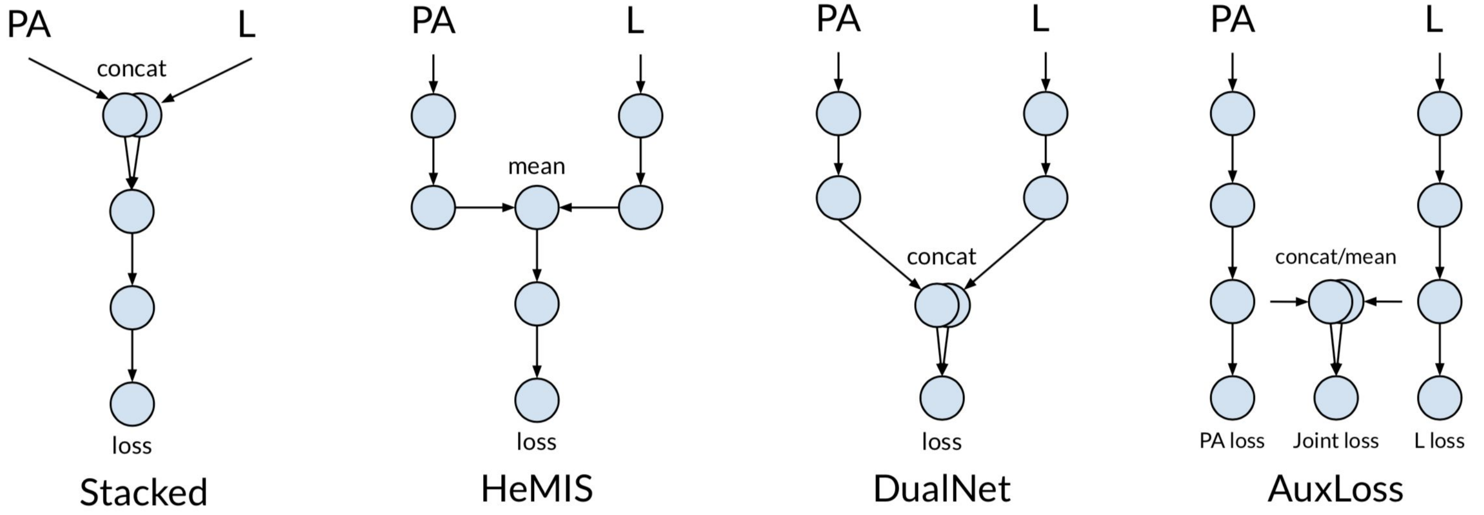
$$f(\mathrm{T1},\ \mathrm{Flair},\ \mathrm{DWI},\ \mathrm{T2})$$

Given:

$$f(\mathrm{T1},\ -\ ,\ \mathrm{DWI},\ -\ )$$

$$f(\ -\ ,\ \mathrm{Flair},\ -\ ,\ -)$$

Integrating multiple views



Combine images right at the input

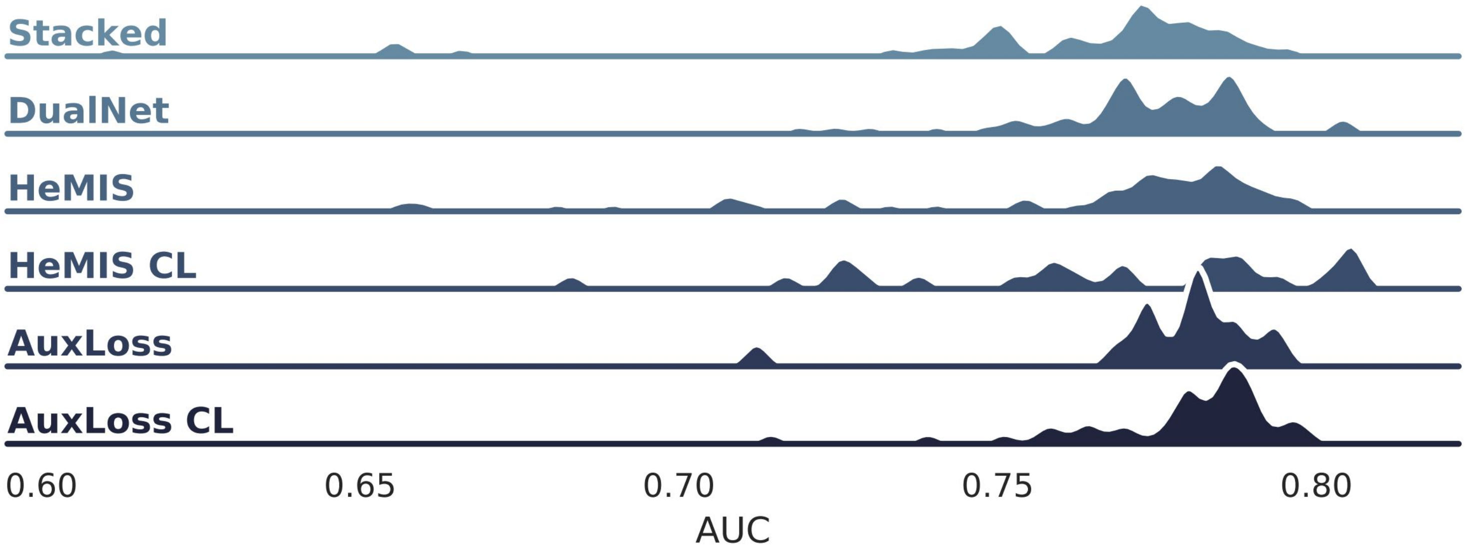
Take mean of activations in the middle of the network

Concat output features of two models with single prediction

Three losses. A network for each modality with losses that regularize each network.

Image: [Hashir, Quantifying the Value of Lateral Views in Deep Learning for Chest X-rays, 2020]

Integrating multiple views (X-ray images)



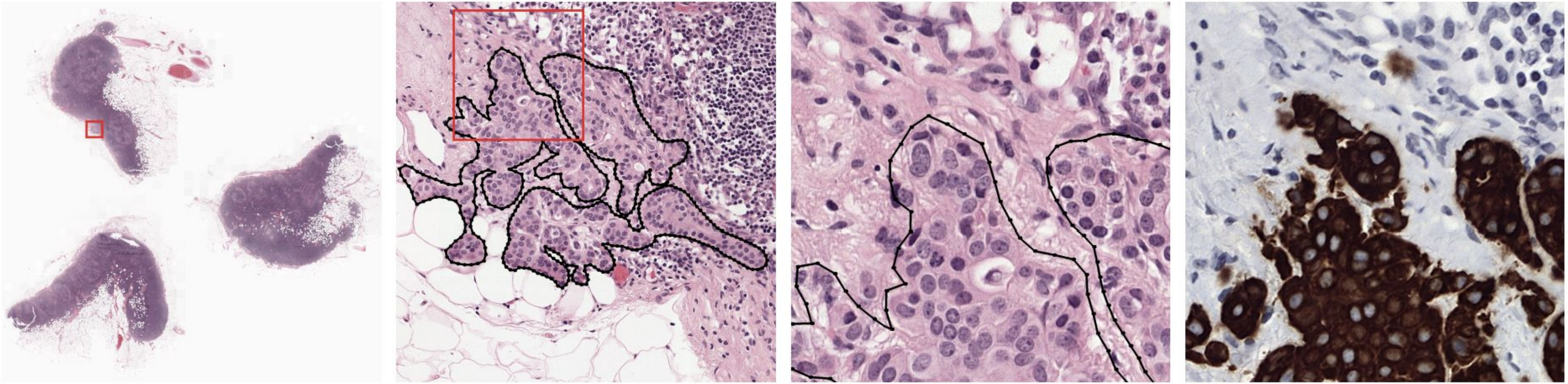
All models are about equal in performance given the right hyperparameters.
Hyperparameter tuning is easier on some models but not others

Image: [Hashir, Quantifying the Value of Lateral Views in Deep Learning for Chest X-rays, 2020]

Chapter 2

Histology and segmentation

CAMELYON17: A large high resolution open histology dataset for cancer detection



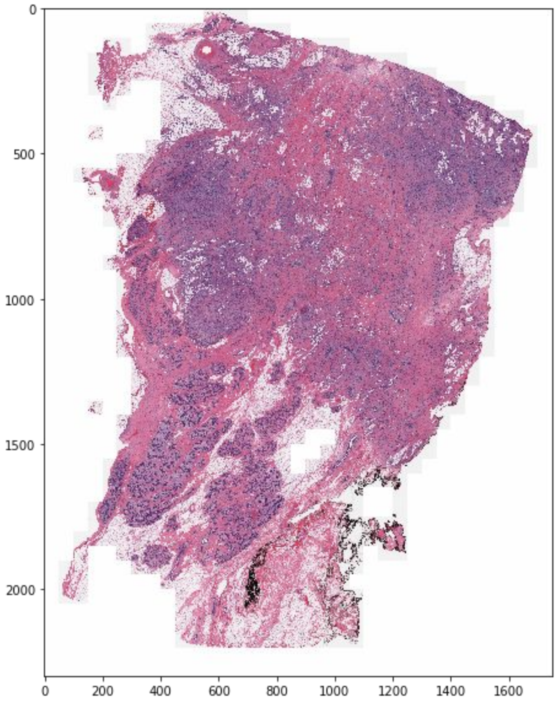
Example of a WSI of a H&E stained section with a delineated micro-metastasis at increasing zoom levels, and the corresponding IHC (cytokeratin 8-18 stained) slide at the same location. The metastasis is outlined with black.

CAMELYON17 Dataset
1000 whole-slide images (WSIs) of sentinel lymph node. (~3GB each!)
5 medical centers. 40 patients from each center. 5 whole-slide images per patient.

Peter Bandi, et al. From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge. IEEE-TMI 2018

Patch wise segmentation

Use case: Invasive Ductal Carcinoma (most common subtype of all breast cancers)



Starting with a full slide image of breast tissue.

