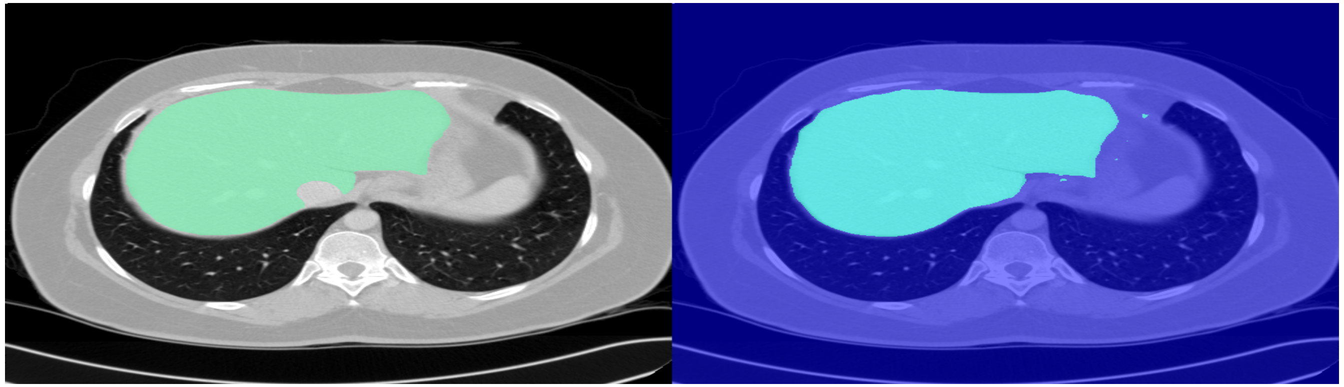
Testing deep segmentation of Computer Tomography scans

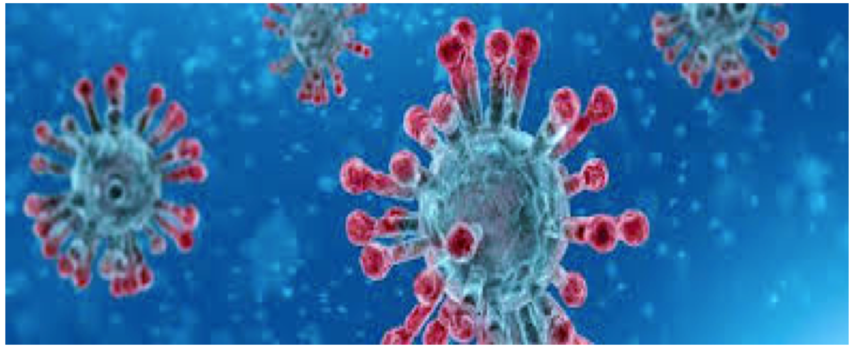


Pedro Furtado @ ICBBT 2020

pnf@dei.uc.pt

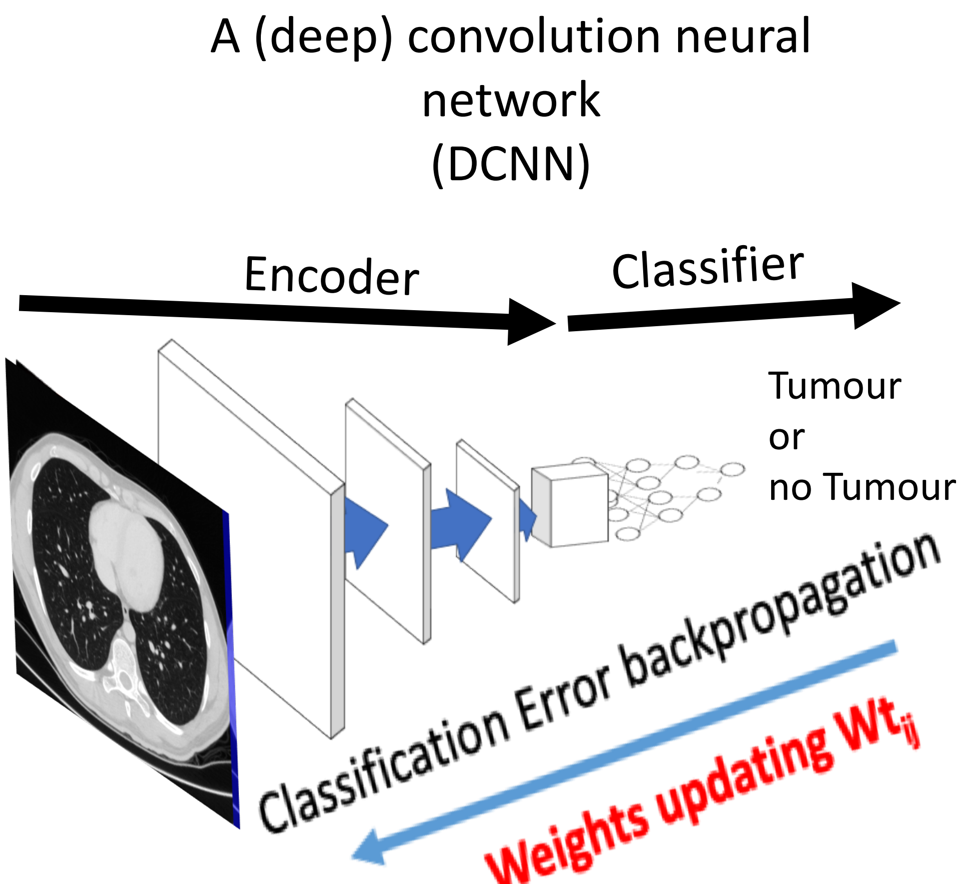
Pedro Furtado
Faculty of Science and Technology, DEI/CISUC,
University of Coimbra. Portugal

