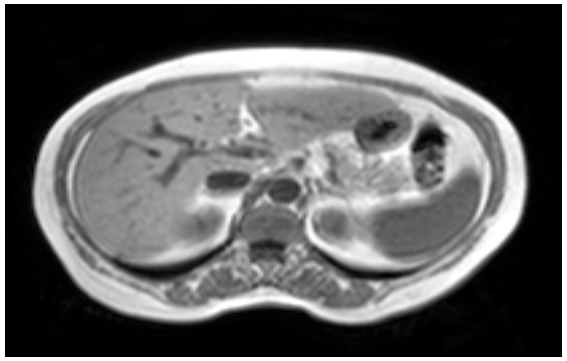
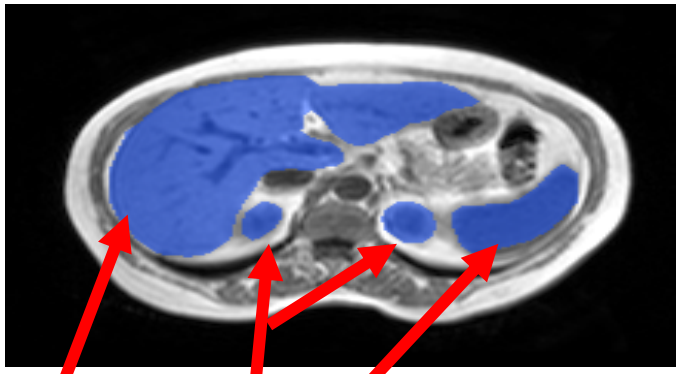
Magnetic Resonance Sequences: Experimental Assessment of Achievements and Limitations





Pedro Furtado @ ICDIP 2020

pnf@dei.uc.pt

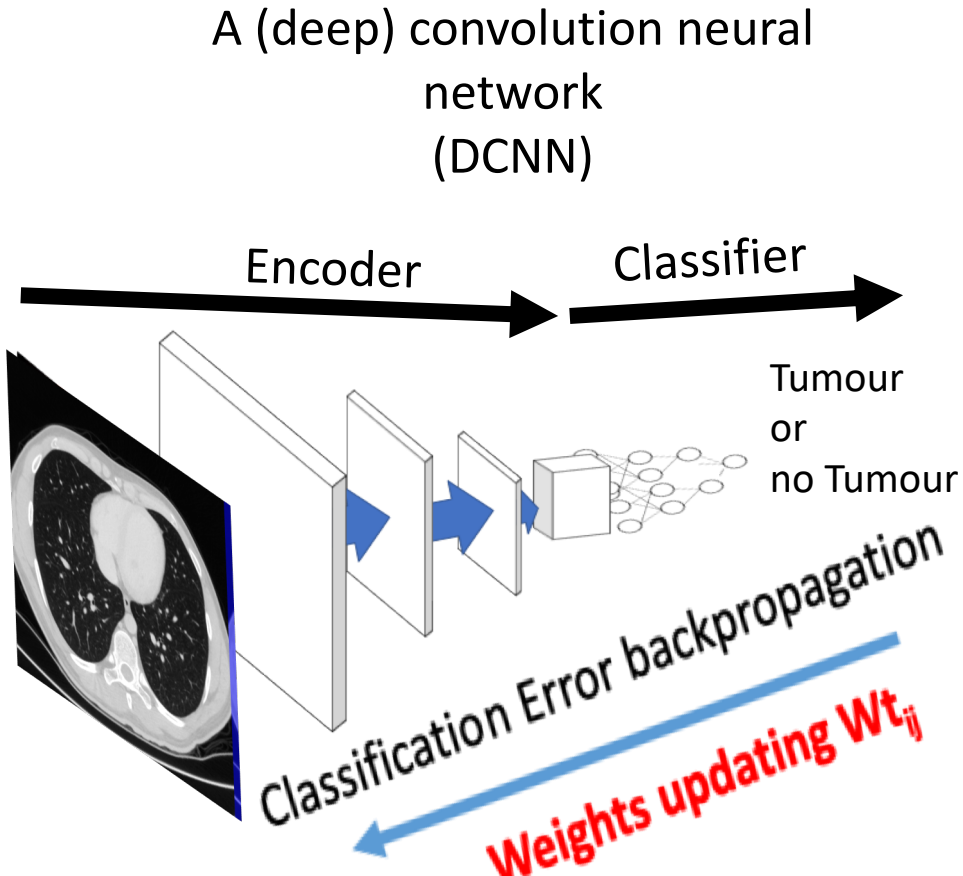
MRI images of abdominal organs:

- Liver
- Spleen
- Left kidney
- Right kidney

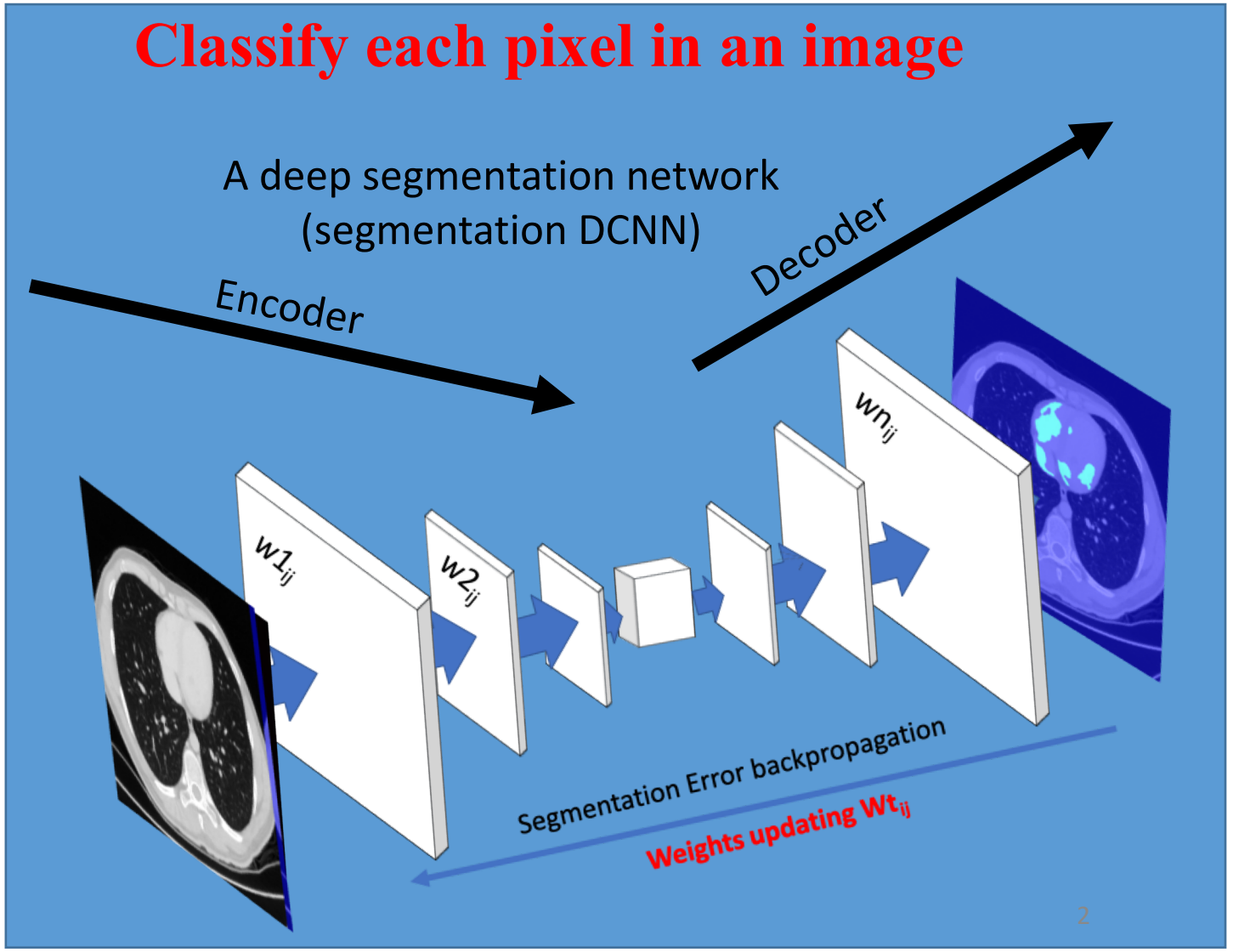
Pedro Furtado
Faculty of Science and Technology, DEI/CISUC,
University of Coimbra. Portugal