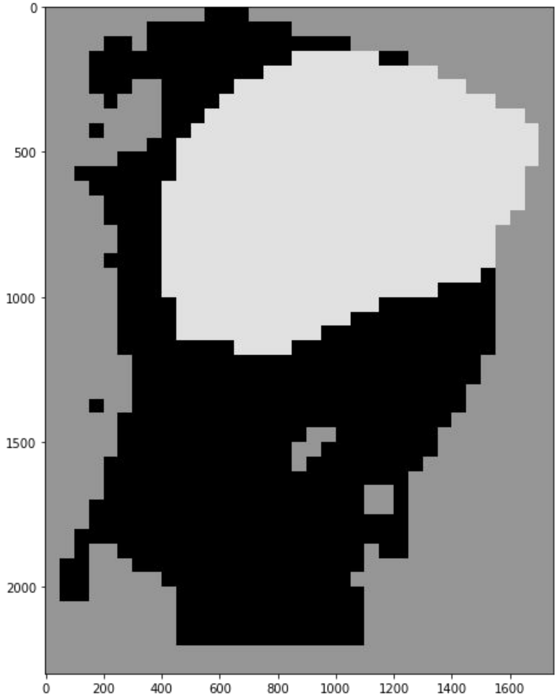


Image is labelled as IDC or not

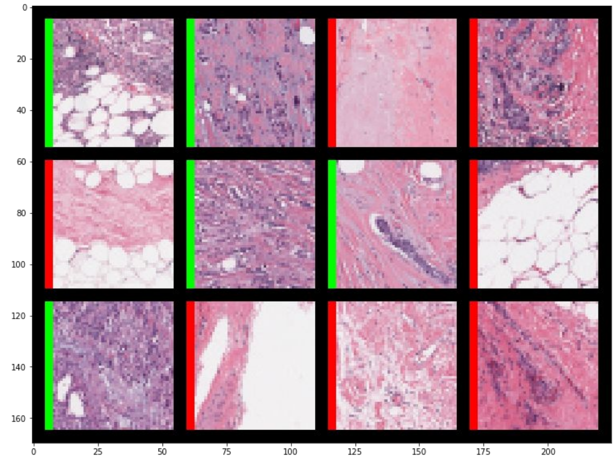
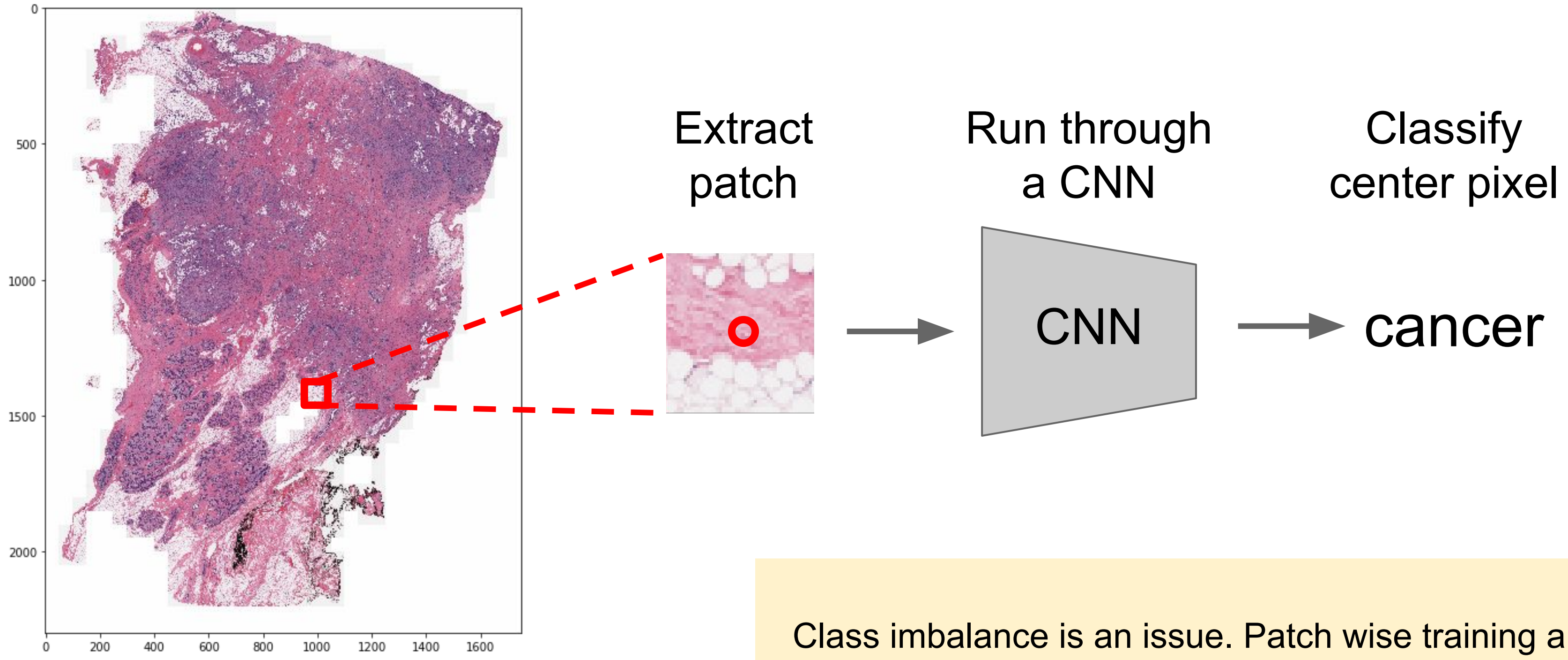


Image is chopped into patches and labelled as IDC or not

https://colab.research.google.com/drive/13T9s3weexAw6YskKoY6c-VvoUgUvWsqf

Patch wise segmentation

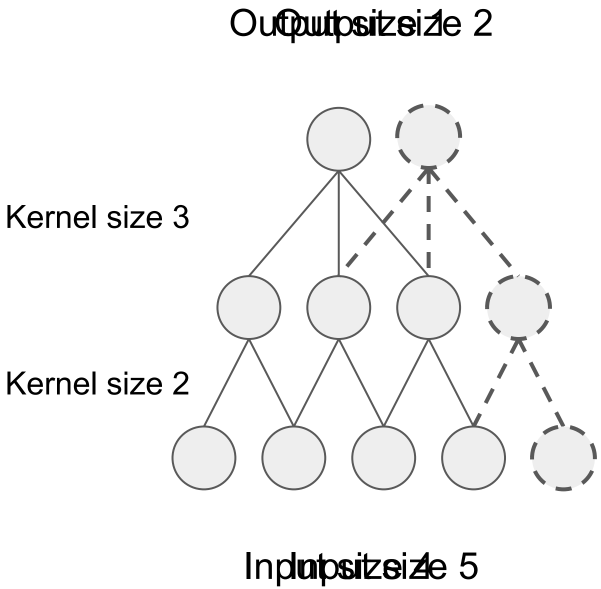
Use case: Invasive Ductal Carcinoma (most common subtype of all breast cancers)



Class imbalance is an issue. Patch wise training allows easy balancing of classes using standard methods.

Slide design: Fei-Fei Li & Andrej Karpathy & Justin Johnson

Fully convolutional processing (FCN)



What is this model's receptive field? 4 nodes

How many multiplications were saved? 4

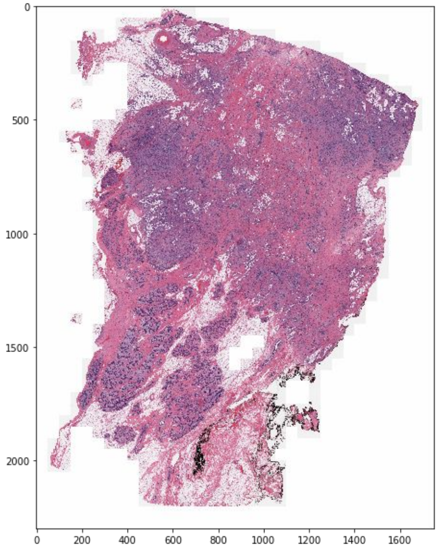
How many saved for an input size of 6? 8

Allows for very fast inference.

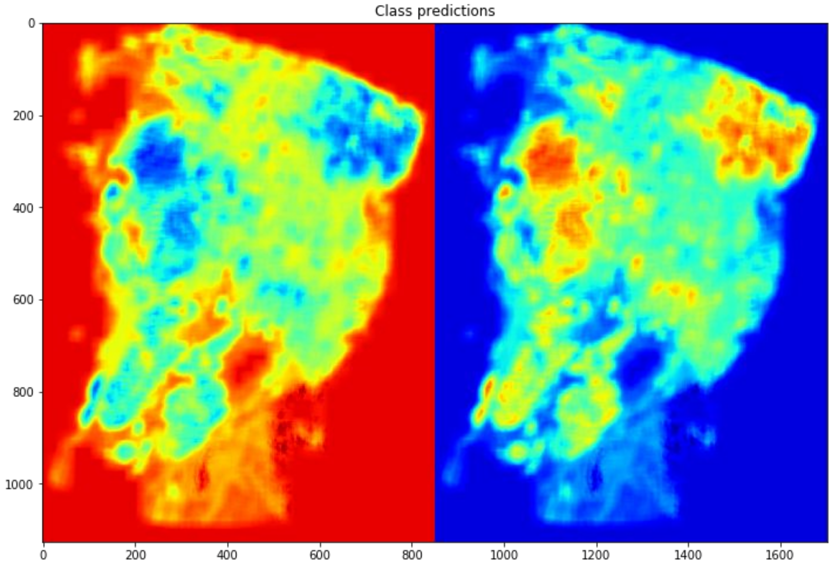
However, training this way requires a lot of memory. Need to save past outputs.

Patch wise training together with FCN inference is a good balance.