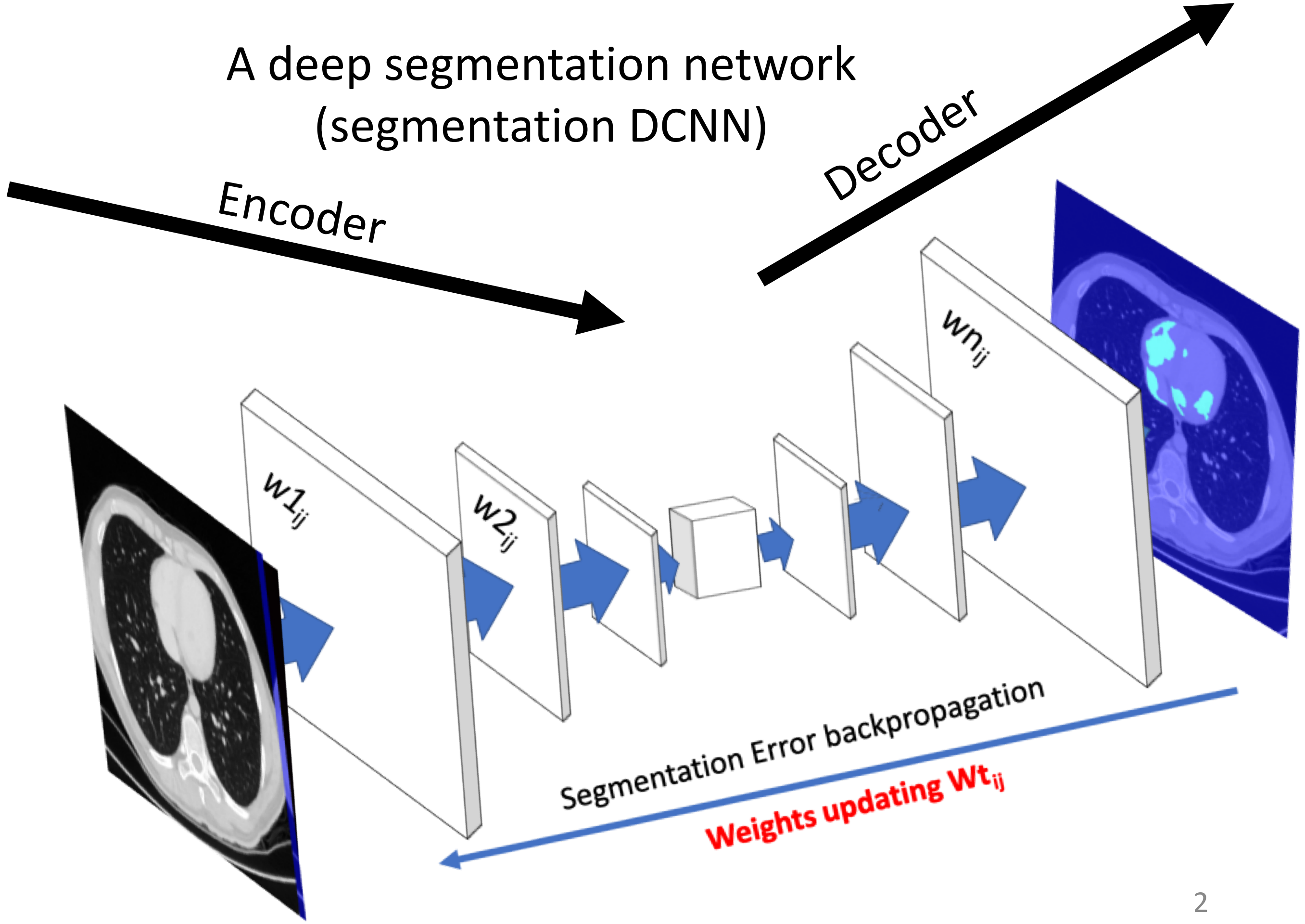
In a time of COVID-19...



Background

- Convolution neural networks are used for classification and segmentation
- **Classify an image** **Classify each pixel in an image**





What is the issue we investigate?

- *STATEMENT: A feedforward network with a single layer is sufficient to represent any function, … but the layer may be infeasibly large and may fail to learn and generalize correctly.*

  —**Ian Goodfellow, DLB**

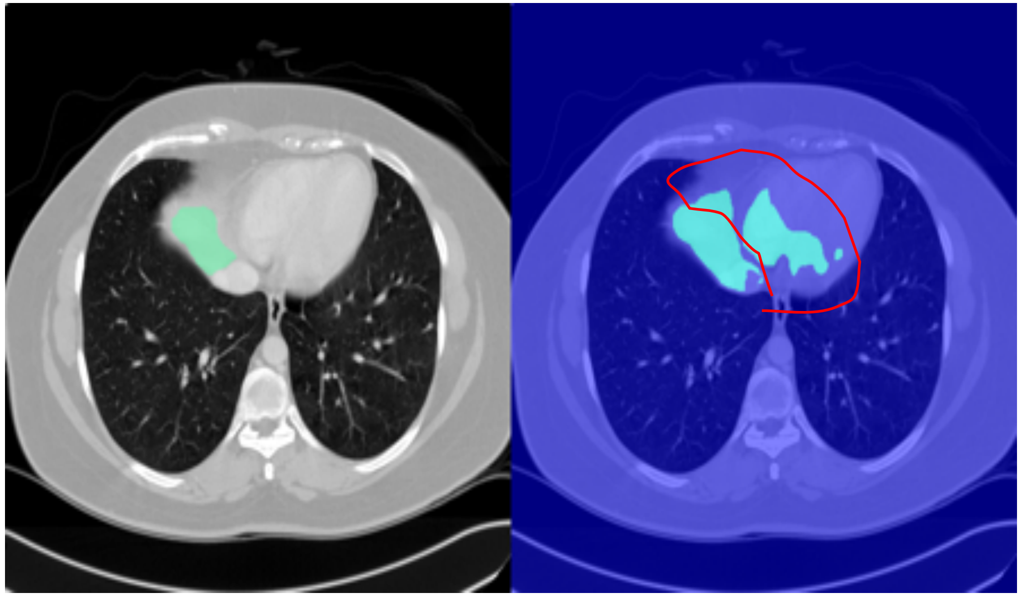
- Convolution neural networks (DCNNs) seem very well adapted to learn from images in a supervised manner, from training images
- Most people like to think that segmentation DCNNs are almost 100% perfect

SORRY, DATA IS NOT PERFECT ➔ ITS A FACT OF LIFE!