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

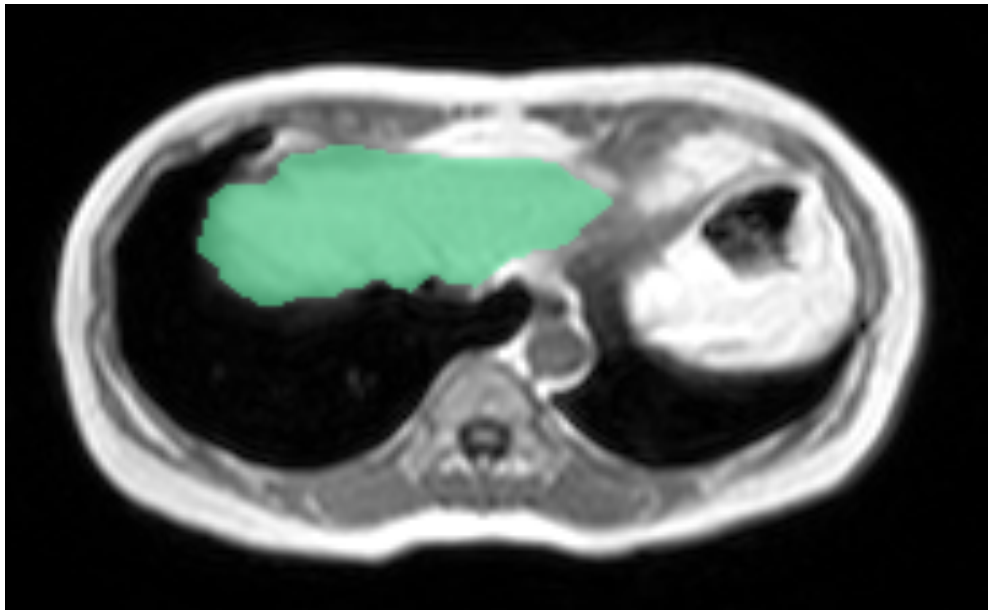
- Convolution neural networks (DCNNs) seem very well adapted to learn from images in a supervised manner, from training images
- Most people like to think that segmentation DCNNs are almost 100% perfect

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

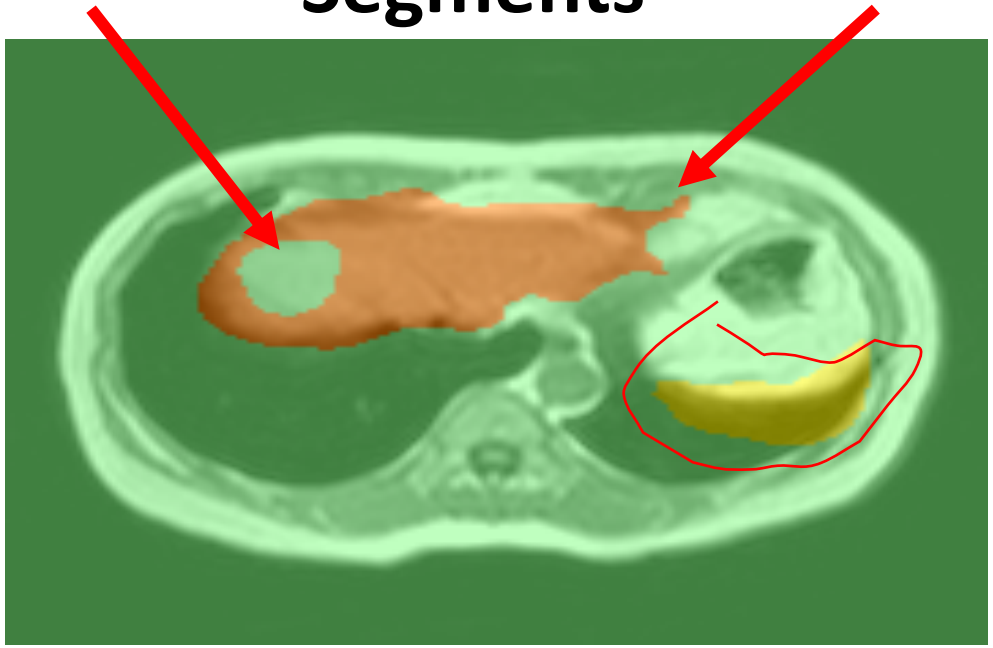
What is the problem?

- Deep convolutional neural network are very useful to segment medical images... but they are not perfect
- And please, don't blame me... and don't blame data size

GNDtruth



Segments



Some related work results segmenting abdominal organs from MRI:

- VERY FAR FROM 100% perfect
- **Sometimes COMPLEX mix of** METRICS is used, with thresholds??