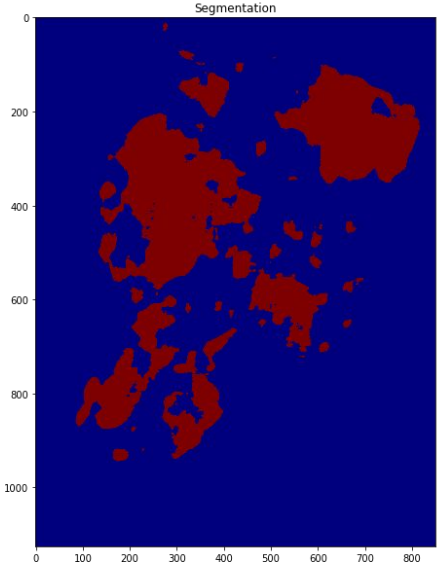
Input image



Output class 0 Output class 1

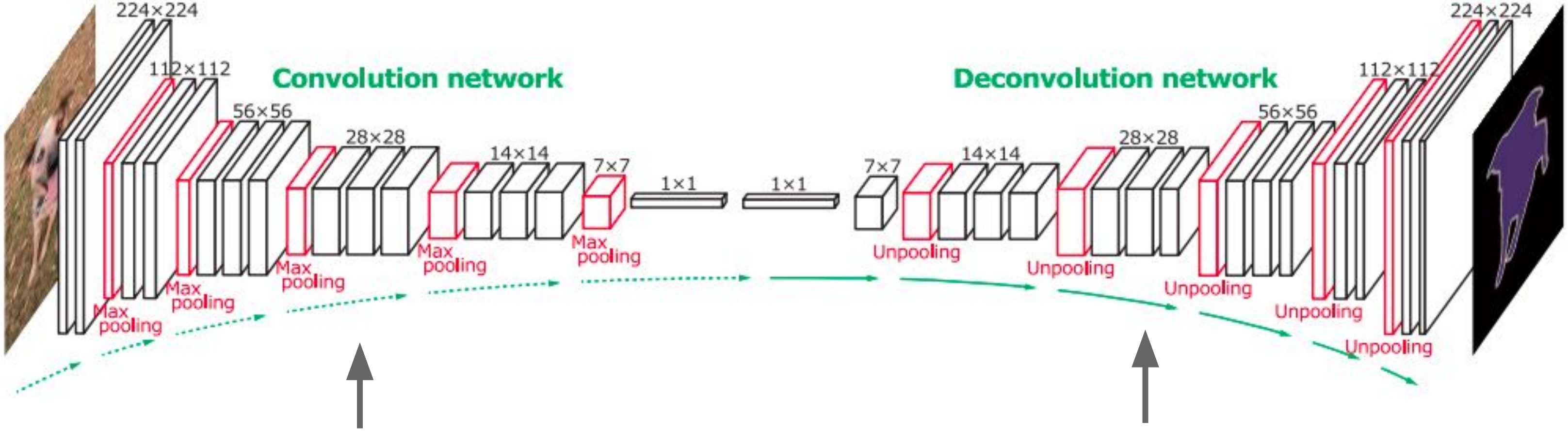


class 1 > class 0



https://colab.research.google.com/drive/13T9s3weexAw6YskKoY6c-VvoUgUvWsqf

Recap: Segmentation using a bottleneck



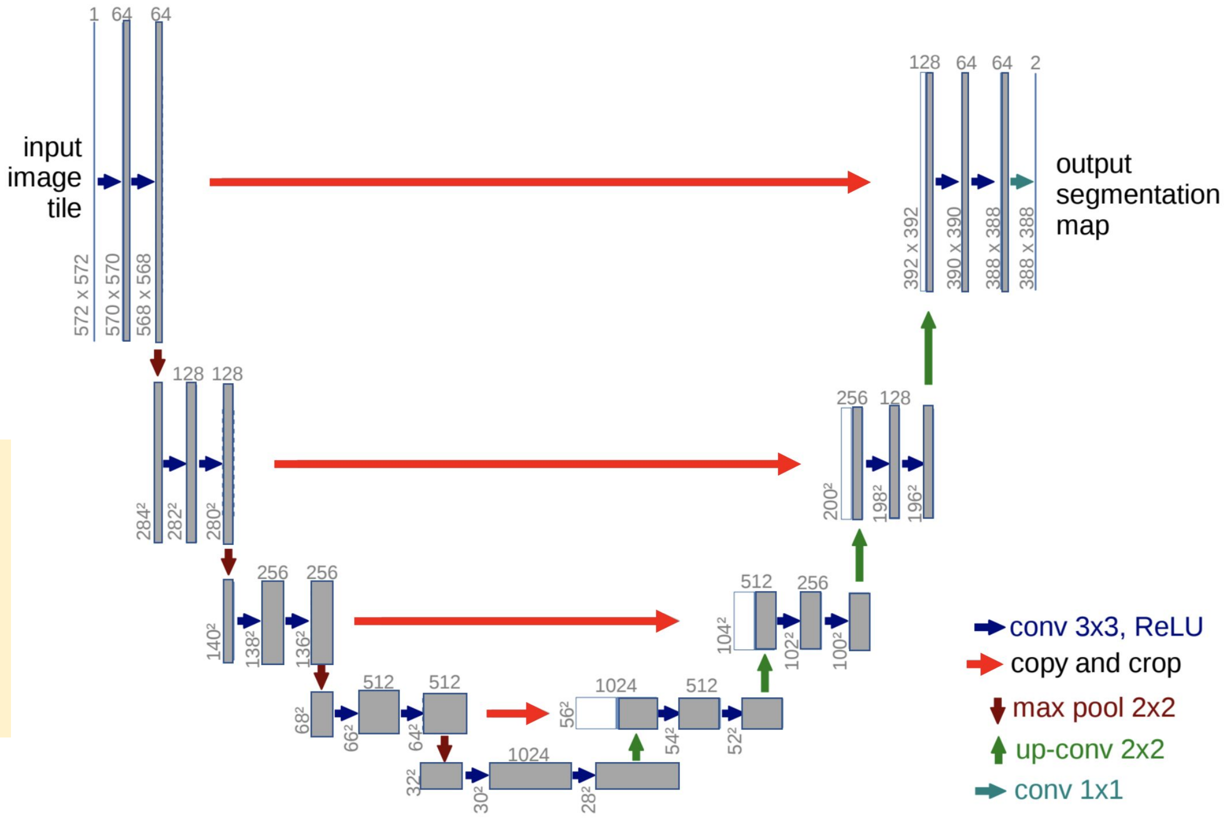
Normal VGG

"Upside down" VGG

Upsampling possible with
- Unpooling
- Transposed convolutions

Noh et al, "Learning Deconvolution Network for Semantic Segmentation", ICCV 2015
Slide design: Fei-Fei Li & Andrej Karpathy & Justin Johnson

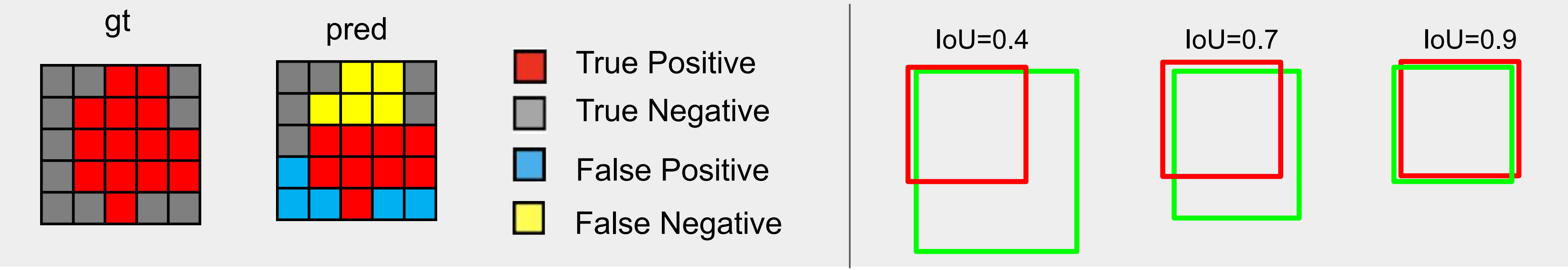
Recap: U-NET



Difference:
Skip connections (like resnet)

Dogma: skips carry spatial information, bottleneck carries high level structure.

Segmentation metrics



$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{IoU} = \text{Jaccard Index} = \frac{TP}{TP + FN + FP} = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

$$\text{Dice Coefficient} = \frac{2 \cdot TP}{(2 \cdot TP) + FN + FP} = \frac{2|X \cap Y|}{|X| + |Y|}$$

Training with dice

What maximizes the numerator?

Using the dot product to compute the intersection allows for a differentiable loss.

$$\mathrm{DL}(p, \hat{p}) = \frac{2\sum_i p_i \hat{p}_i}{\sum_i (p_i + \hat{p}_i)}$$

For multiple classes a basic approach is to average over all classes

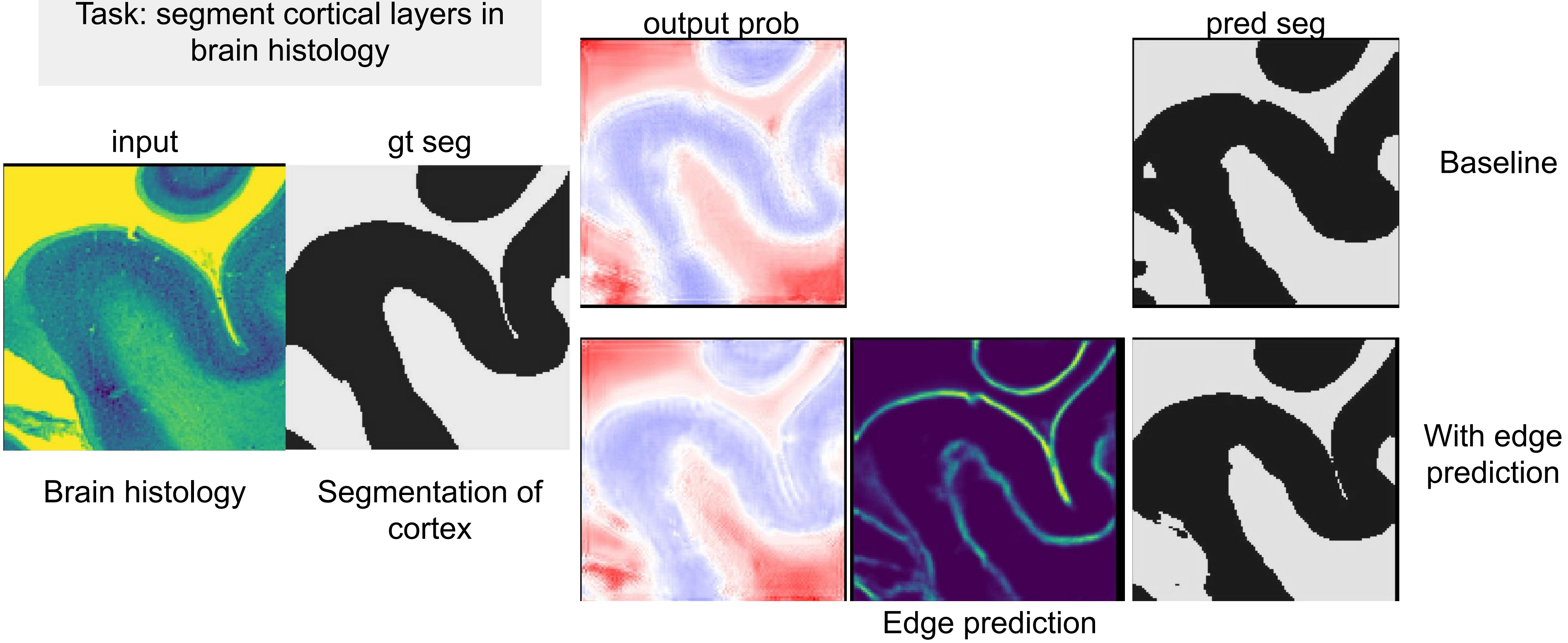
$$\mathrm{DL}_{mean}(p, \hat{p}) = \frac{1}{|C|} \sum_{c \in C} \frac{2\sum_i p_i^c \hat{p}_i^c}{\sum_i (p_i^c + \hat{p}_i^c)}$$

Use a sigmoid or a softmax to restrict output.