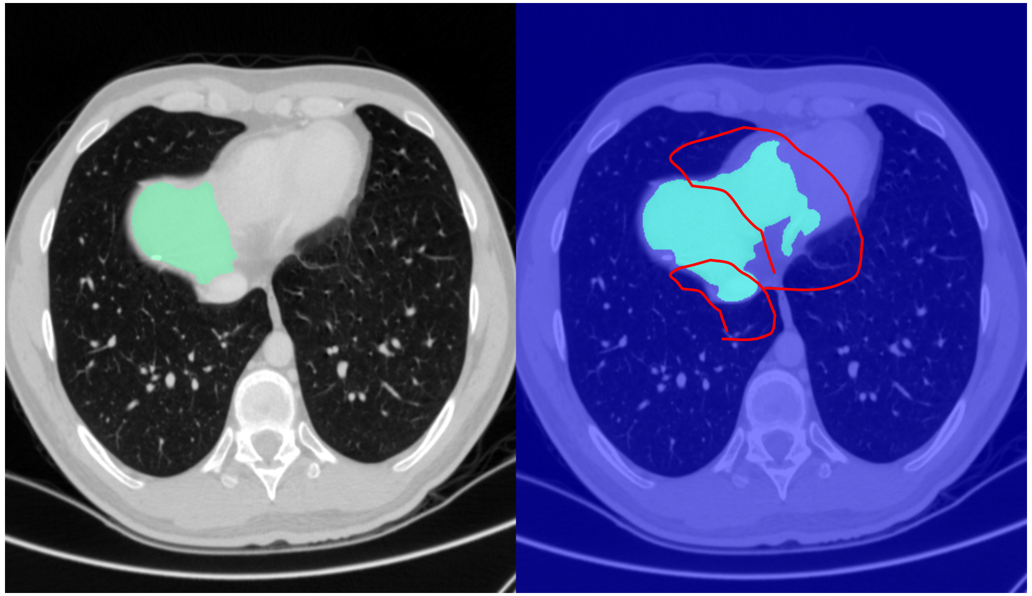
What is the problem?

- After a whole lot of training, with one of the best possible DCNNs I could find...
- And please, don't blame me... and don't blame data size

GNDtruth **Segments**



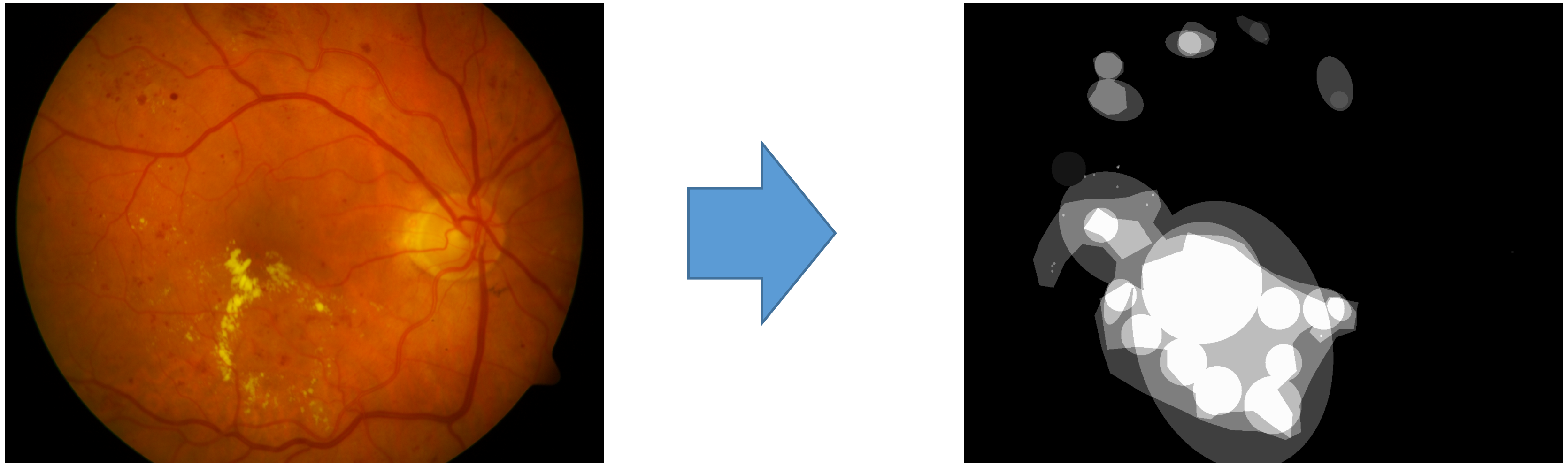
GNDtruth **Segments**



Semantic segmentation

- I have often seen tolerant assumptions used in publications, such, for instance,
  - *that a lesion found within a large GNDTRUTH region is a TRUE POSITIVE*
  - *a pixel less than 500% of its size away from the GNDTRUTH region is TRUE POSITIVE...*
- Same with TN, FP, FN ->

results become mush better, nearing perfection... And people believe...

