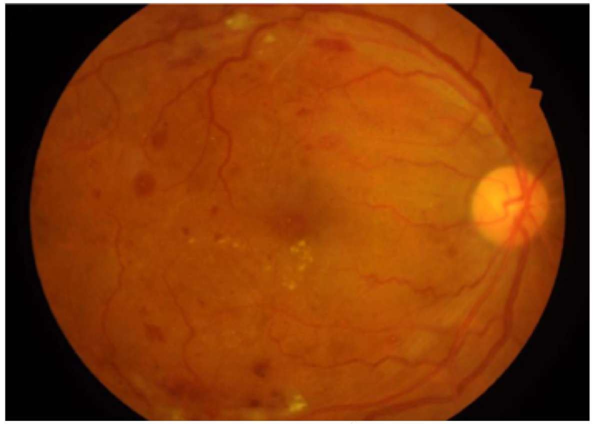
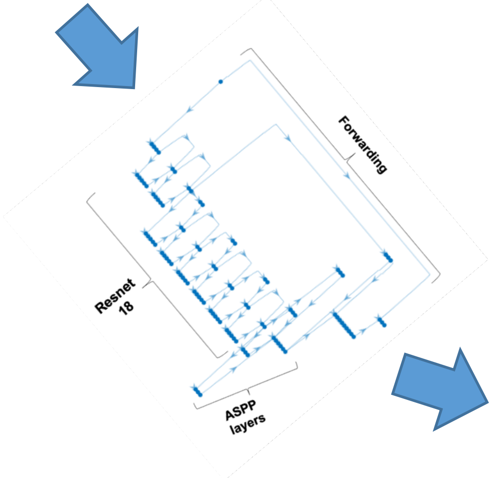
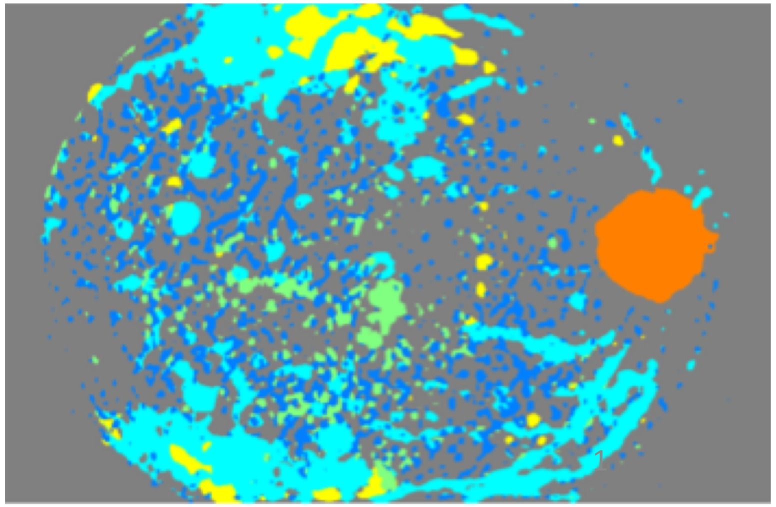
*Deep semantic segmentation of diabetic retinopathy lesions: **what metrics really tell us***







Pedro Furtado @ Medical Imaging 2020

pnf@dei.uc.pt

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

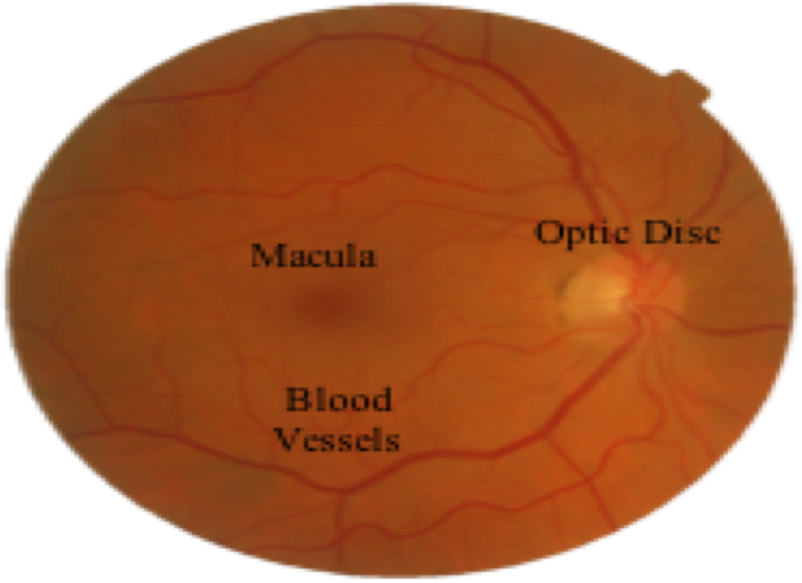
A collaboration with,
Endocrinology Diabetes and Metabolism Department.
Coimbra University Hospital Centre. Portugal