More reading: https://arxiv.org/abs/1707.00478

Tricks: Improving edges in segmentations by predicting edges

Task: segment cortical layers in brain histology



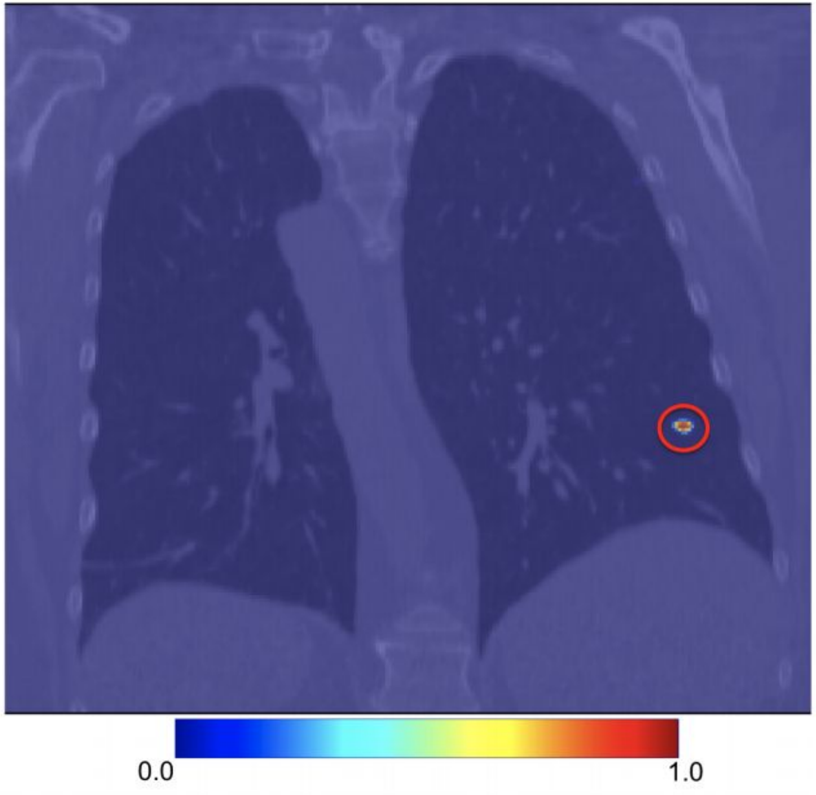
Images provided by Konrad Wagstyl (University College London) 2020

More reading about idea: [Polzounov, WordFence: Text Detection in Natural Images with Border Awareness, 2017]

Challenge: extreme class imbalance (e.g. lung nodule)

Background classes can dominates the loss and cause learning instability do to large gradients.

Balanced sampling may not work as well because patches which could yield false positives are rarely seen to train on.



CASED importance sampling for large images

General Idea:

Store a probability for each patch.

Generate patches based on this probability.

Probability is inverse of how well your model performs on that patch.

Samples are stratified by class.

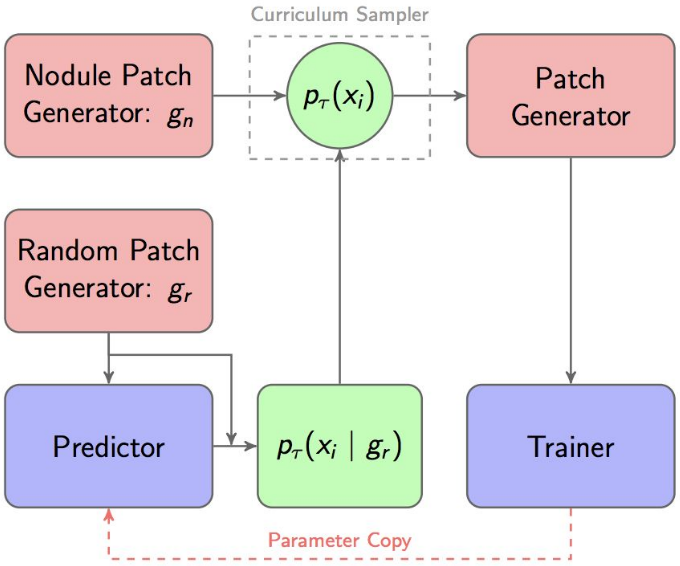


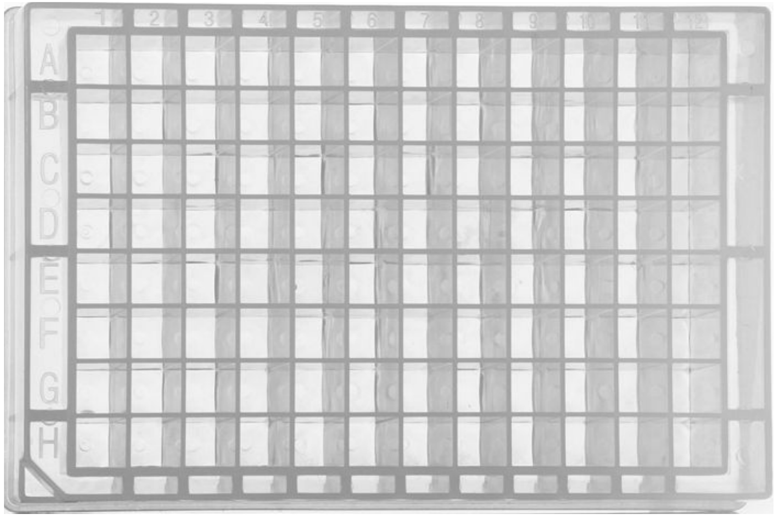
Fig. 1. Schematic diagram of CASED framework

[Jesson, https://arxiv.org/abs/1807.10819]