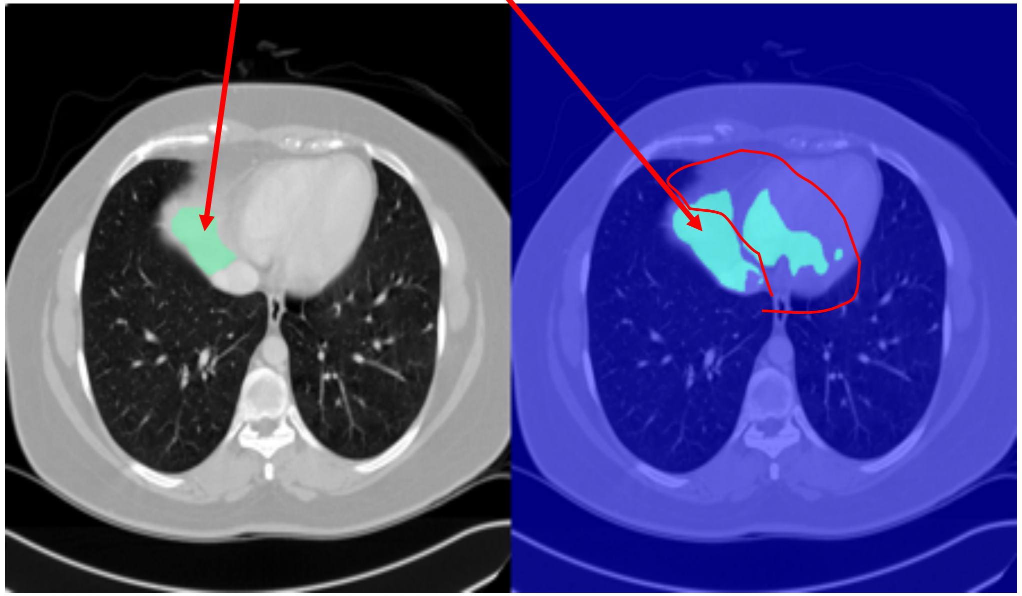


BUT Semantic segmentation… What about if we evaluate exactly…

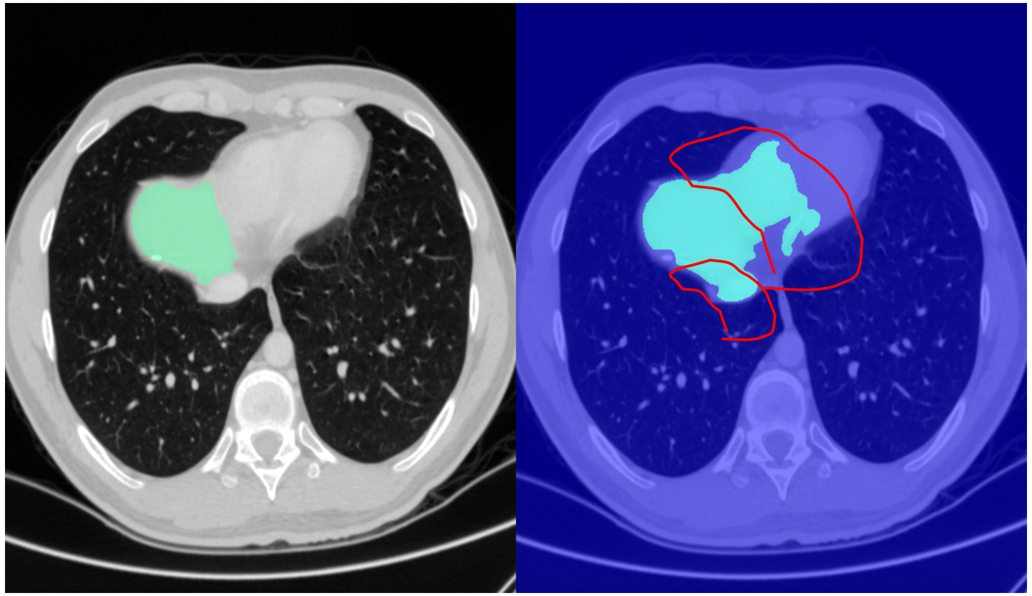
What is a True Positive (TP)/Negative (TN) and a False Positive (FP)/Negative (FN) to me?

- TP liver is a liver output pixel that is a liver pixel in GNDTRUTH as well and was recognized as liver
- Same care with TN, FN, FP

GNDtruth Segments



GNDtruth Segments



Some related work results segmenting the liver CT: FAR FROM 100% perfect

- Scale 0 to 100 (no good -> perfect)

CT-Liver segmentation results:

PKDIA	-	82.457
MedianCHAOS6	-	80.446
MedianCHAOS3	-	80.428
MedianCHAOS1	-	79.91
MedianCHAOS2	-	79.778
MedianCHAOS5	-	73.392
OvGUMEMoRIAL	-	61.129

What we do:

- We believe it is important to try to UNDERSTAND where things fail...
- ... And teh exact quality...

Investigative Method:

Build Deep Segmentation Networks
Pre-trained

↓

Setup Dataset with gndtruths
Define Evaluation Metrics

↓

Train Seg Nets with images

↓

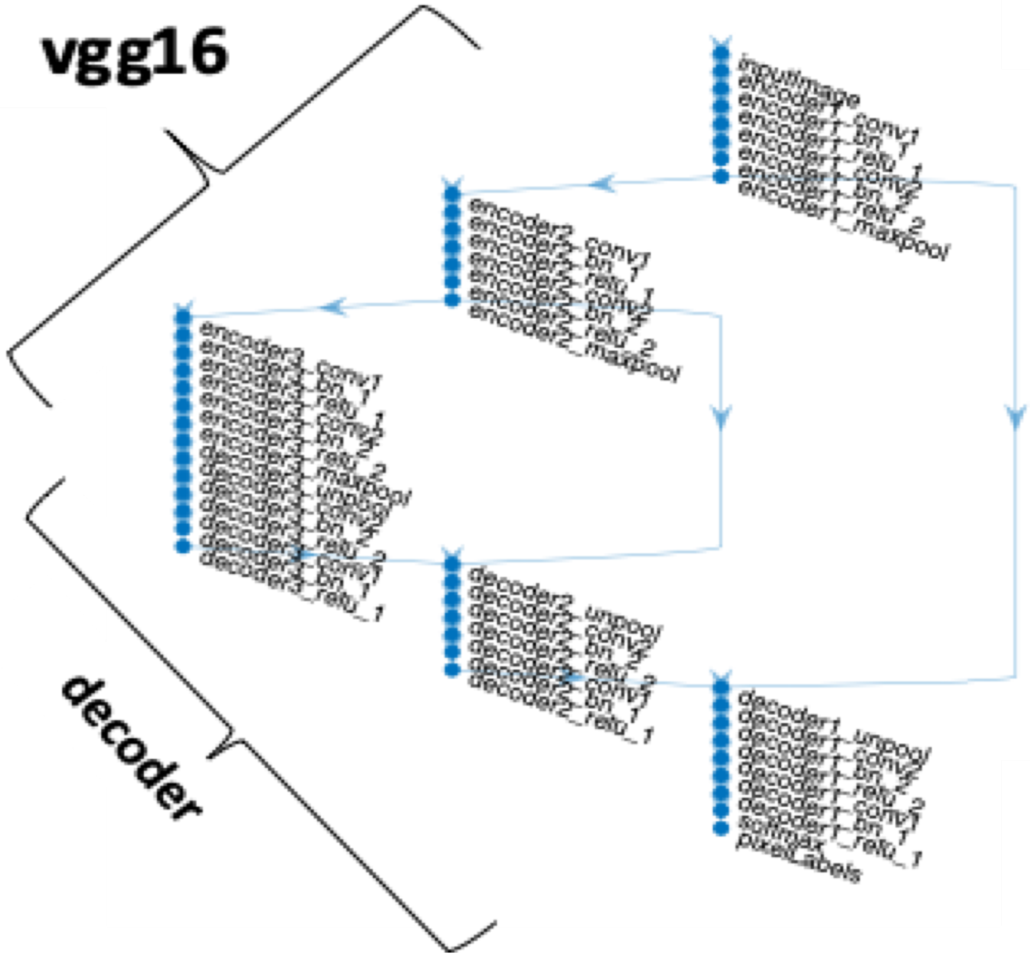
Evaluate with Test Data

↓

Interpret Metrics to Conclude

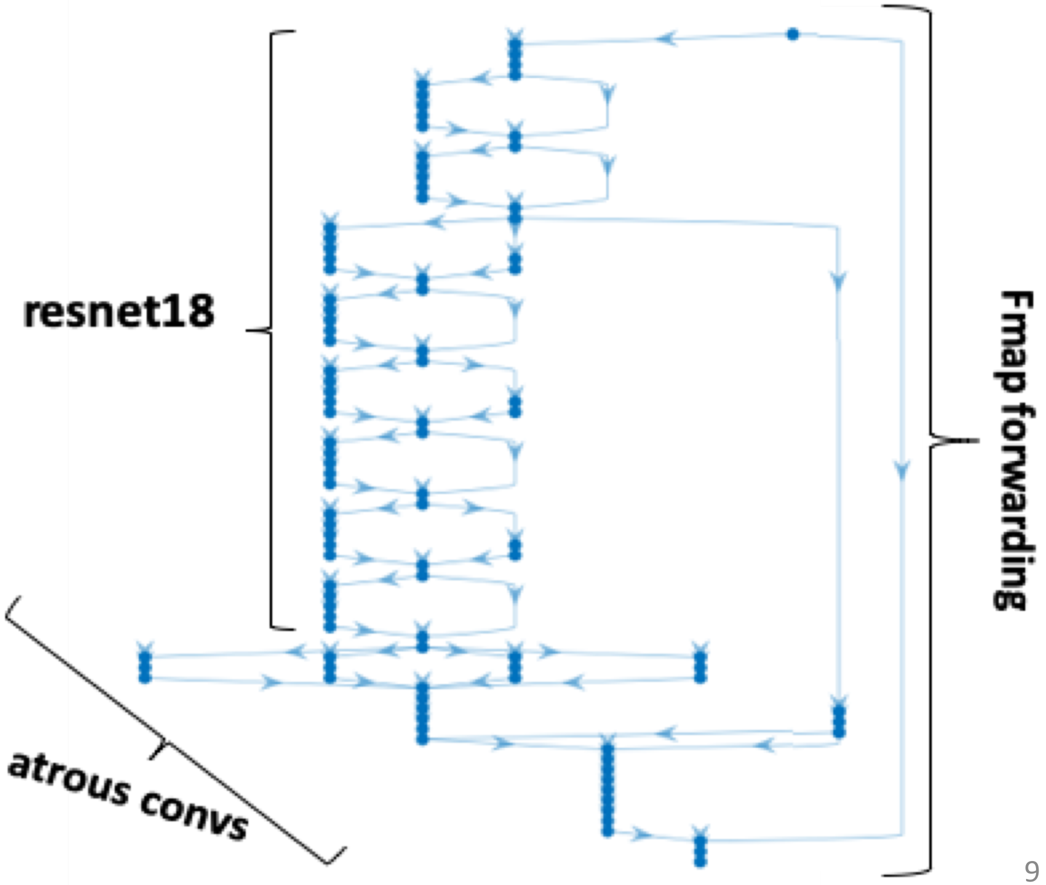
- After that work on optimizing the approaches is better guided by knowledge

Segnet:



DeepLabV3:

* 100 layers, Resnet-18 feature extractor = 71 layers pre-trained network
* Atrous Spatial Pyramid Pool- ing (ASPP) layers
  improves segmenting of objects at multiple scales.
* Outputs combined with Conditional Random Field (CRF) for
  improved localization of object boundaries



Computer tomography data used ...

- CT scans of **40** different patients, **77 to 105 slices per patient**,
- Philips SecuraCT with 16 detectors, Philips Mx8000 with 64 detectors and Toshiba AquilionOne with 320 detectors.
- **For evaluation, the patients were divided randomly into train and test folds (80%/20%).**
- CHAOS data ([6][7][13])

First worthless metrics results…

- ACCURACY SEEMS PERFECT…weighted IoU also… Recall also…sensitivity also… but **why are those not enough as metrics?**

	Global Accuracy	Mean Accuracy	Weighted IoU
DEEPLAB	0.98	0.98	0.96
SEGNET	0.92	0.95	0.88