E.g. CHAOS segmentation metrics:

- Sørensen–Dice coeficiente = degree of overlapping (0-100)
- Relative absolute volume difference (RAVD): (0-100)
- Average symmetric surface distance (ASSD), Average symmetric surface distance (ASSD), Maximum symmetric surface distance (MSSD): converted to (0-100)

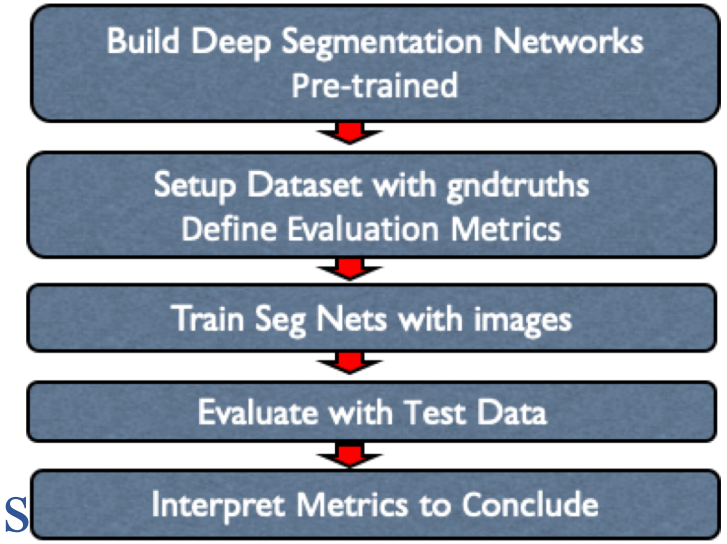
Other abdominal organs- segmentation results:

PKDIA	66.463
mountain	60.200
ISDUE	56.251
OvGUMEMoRIAL	44.343
IITKGP-KLIV	25.632

What we do:

- We believe it is important to try to UNDERSTAND where things fail

- We study the quality of segmentation of the Abdominal organs on MRI sequences, trying to determine where it is successful and where it is not so successful
  - *(1) have an initial quantification of segmentation of the organs*
  - *(2) compare DCNN networks;*
  - *(3) understand limitations of some metrics*
  - *(4) Understand where approaches still need improvements*



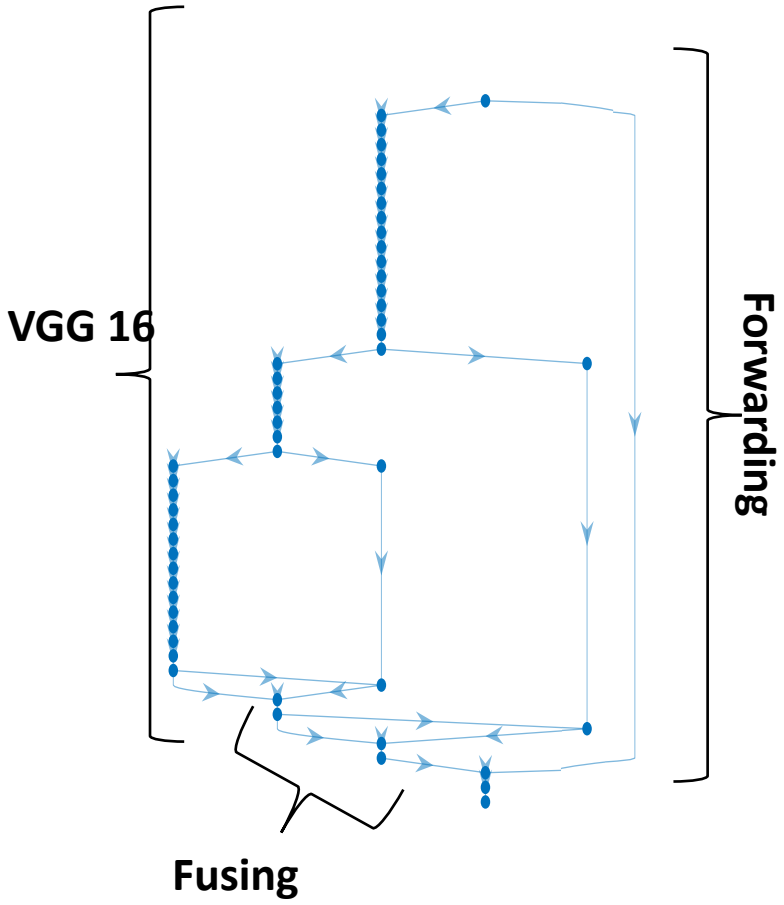
- Our own current work includes optimizing the approaches

Magnetic Resonance Imaging data used …

- MRI sequences of the abdominal region including the **kidneys, spleen and liver**, 256x256 resolution.
- Acquisition used T1-DUAL modality, a fat suppression sequence using difference of T1 times between fat and water protons.
- **1594 slices** from four patients sequences acquired and used in the experiments
- **For evaluation, the slices were divided randomly into train and test folds (80%/20%).**

- CHAOS data ([6][7][13])

FCN: ~50 layers



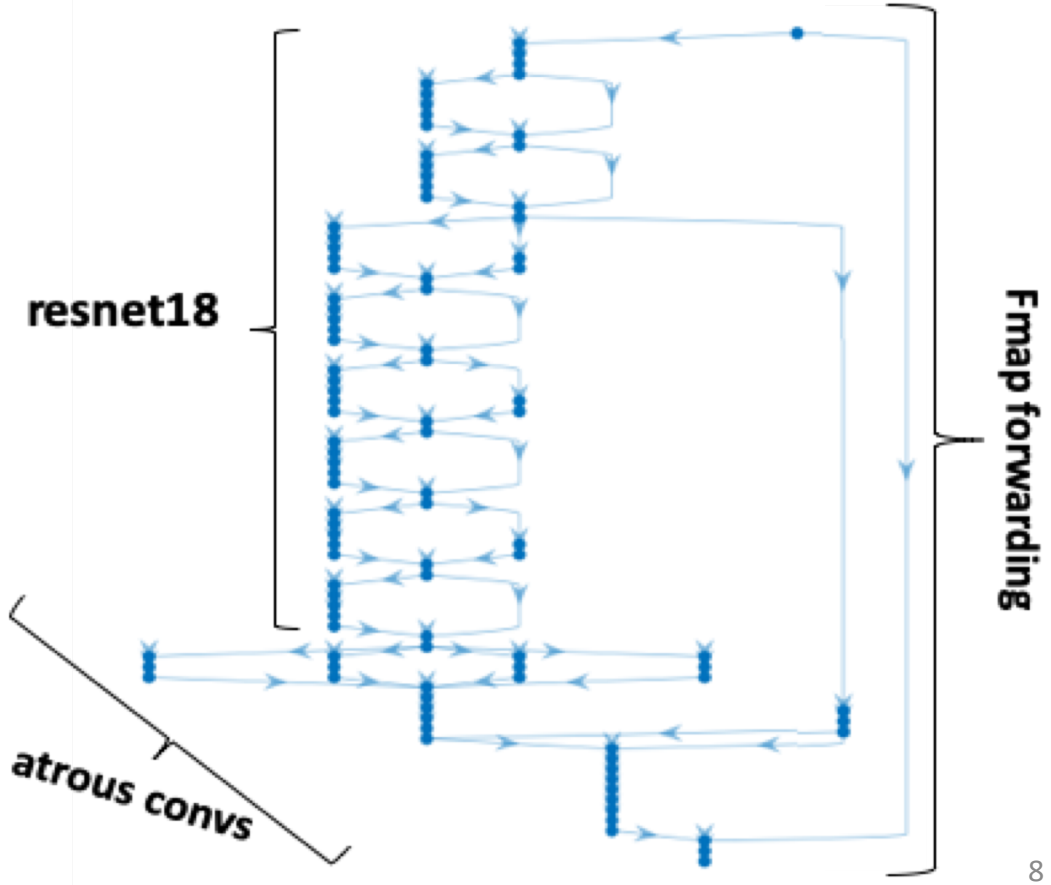
DeepLabV3:

~100 layers

Resnet-18 feature extract = 71 layer

Atrous Spatial Pyramid Pooling (ASPP), Conditional Random Field

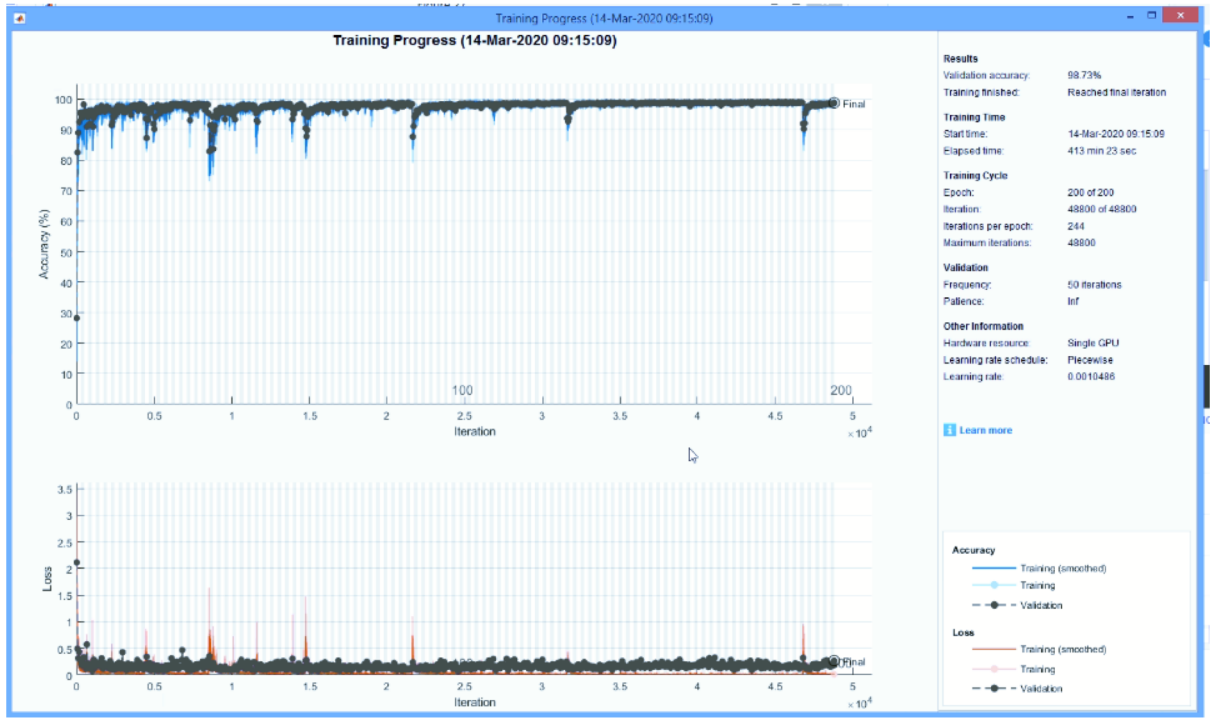
Improve segm multiscales + local object boundaries



We tested and adjusted best training options, etc…

- Type of learn rate schedule = 'piecewise' (rate decreases after period)
- Learn rate drop period=10 (time to decrease), factor=0.8 (how much), Momentum=0.9, Initial learn rate= 0.001;
- Maximum number of epochs= 500, batch size= 8 (4 to 64), Shuffle every epoch,…

- We adjusted class weights as required
- Data augmentation
- We tested several loss functions
- ….



Metrics results...

- Accuracy and weighted IoU seem perfect... Most people already know accuracy is not a good metric... but **why, and why this IoU?**

	Global Accuracy	Weighted IoU	Mean Accuracy
DEEPLABV3	0.98	0.97	0.98
FCN	0.97	0.96	0.91

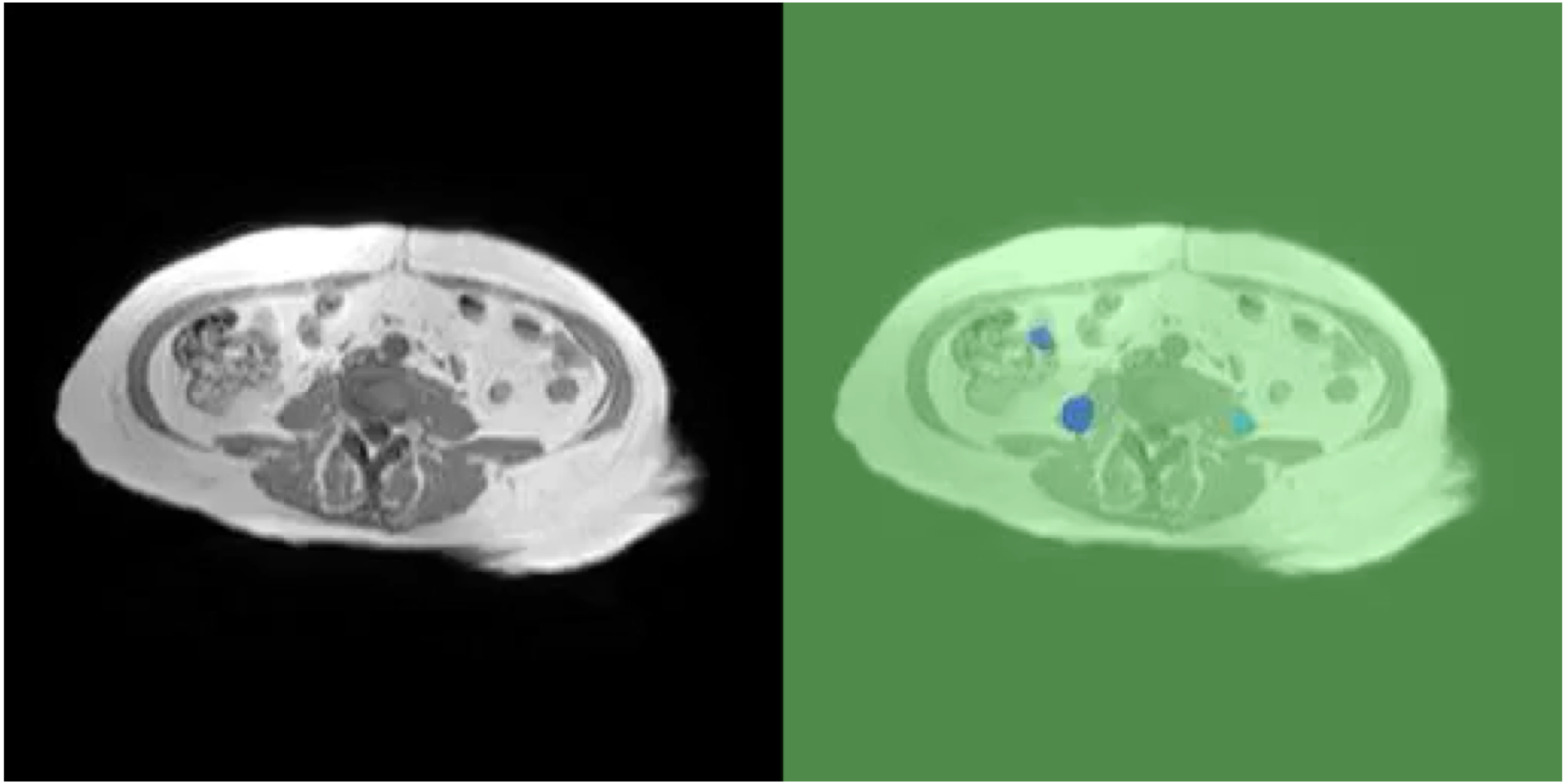
A "normalized" confusion matrix output by tool:

- Also seems almost perfect ... Because it is given as relative to rows !!!!...
  - *Meanw 99% of all lkidney were classified as such*
  - *BUT What about other things (e.g. bkgnd, others) classified as lkidney??? ($FP_{lkidney}$)*

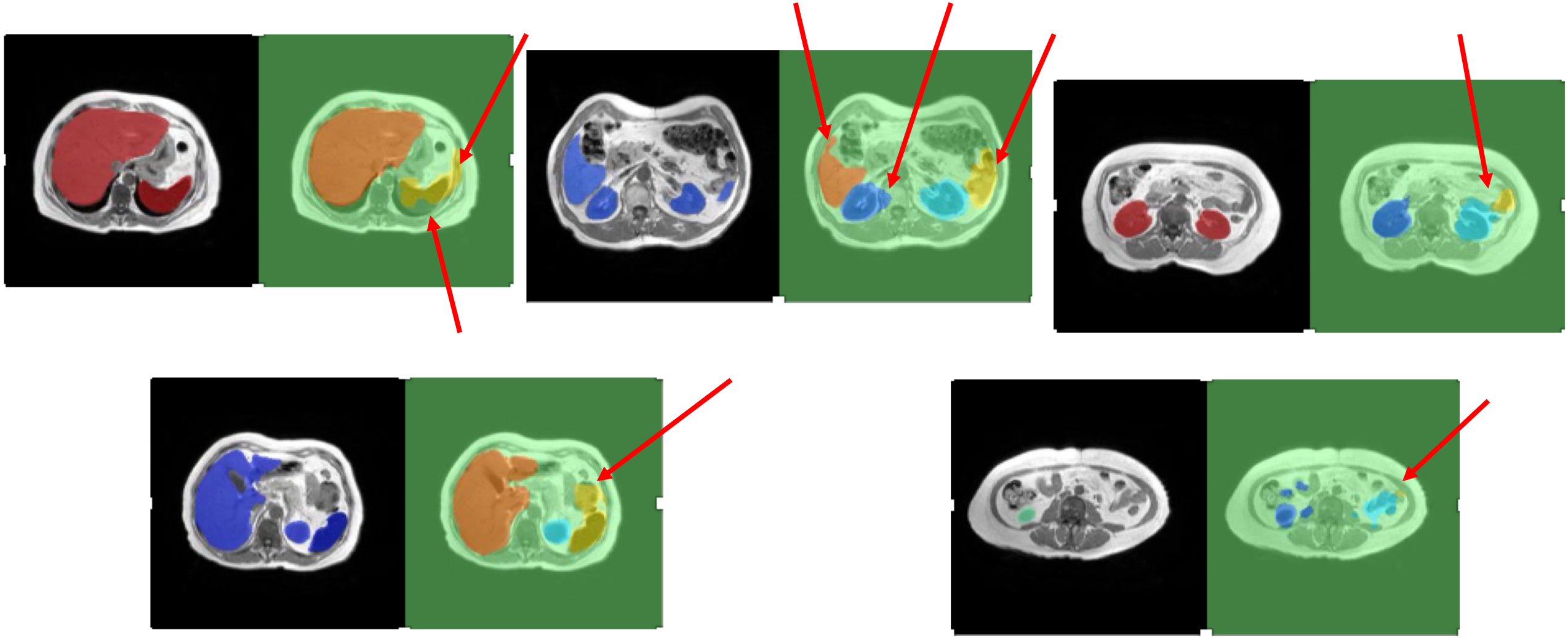
TABLE II. CONFUSION MATRIX DEEPLABV3

DEEPLAB	BackGrd	liver	spleen	rkidney	lkidney
BackGrd	**0.98**	0.01	0.00	0.00	0.01
liver	0.03	**0.97**	0.00	0.00	0.00
spleen	0.01	0.01	**0.98**	0.00	0.00
rkidney	0.04	0.00	0.00	**0.95**	0.01
lkidney	0.01	0.00	0.00	0.00	**0.99**

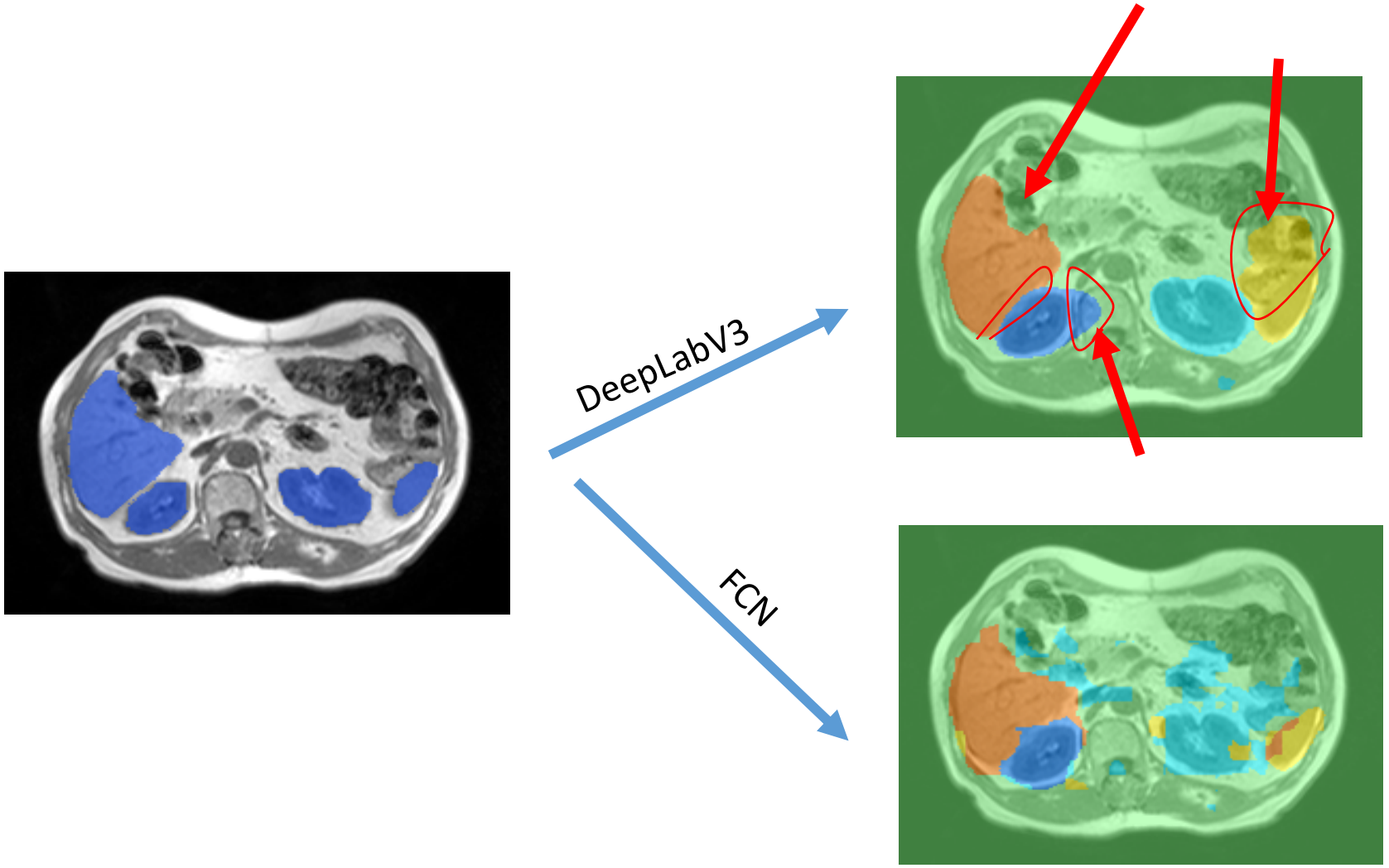
Some sequence of slices...