Chapter 3

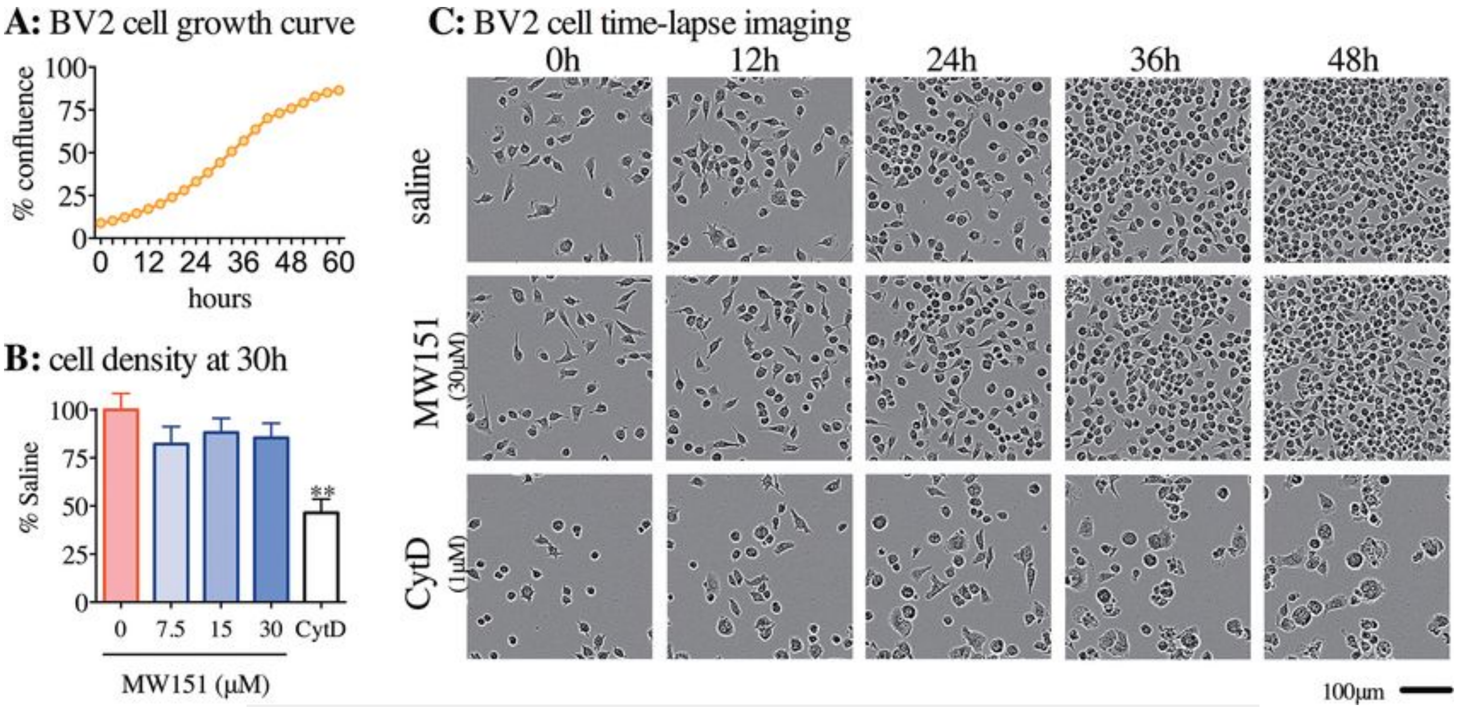
Counting

Use case: Proliferation/Cell growth studies



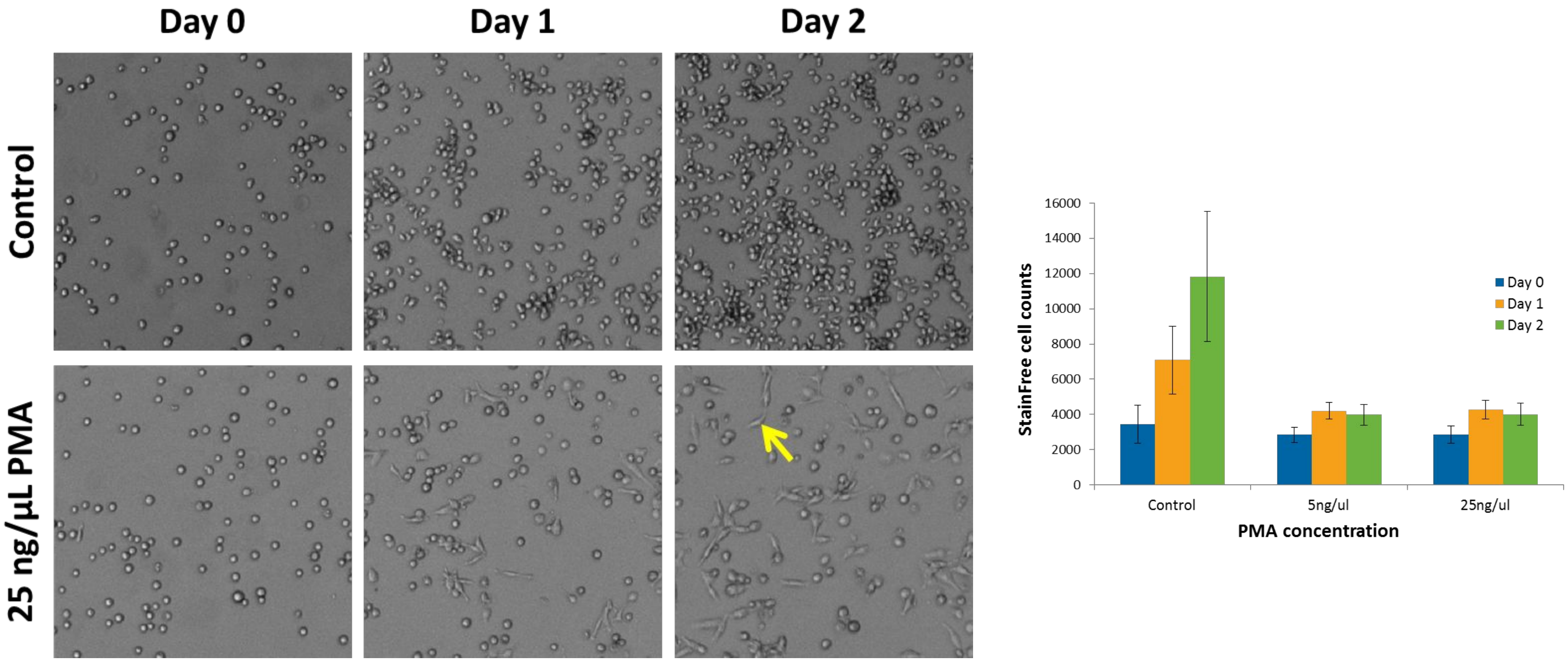
Standard 96-well plate

Treat cells with different compounds and observe proliferation over time



Bachstetter, MW151 Inhibited IL-1? Levels after Traumatic Brain Injury with No Effect on Microglia Physiological Responses, PLOS ONE, 2017

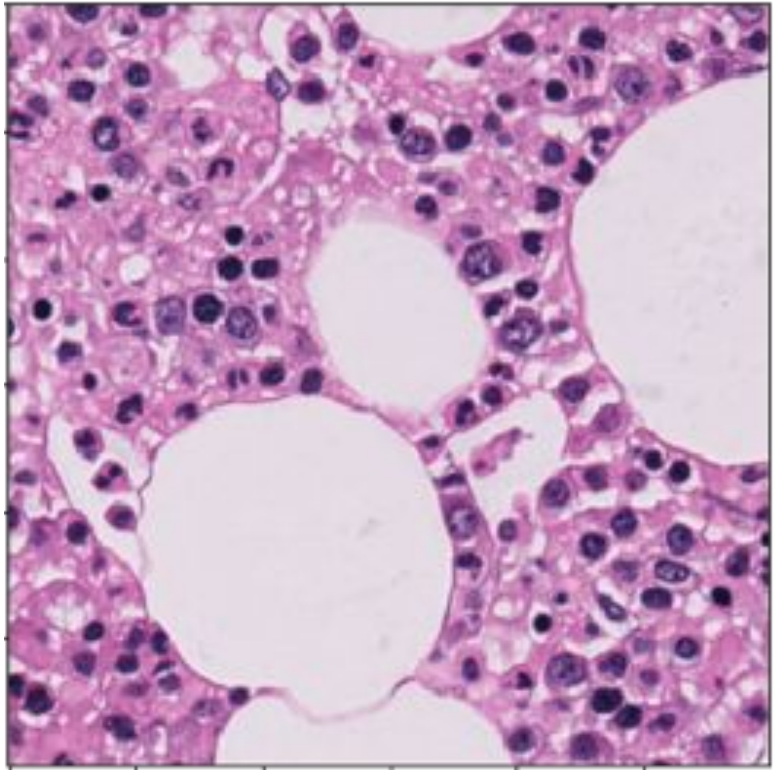
Use case: Proliferation/Cell growth studies

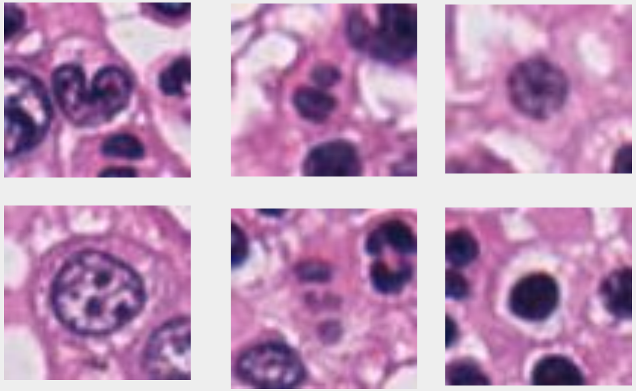


At the Cell Counter: THP-1 Cells, Molecular Devices
https://www.moleculardevices.com/cell-counter-thp-1-cells

Use case: Proliferation/Cell growth studies

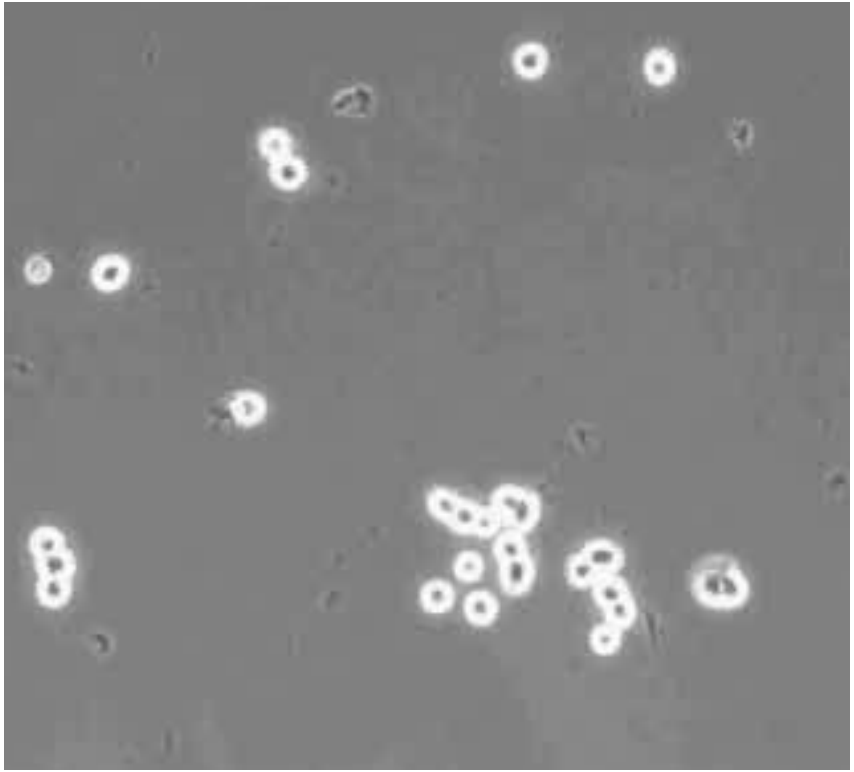
Use case: Counting in histology slides





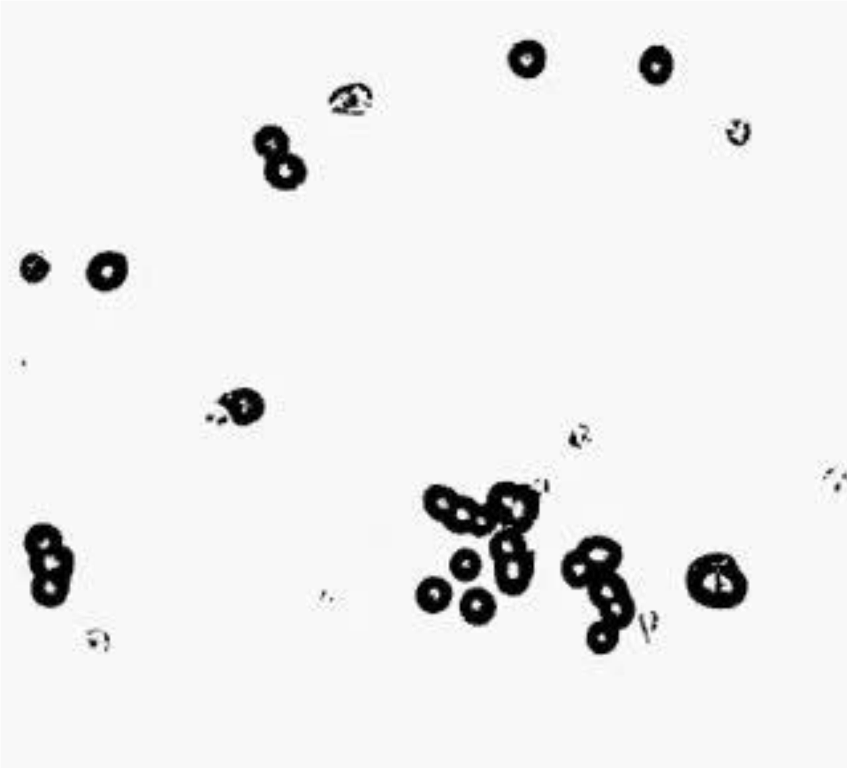
Complicated cell structure

Cell counting (classic CV)