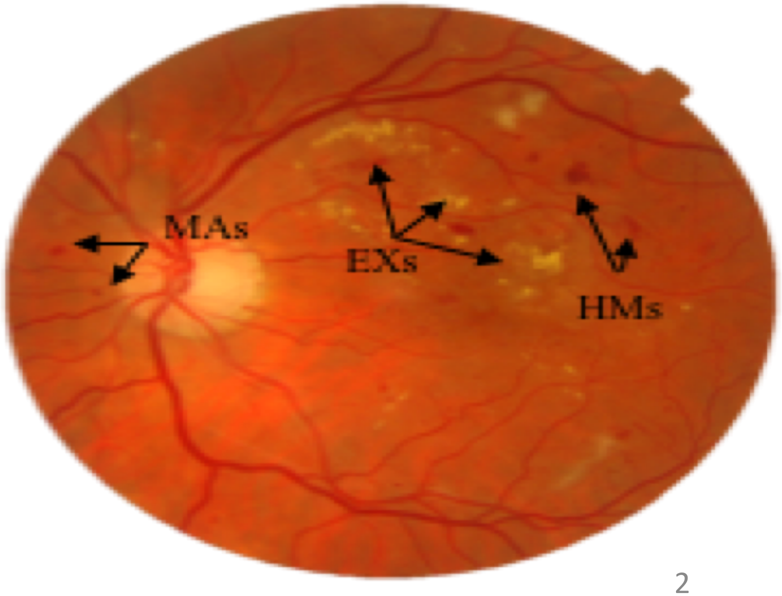
Reviewing the problem...

Diabetic Retinopathy (DR) *is an eye condition related to* ***microvascular changes in the retina*** *that affects people with diabetes.*

*...**leakage of extra fluid** and **small amounts of blood in the eye** (microaneurysms and hemorrhages) and **deposits of cholesterol and other fats** (exudates) [1].*