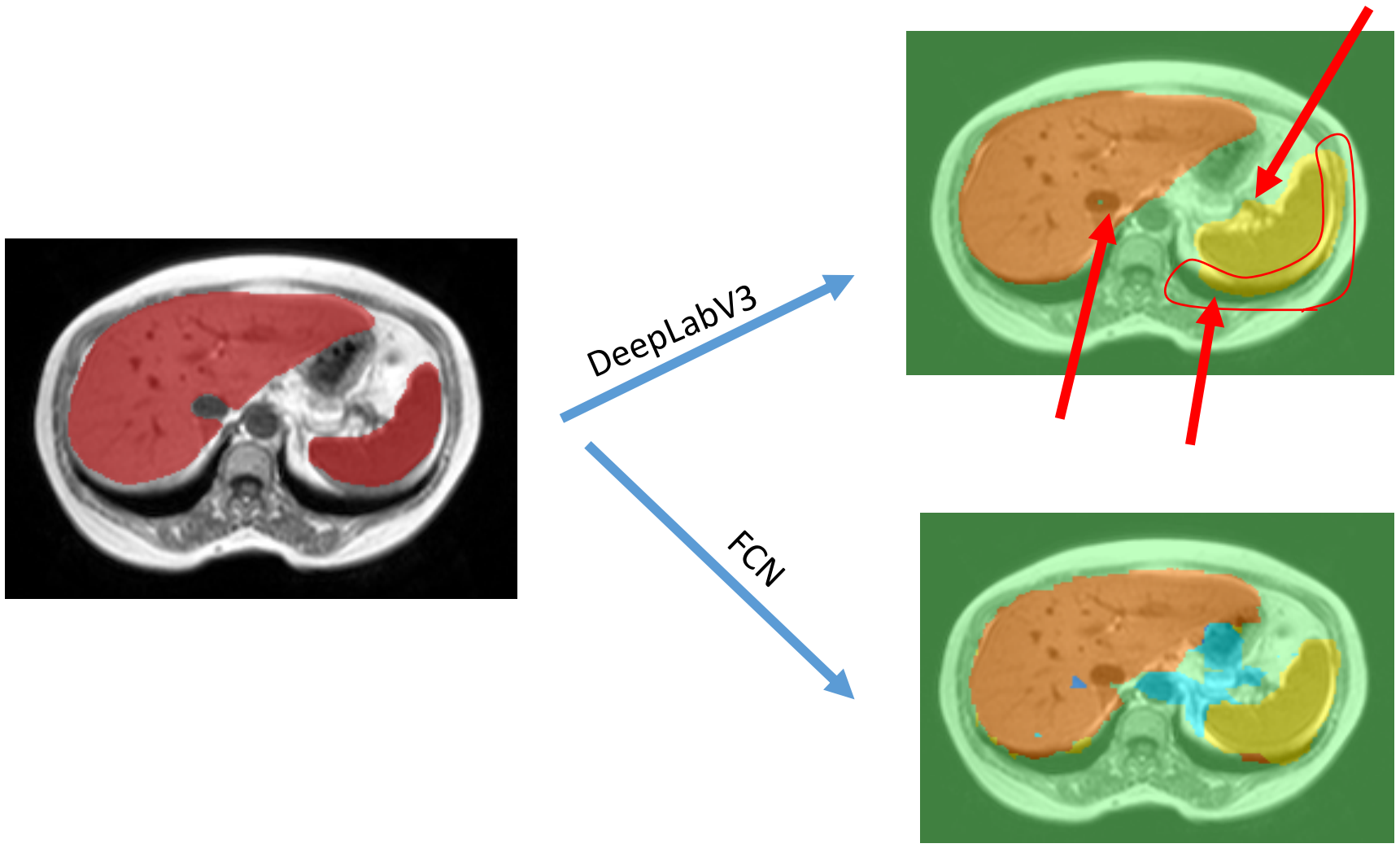
DeepLabV3 example slices..



Visualization of some results...



Visualization of some results... Worse in FCN

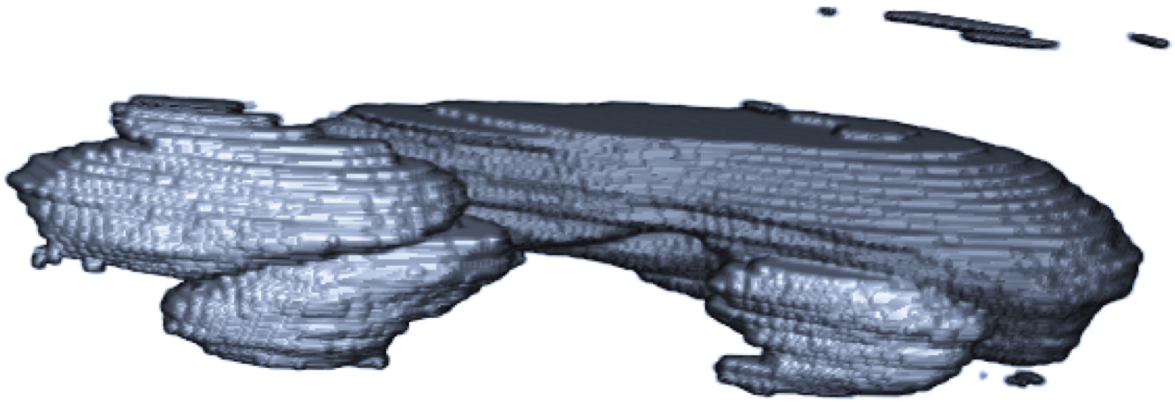


Another visualization:

Liver:

Groundtruth GND

DeepLabV3 seg DL



Why we need IoU…

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

  acc= (TP+TN)/ALL **Background is BIG=>** **98%** **well classified**

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

  acc(organ X) = recall(x) = $TP_x/(TP_x+FN_x)$ => ~97%, **GOOD also,**

  organ pixels are well classified

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

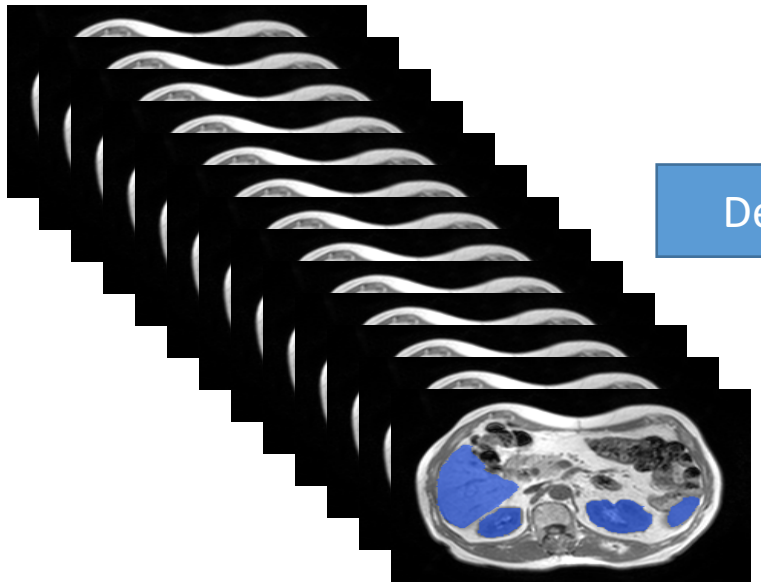
  IoU(organ x)= $TP_x / (TP_x + FN_x +$ **FP_x**) => adds **FP_x = OTHER PIXELS as ORGAN**

- **BUT WE MUST USE IoU of each organ**

- **Note: other metrics can also help: Precision, BF-Score, dice would also reveal FP_X**

Even with IoU we need to be careful!!

- In the following real case segmenting MRI with deepLabV3...
  - *IoU, measured as "weighted IoU", was 97%*
  - *IoU measured as "mean IoU" is 69%*
  - *true "quality" segmenting organs ranges between* ***41% and 78%***



DeepLabV3

Weighted IoU ~=97%

left kidney IoU=41%

Quantification of those problems…

TABLE IV. PER CLASS PERFORMANCE DEEPLABV3

DEEPLABv3	Accuracy(%)	IoU(%)	BFScore(%)
BackGrd	97	97	90
liver	96	78	60
spleen	97	59	56
rkidney	94	66	56
lkidney	98	41	36

- Accuracy of any class and IoU of bckgnd are perfect -> BCKGND 97%, easy
- IoU of each organ not very good 41%, 59%, 66%, 78%
- BF-Score also reveals problems
- FCN is much worse

TABLE V. PER-CLASS PERFORMANCE FCN

FCN	Accuracy(%)	IoU(%)	BFScore(%)
BackGround	0.96	0.96	0.81
liver	0.90	0.68	0.46
spleen	0.90	0.49	0.39
rkidney	0.88	0.18	0.18
lkidney	0.86	0.35	0.36

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 and organ pixels were classified as other organs...
- DeepLabV3 was much better than 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 positives filtering

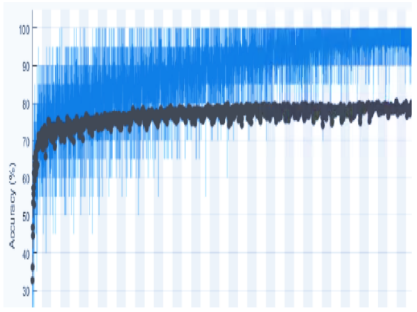
Thank you!