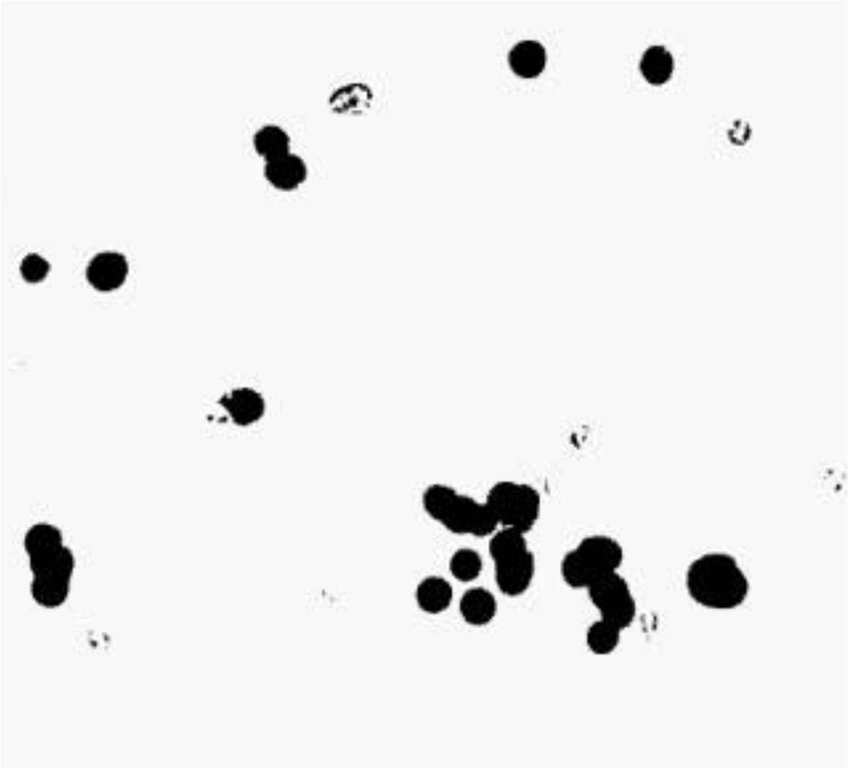
1. Create binary segmentation image
2. Watershed segmentation
3. Isolate and count

Cell counting (classic CV)



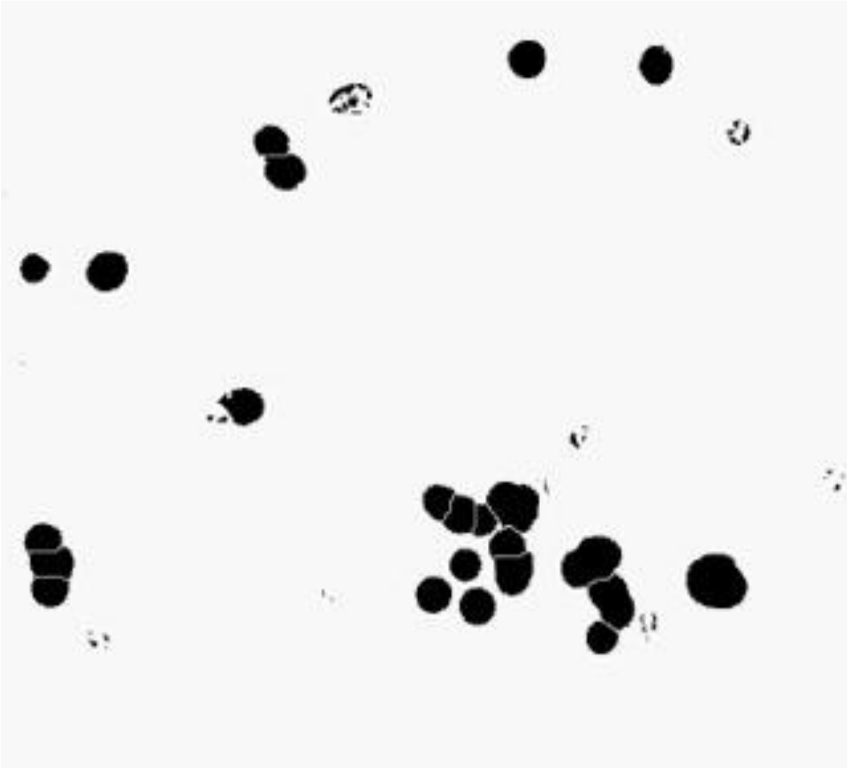
1. **Create binary segmentation image**
2. Watershed segmentation
3. Isolate and count

Cell counting (classic CV)



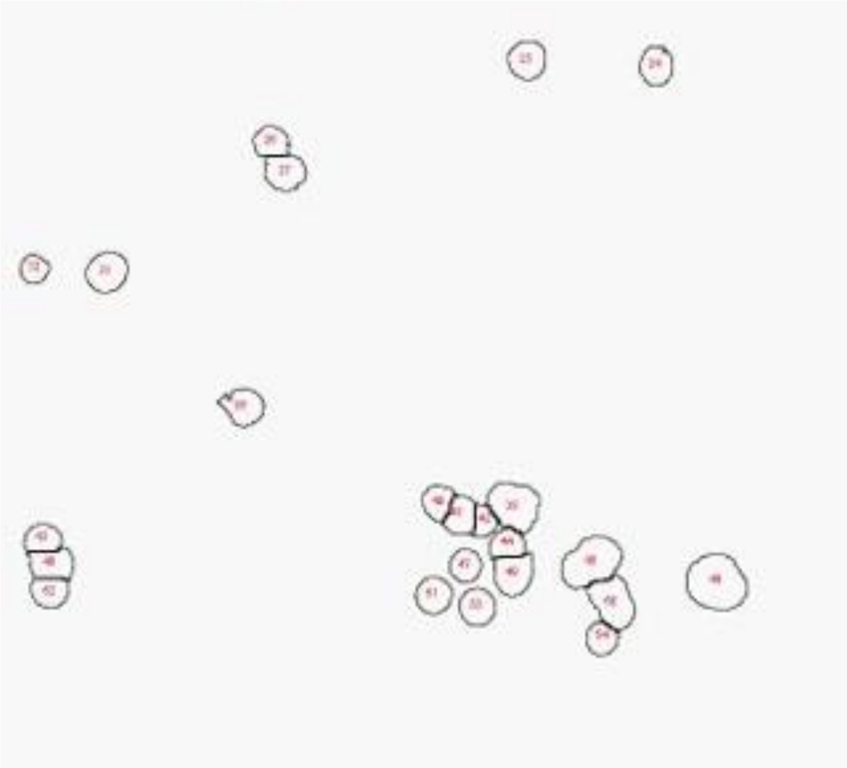
1. **Create binary segmentation image**
2. Watershed segmentation
3. Isolate and count

Cell counting (classic CV)



1. Create binary segmentation image
2. **Watershed segmentation**
3. Isolate and count

Cell counting (classic CV)

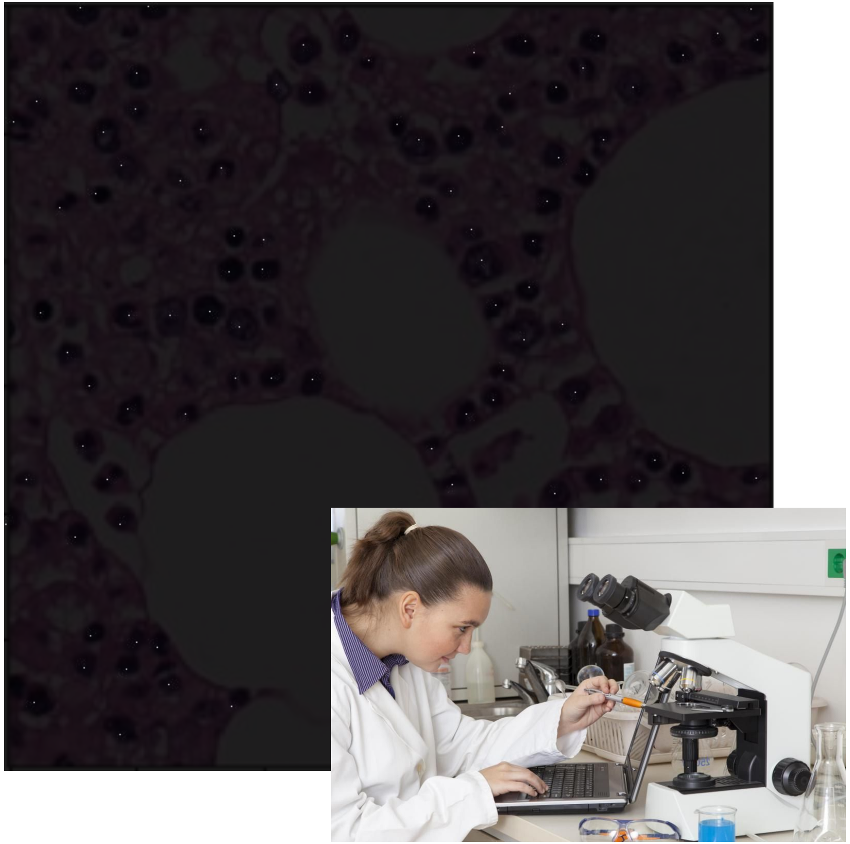


1. Create binary segmentation image
2. Watershed segmentation
3. **Isolate and count**

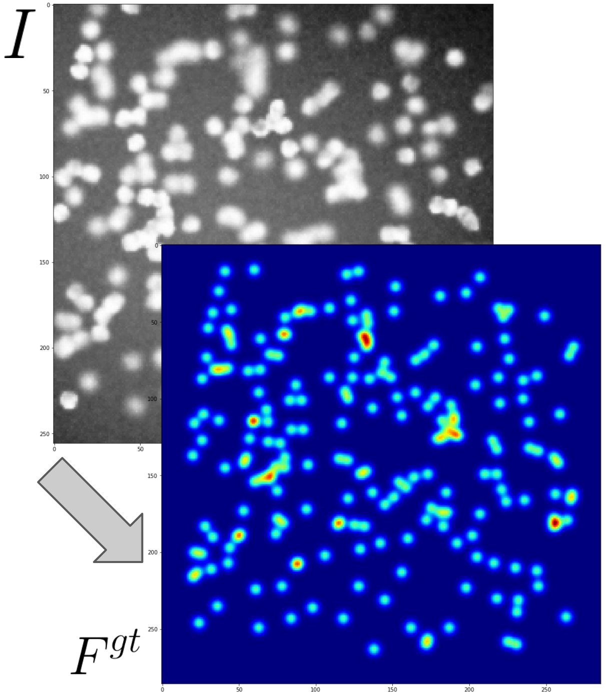
This works well on easy tasks but doesn't scale.