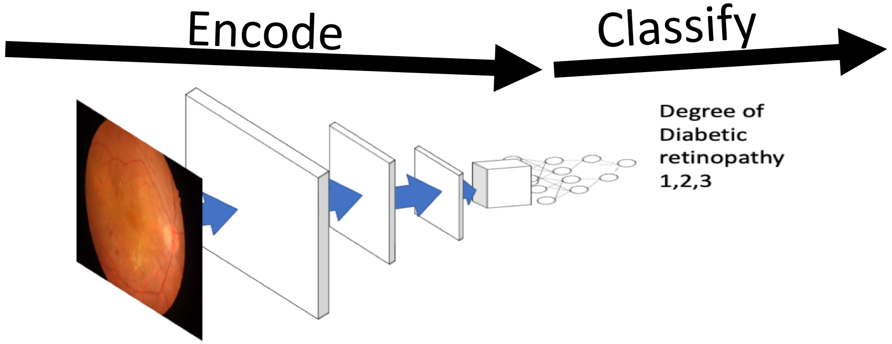
EFI=eye fundus image



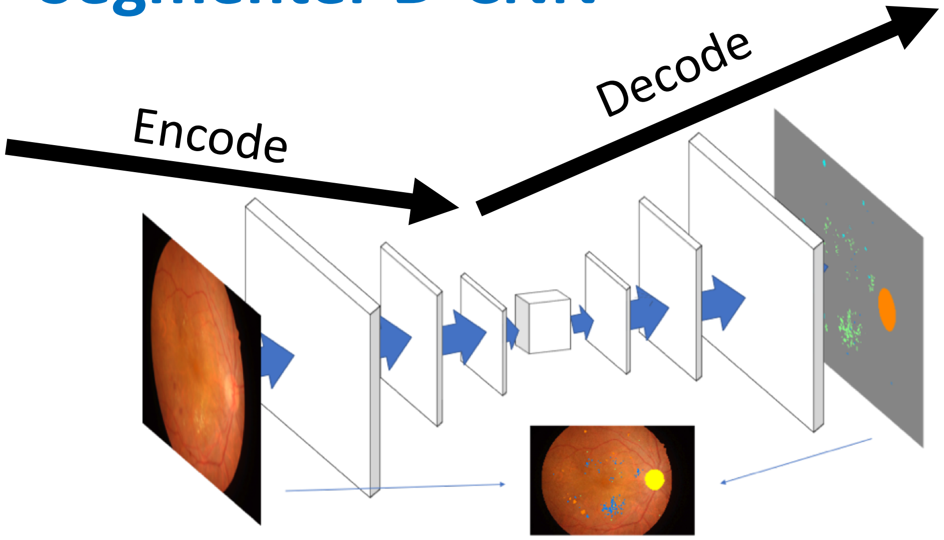


Background (Classification of DR and segmentation of lesions)

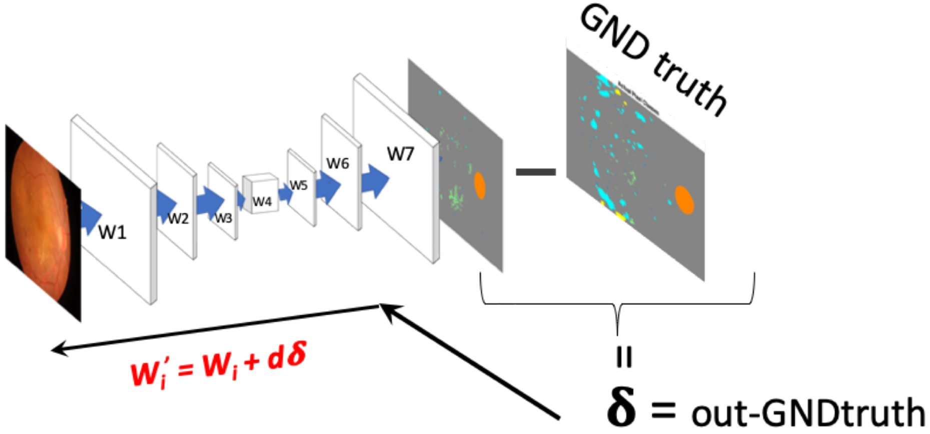
Classifier D-CNN



Segmenter D-CNN

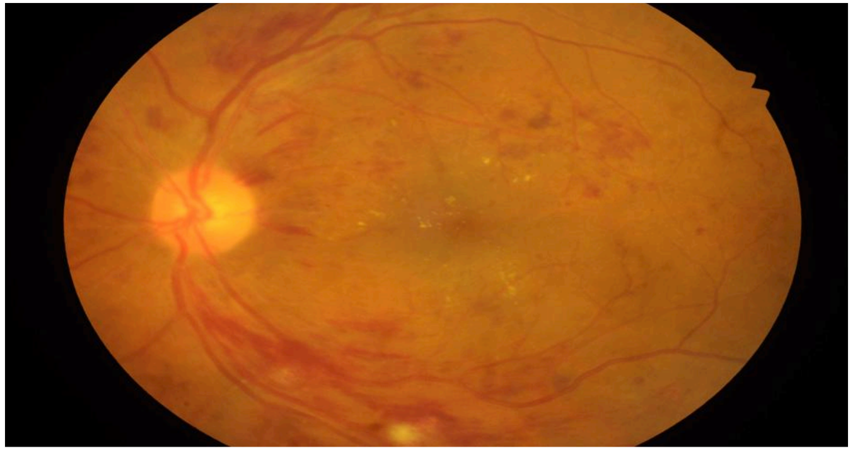


Backpropagation: Ajust hundreds of thousands of weights...

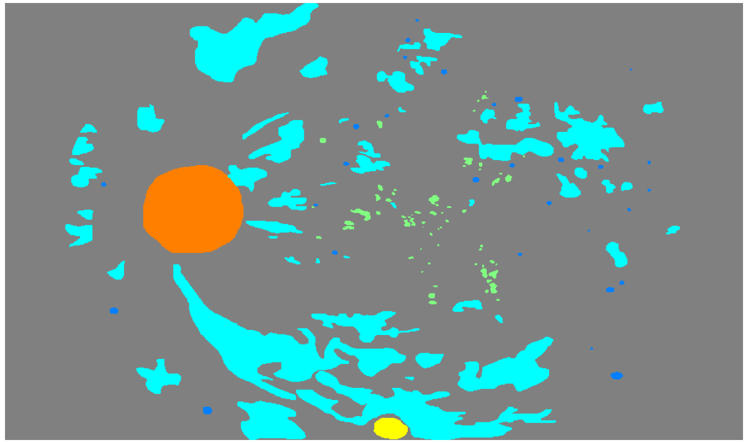


Context *(Segmentation of lesions in eye fundus images EFI)*

- Difficult problem, due to **"very plastic conformation" of lesions**, **small sizes, similarity and lack of contrast**.