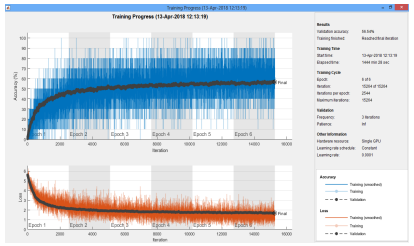
pnf@dei.uc.pt

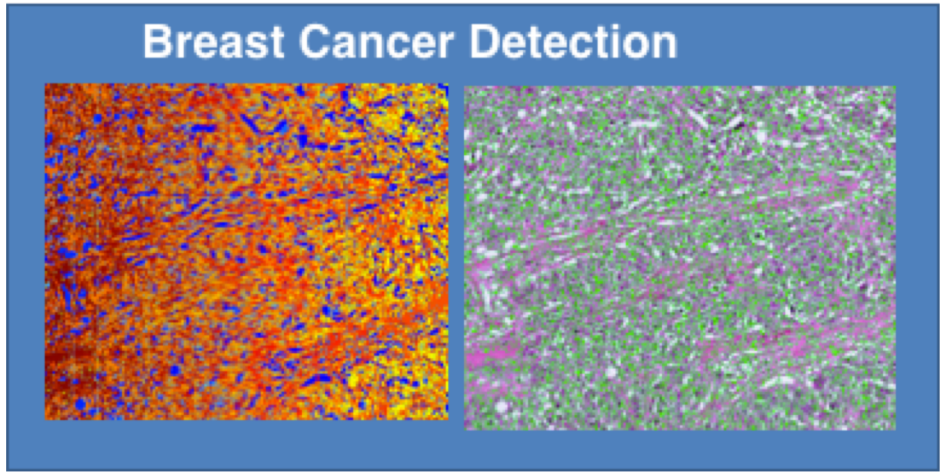
https://eden.dei.uc.pt/~pnf/

Pedro Furtado,

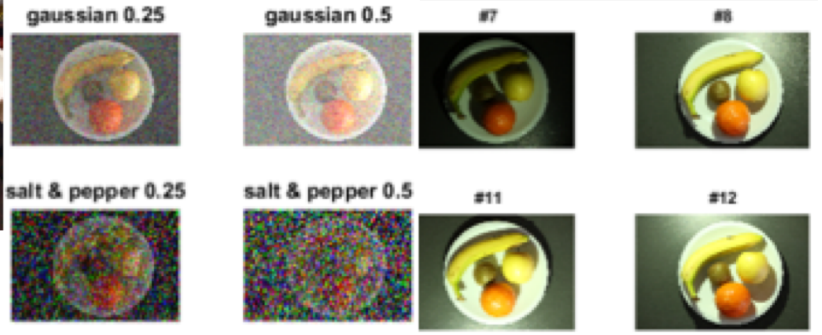
U. Coimbra, Portugal



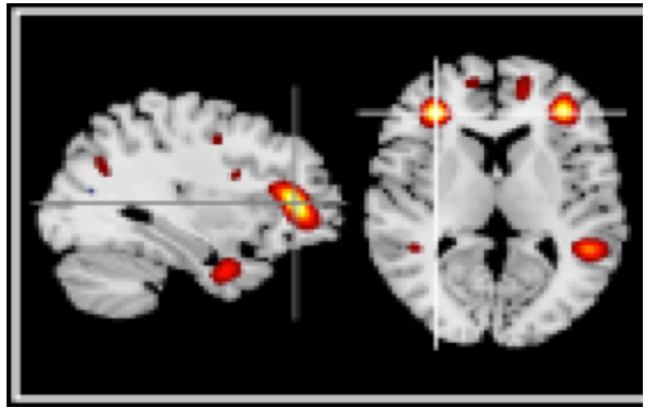




Automated CHC In Self-management Of Diabetes

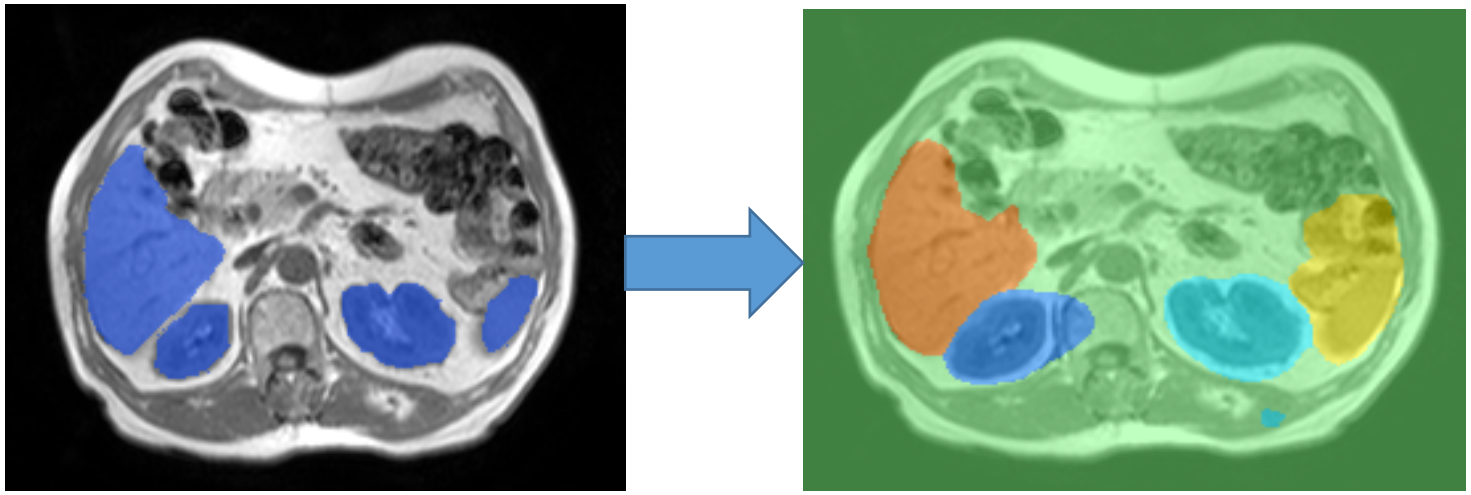


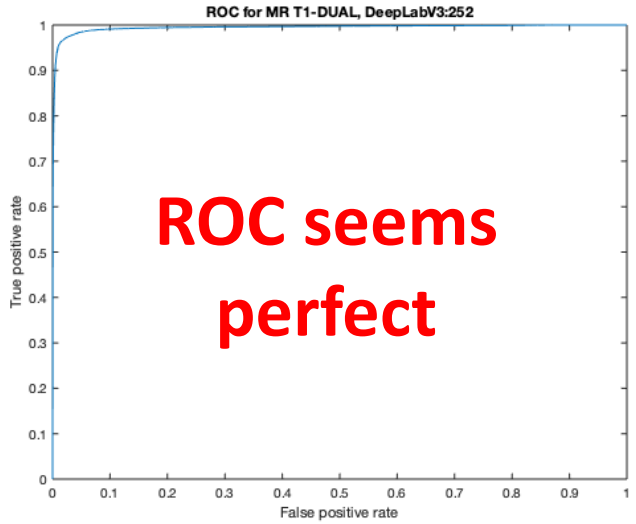
Seizures Detection



ROC also strange!!

- Everybody loves ROC with TPR/FPR...
- In the next case segmenting the spleen in MRI scans using DeepLabV3, ROC and AUC was 91%, but the true quality was 12%





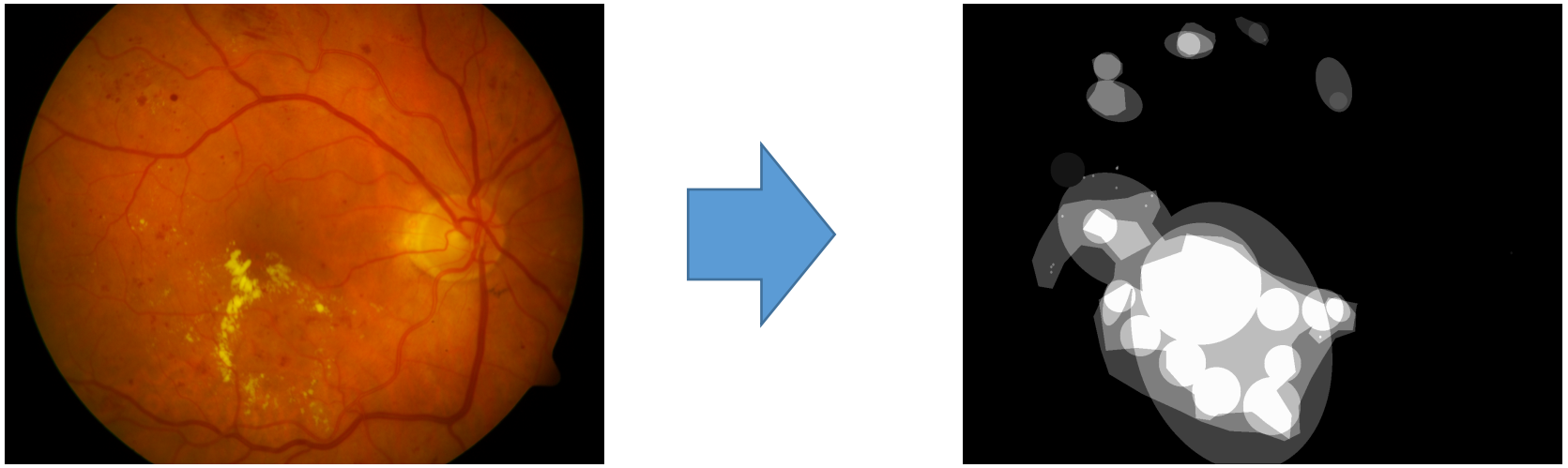
ROC seems perfect

AUC =0.9091

Semantic segmentation and rough 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 in the GNDTRUTH background region "close" to the lesion (less than 10x the size of th elsion) is a lesion TRUE POSITIVE...*
- Same with TN, FP, FN ->

results nearing perfection... But its not Semantic Segmentation

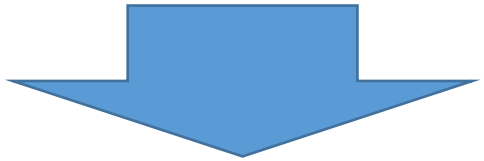


Why do some metrics make it seem so perfect?

- 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 **EACH INDIVIDUAL ORGAN**
- And metrics must be used carefully