"Pipelines" end up breaking on new images with different lighting or stain.

How to get labels?



Counting via Segmentation



Sigma is typically small like a few pixels

Targets for regression

$$F^{gt}(x,y) = \sum_{i}^{\#\ \text{Cells}} \mathcal{N}([x,y];[x_i,y_i],\sigma^2)$$

Train model to regress

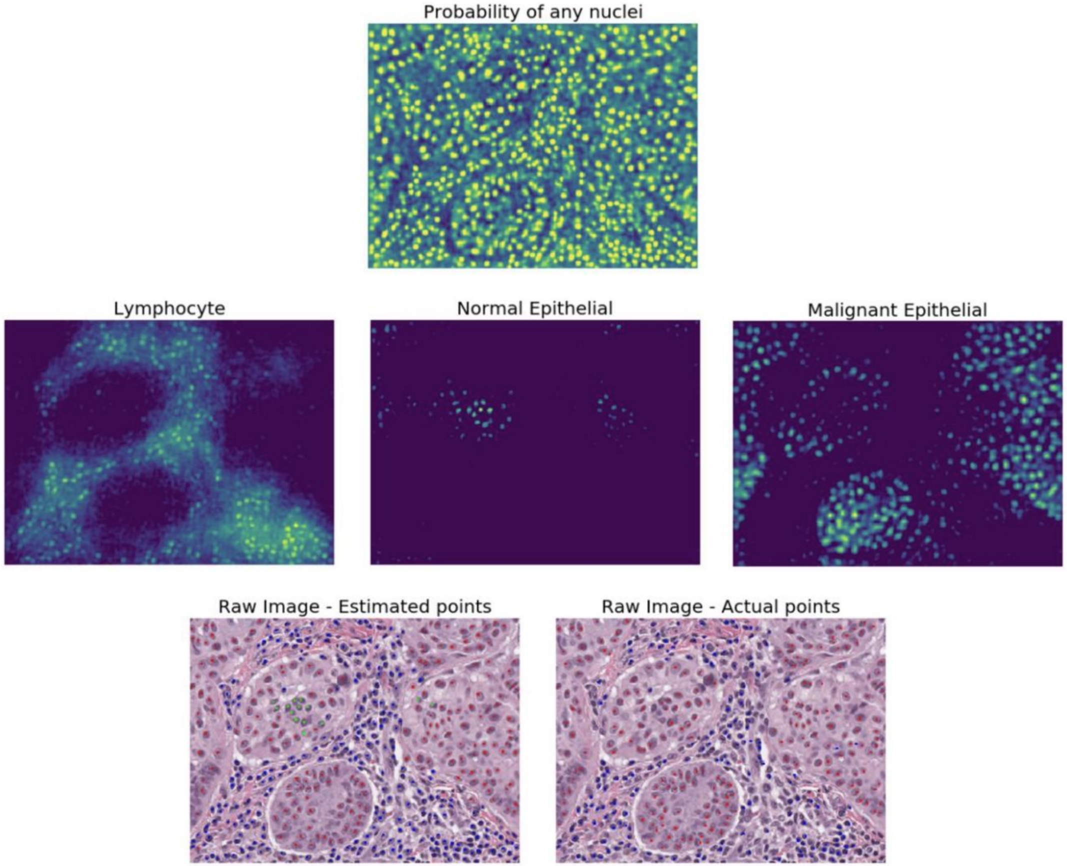
$$L = \sum_{x,y} |F^{gt}(x,y) - F(I)(x,y)|$$

To recover count:

$$\text{count} = \sum_{x,y} F(I)(x,y)$$

V. Lempitsky and A. Zisserman, “Learning To Count Objects in Images,” **2010**.

Multiple output classes



Count and classify different cell types [Bidart 2018]

Counting and classifying also possible using multiple output channels.

Combine losses together

$$\lambda_1 L_{lymph} + \lambda_2 L_{norm} + \lambda_3 L_{mal}$$

Max prediction over output channels for each cell identified

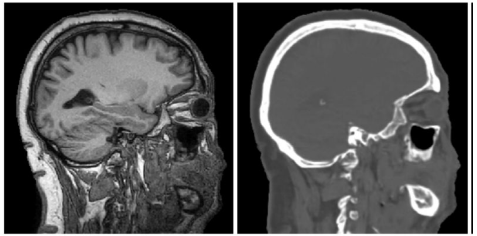
Chapter 4

GANs

Medical image-to-image translation considered harmful

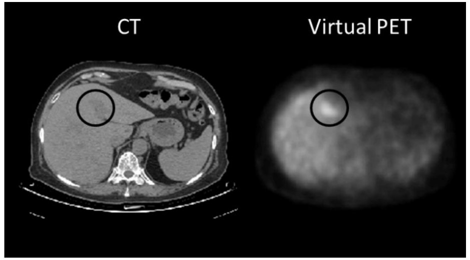
Many papers have proposed methods that can "translate between modalities"