Actual Pixel Classes



- **Metrics can be wrongly interpreted, e.g. 90% global accuracy of FCN does not mean it segments lesions very well.**

Some questions…

1. How successful is segmentation of SMALL lesions & LARGER optic disk with standard, off-the-shelf Deep Segm Nets?

2. How successful are different network architectures?

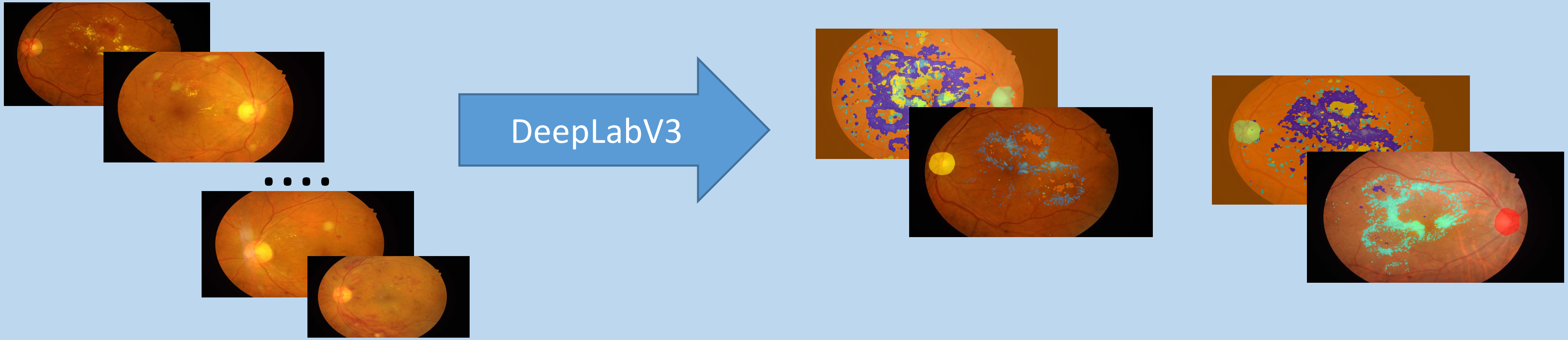
3. How advantageous is it to apply PATCHING on enlarged images? How does a REGION-PROPOSAL method (RCNN) fare?

4. What needs to improve in the future?

Difficulties with Evaluation (Metrics)

- In segmentation, metrics can be deceiving if not fully understood...
- What does each metric really mean? What should we use?

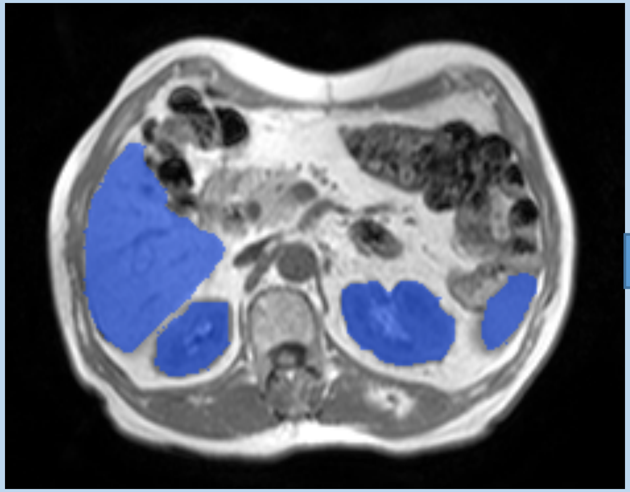
- Global acc, Mean acc were **81 to 84%** ...
- Weighted IoU was **88%**