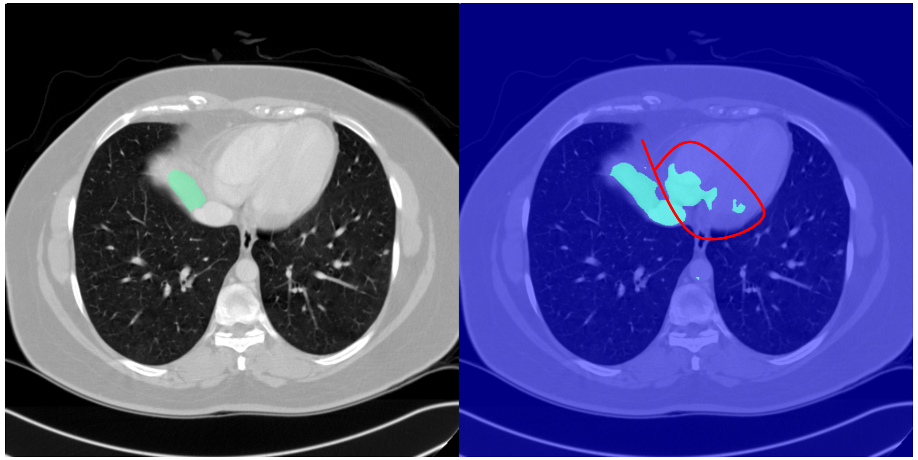
AND
Per class

Background	Accuracy
DEEPLAB	0.98
SEGNET	0.91

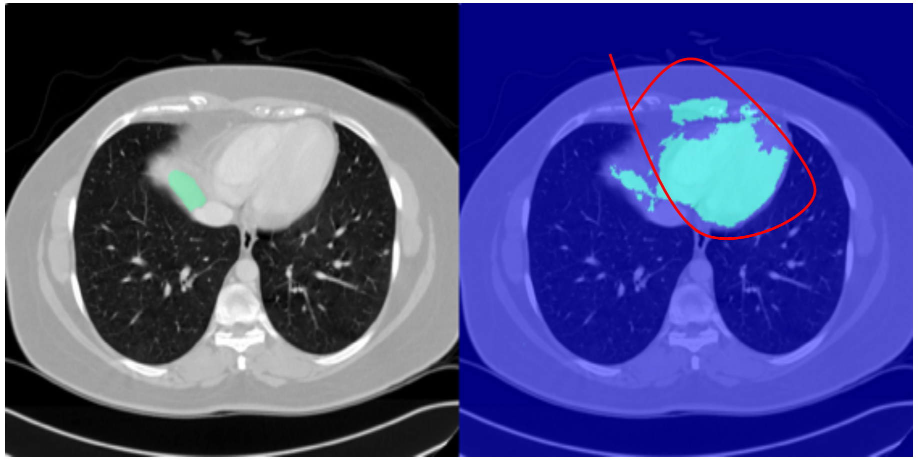
Liver	Accuracy
DEEPLAB	0.97
SEGNET	0.99

You can see the errors by visualization examples...