MR -> CT

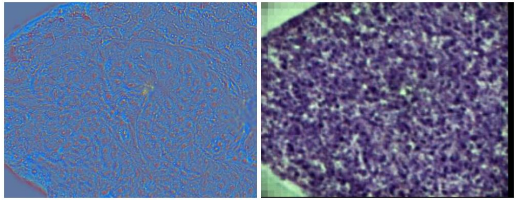


I_{MR} $Syn_{CT}(I_{MR})$

CT -> PET

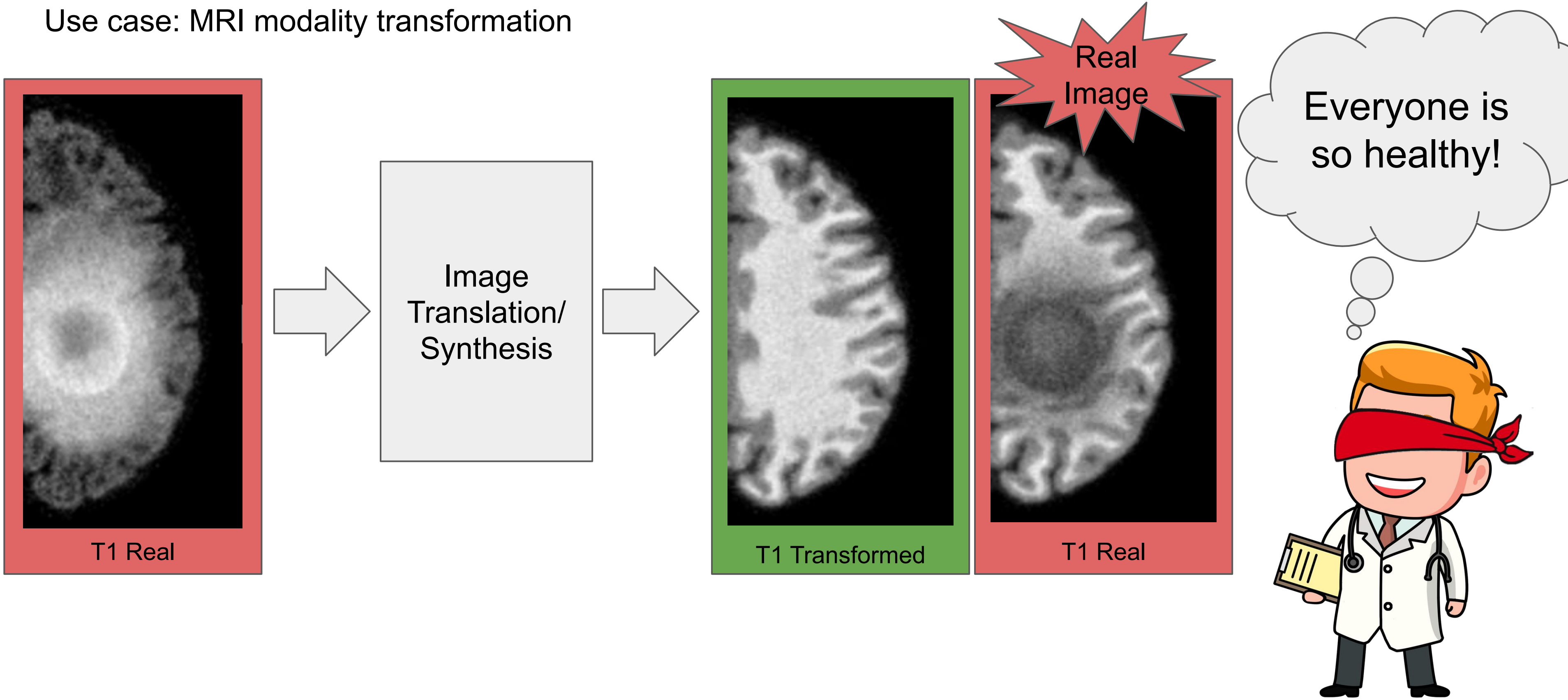


Synthesized H&E staining



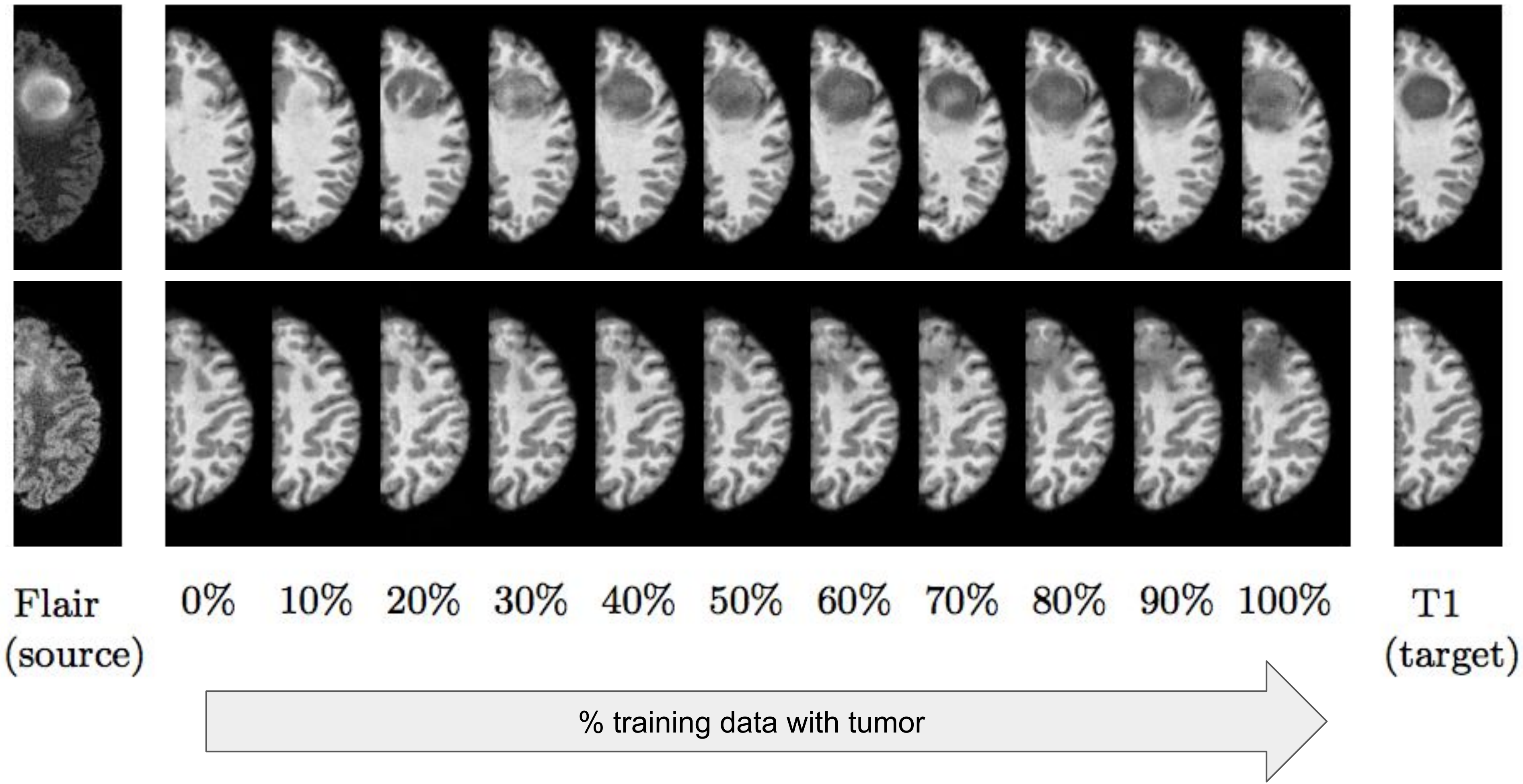
Adversarial losses are very good at distribution matching (e.g. CycleGAN).
But artifacts could be introduced and then used in diagnosis which can be dangerous.

Use case: MRI modality transformation

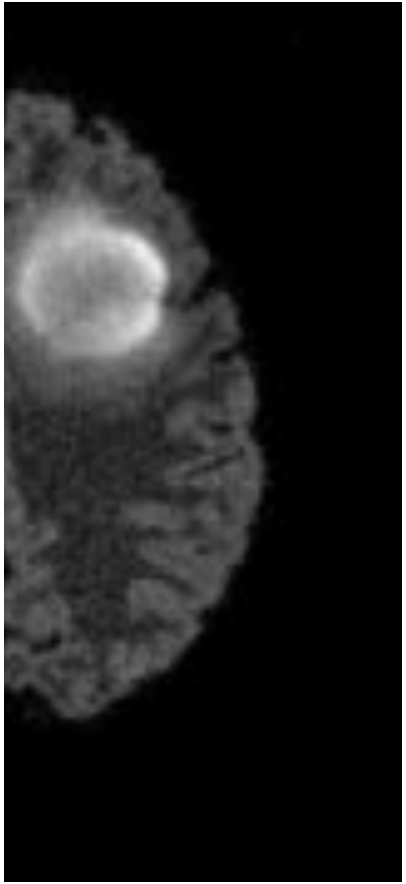


But a bias in training data can lead to incorrect translation

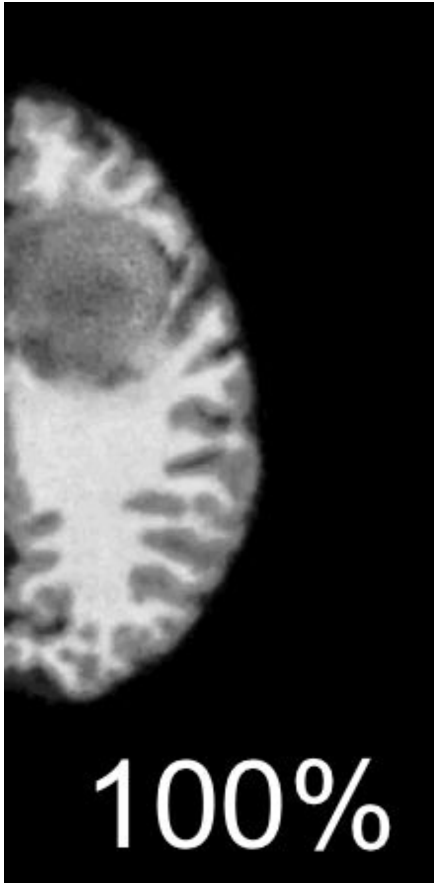
CycleGAN results



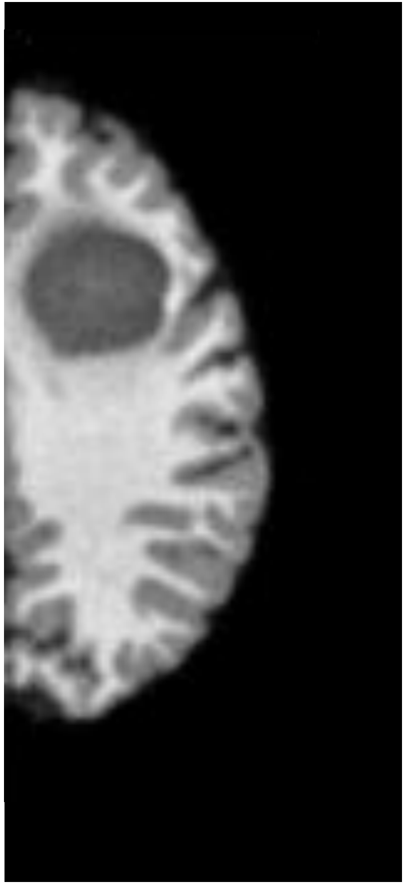
Real Flair



Biased Transformations



Real T1



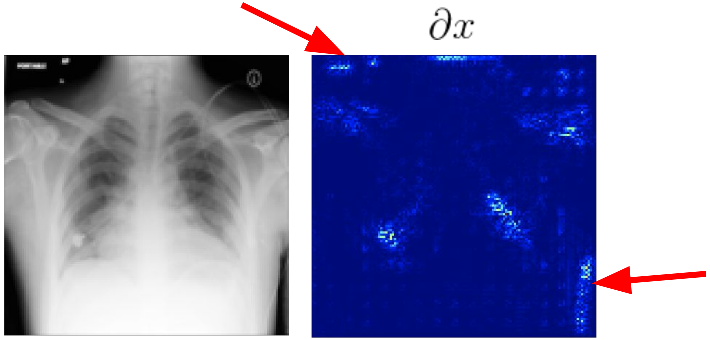
Chapter 5

Right for the right reasons

Incorrect feature attribution

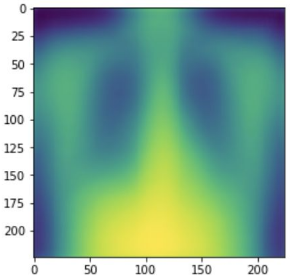
Models can overfit to confounding variables in the data.

$$\frac{\partial f}{\partial x}$$

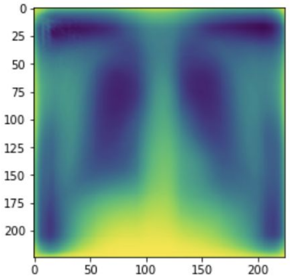


Overfitting while predicting Emphysema [Vivano 2019]

Example: Systematic discrepancy between
average image in datasets



NIH



PADCHEST

- Merging datasets with different class imbalance (confounding artifacts from each hospital)
- Labels confounding with each other
- Demographics confounding with labels

[Zeck, Confounding variables can degrade generalization performance of radiological deep learning models, 2018]
[Viviano, Underwhelming Generalization Improvements From Controlling Feature Attribution, 2019]
[Simpson, GradMask: Reduce Overfitting by Regularizing Saliency, 2019]
[Ross, Right for the Right Reasons, 2017]

Mitigation approaches

Feature engineering

- **Range normalization** (/max)
- **Subspace alignment** (align data using their eigenbases)

During training

- **Reverse gradient** [Ganin & Lempitsky, Unsupervised Domain Adaptation by Backpropagation, 2014]
- **Right for the Right Reasons regularization** [Ross, Hughes, & Finale Doshi-Velez, 2017]
- **GradMask contrast loss** [Simpson, 2019]
- **ActivDiff** [Viviano, 2019]

What if feature artifact is correlated with target label?
Is the reason that should be used for prediction known?
What if it is not known?

Not discussed

Image Registration

Cell morphology representation (e.g. BBBC021)