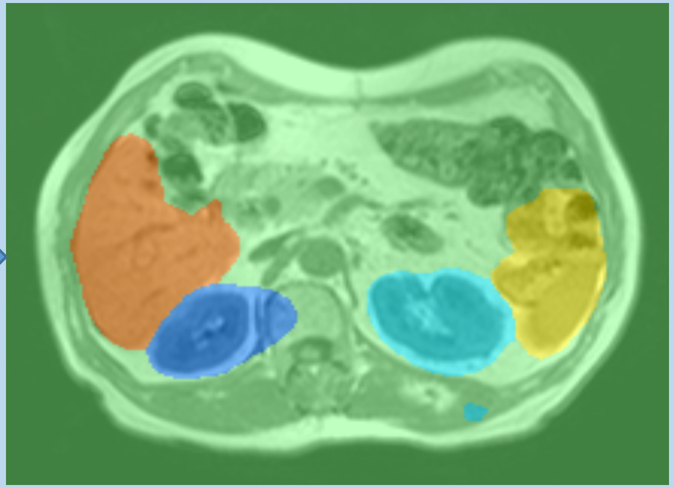


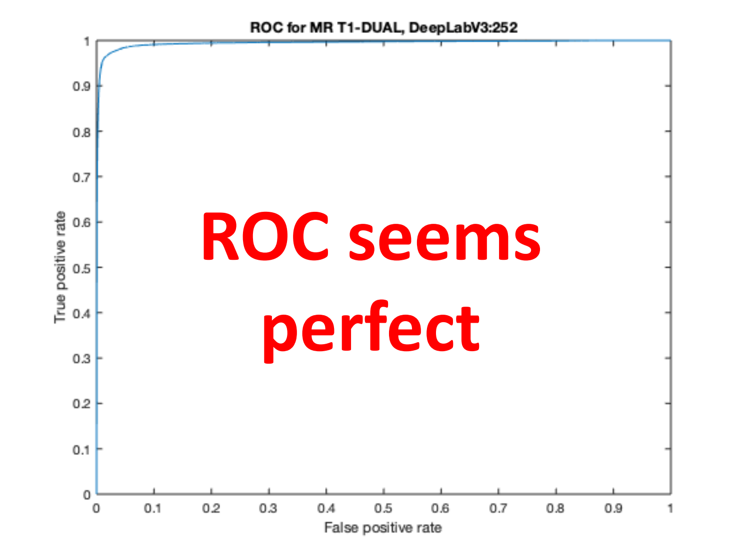
- Actual "quality" of **segmentation of lesions: 2 to 13%...**
- **"quality" ~ % of regions match**

ROC and AUC do not help either...

- AUC over all MRI slices was **91%**







7

- Actual "quality" of segmentation of organs was **12%...**

Methods and dataset…

DIABETIC RETINOPATHY
SEGMENTATION AND GRADING CHALLENGE
IDRiD
INDIAN DIABETIC RETINOPATHY IMAGE DATASET

Investigative Method:

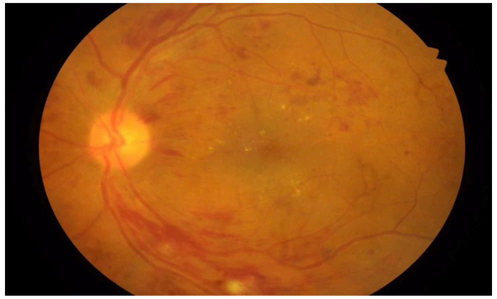
Build Deep Segmentation Networks
Pre-trained

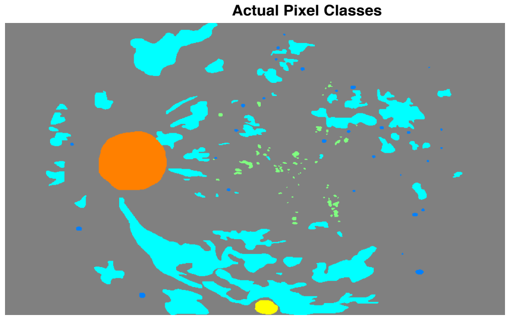
Setup Dataset with gndtruths
Define Evaluation Metrics

Train Seg Nets with images

Evaluate with Test Data

Interpret Metrics to Conclude





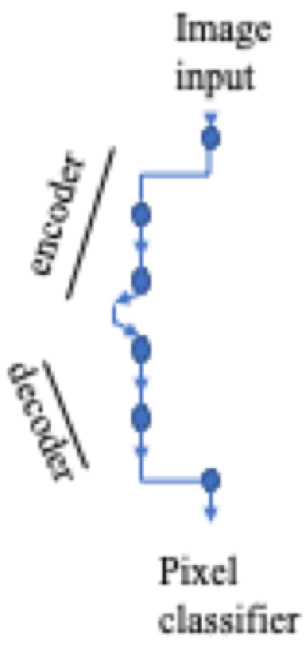
- 83 Eye Fundus Images (EFI) with groundtruth pixelmaps
- Most images have a large number of instances of each specific lesion
- Lesion segmentation task = segment retinal lesions and optic disc as well.