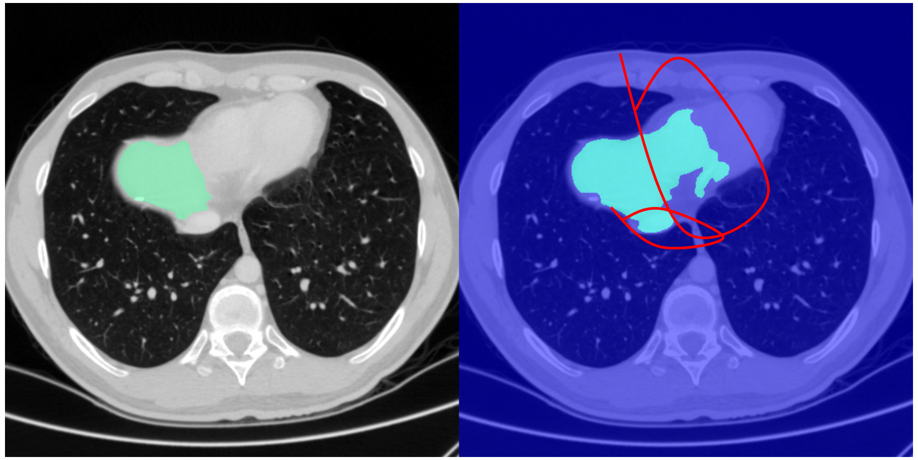
DeepLab V3



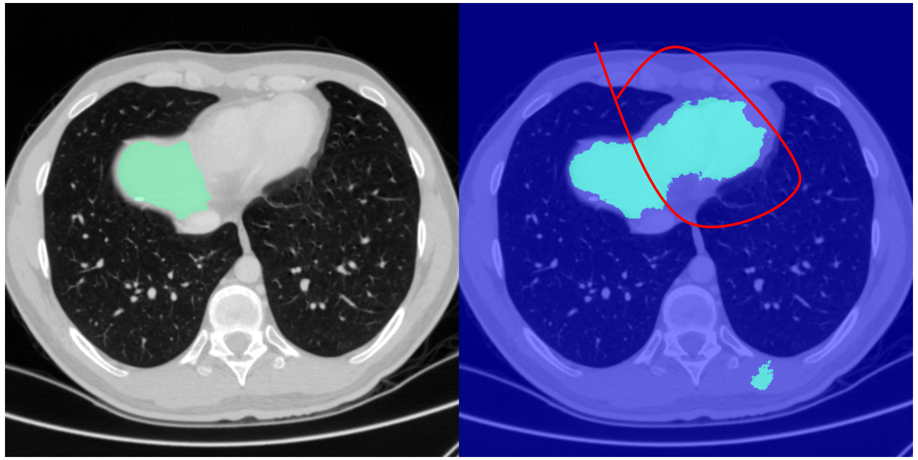
Segnet



DeepLab V3

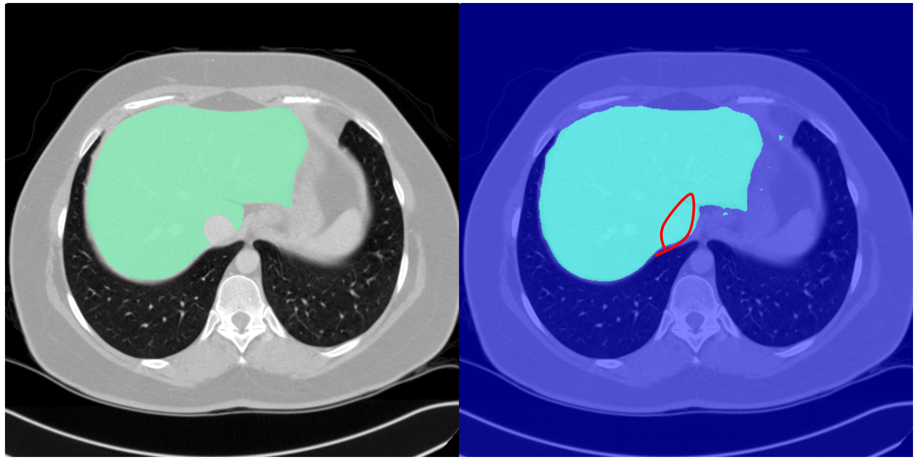


Segnet

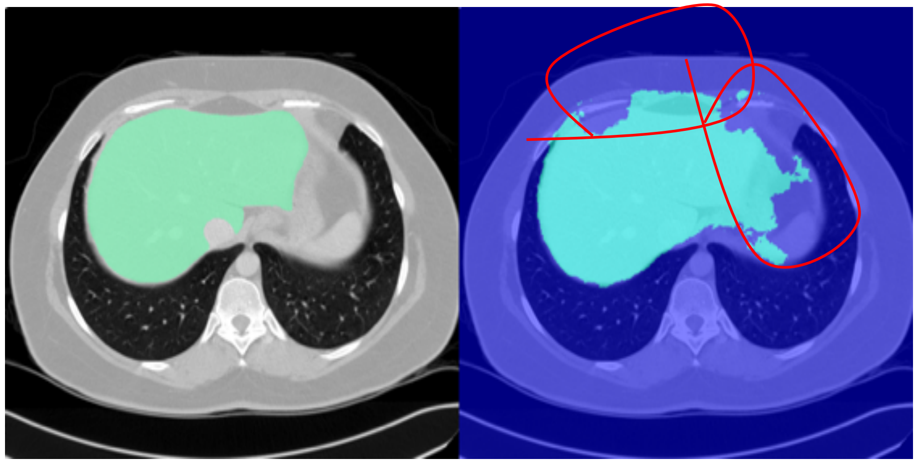


More visualization examples…

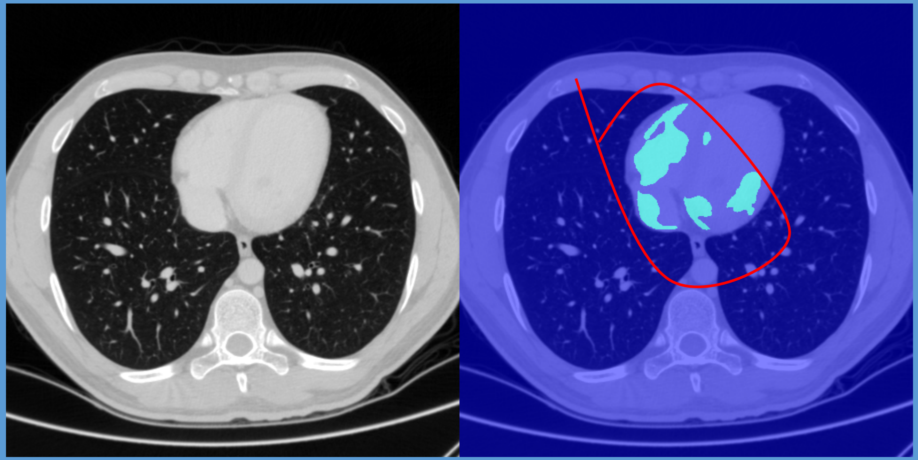
DeepLab V3

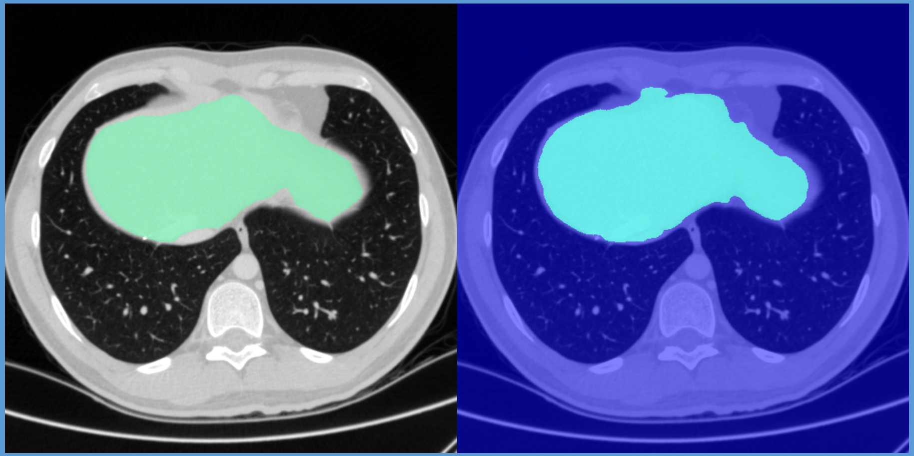


Segnet



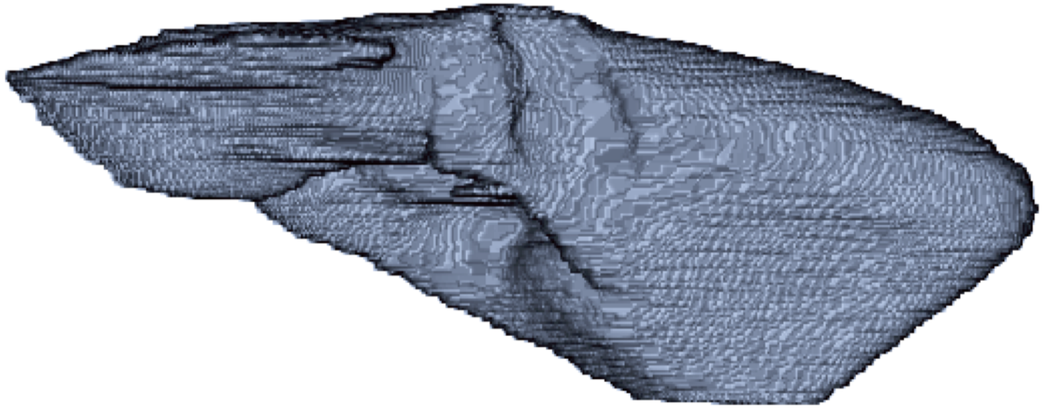
DeepLab V3



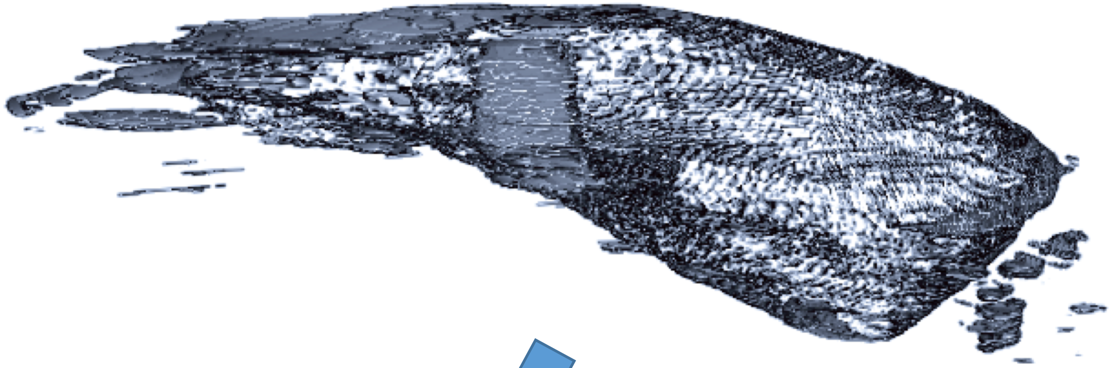


Another visualization:

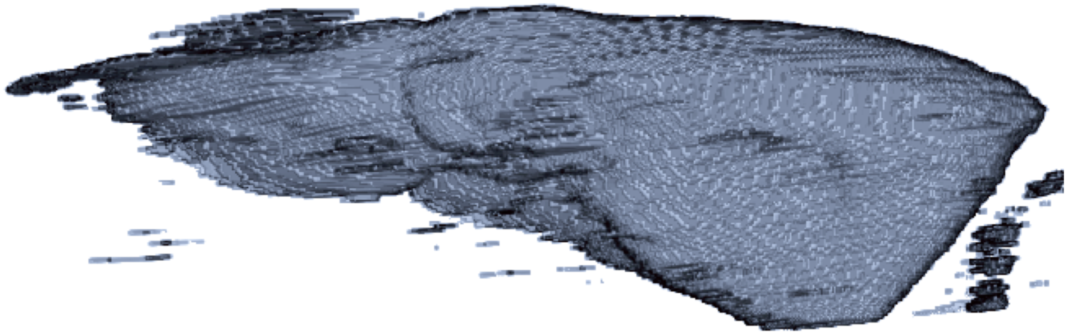
Groundtruth GND



|DL – GND|



DeepLabV3 seg DL



Why do some metrics make it seem so perfect?

- The background is about 93% of all slides
- Vast majority of the background is easy to segment well by learning

=> Any pixel or class aggregate metric is going to eval mostly background

- Accuracy is especially bad...

- We needs metrics on the **LIVER**
- And metrics must be used carefully

Why we need IoU...

- **Accuracy (over all pixels)** = recall = fraction of correct pixels classifications

  acc= (TP+TN)/ALL **Background is BIG=>** **99% well classified**

- **Accuracy or recall of liver** = fraction of correct classifications of liver pixels

  acc(liver) = recall(liver) = $TP_{liver}/(TP_{liver}+FN_{liver})$ => 97%, **very GOOD also,**

  liver pixels are well classified

- **IoU** = degree of "exact matching" of regions = ratio of pixels of object well classified by all

  IoU(liver)= TP_l / (TP_l + FN_l + $\mathbf{FP_l}$) => adds FP_{liver} **= BKGND PIXELS as LIVER**

- **Precision, BF-Score, dice would also reveal FP_{liver}**

Experimental conclusion:
We need more focus on improving....
*Filtering **bkgnd false positive LIVER***

Deeplab	precision	recall
BackGround	0.99	0.98
liver	**0.80**	0.97
Segnet	precision	recall
BackGround	0.99	0.91
liver	**0.46**	0.99

Background	IoU
DEEPLAB	0.98
SEGNET	0.91

Liver	IoU
DEEPLAB	**0.78**
SEGNET	**0.46**

Conclusions

- Deep segmentation networks are amazing, they can learn to segment everything and with good quality...
- But they are not perfect, far from that...
- Significant number of BKGROUND pixels were classified as liver...
- DeepLabV3 was much better than Segnet (and others)... Probably the innovations in DeepLabV3, e.g. to improve object boundaries, improved the results
- **Conclusion: more research is needed into ways to improve current DCNNs further**

Our current work: loss functions, architectures, post-processing, False positves filtering

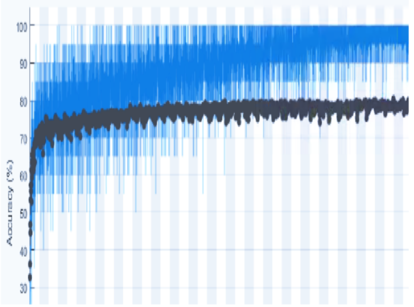
Thank you!

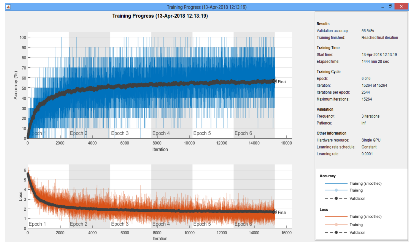
pnf@dei.uc.pt

https://eden.dei.uc.pt/~pnf/

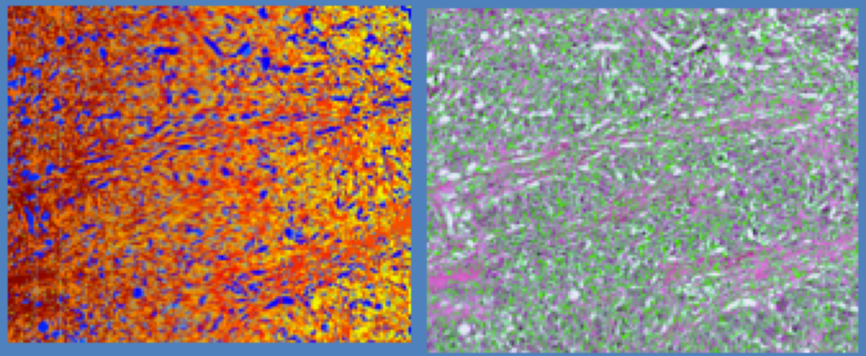
Pedro Furtado,

U. Coimbra, Portugal

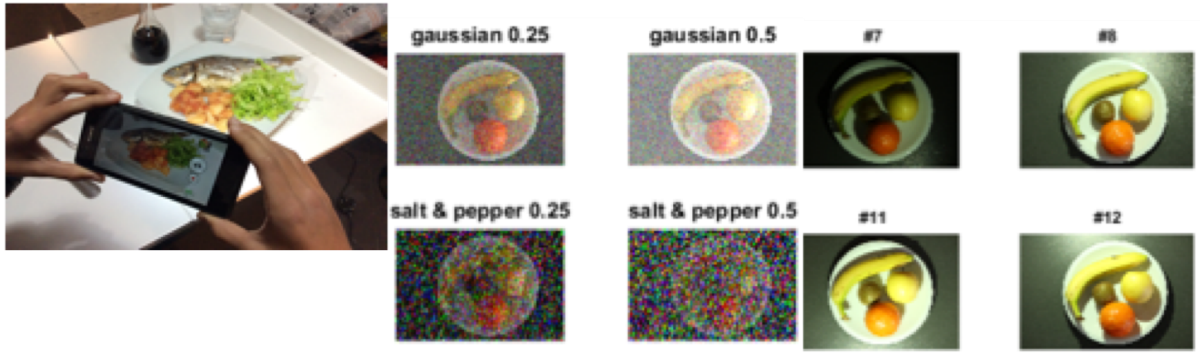




Breast Cancer Detection



Automated CHC In Self-management Of Diabetes



Seizures Detection