•**Data:** Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe and Fabrice Meriaudeau, "Indian Diabetic Retinopathy Image Dataset (IDRiD)", IEEE Dataport, 2018. [Online]. Available: http://dx.doi.org/10.21227/H25W98.
•**Data Descriptor:** Porwal P, Pachade S, Kamble R, Kokare M, Deshmukh G, Sahasrabuddhe V, Meriaudeau F. Indian Diabetic Retinopathy Image Dataset (IDRiD): A Database for Diabetic Retinopathy Screening Research. *Data*. 2018; 3(3):25. Available (Open Access): http://www.mdpi.com/2306-5729/3/3/25
•**Challenge Summary Paper:** Prasanna Porwal, Samiksha Pachade, Manesh Kokare, Girish Deshmukh, Jaemin Son, Woong Bae, Lihong Liu, et al. "IDRiD: Diabetic Retinopathy–Segmentation and Grading Challenge." Medical image analysis 59 (2020): 101561. DOI: https://doi.org/10.1016/j.media.2019.101561

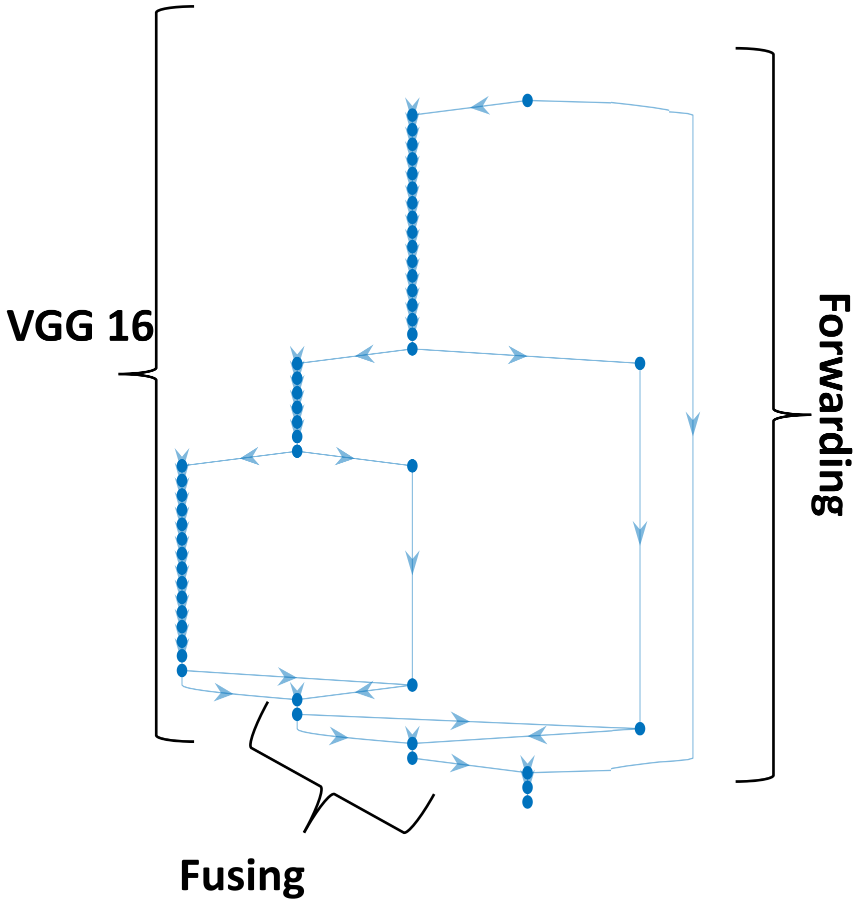
The networks….

A simple encoder stage	●	= [conv + relu + maxpool (to DNsample)]
a simple decoder stage	●	= [transposed conv (to UPsample-2x) + relu]

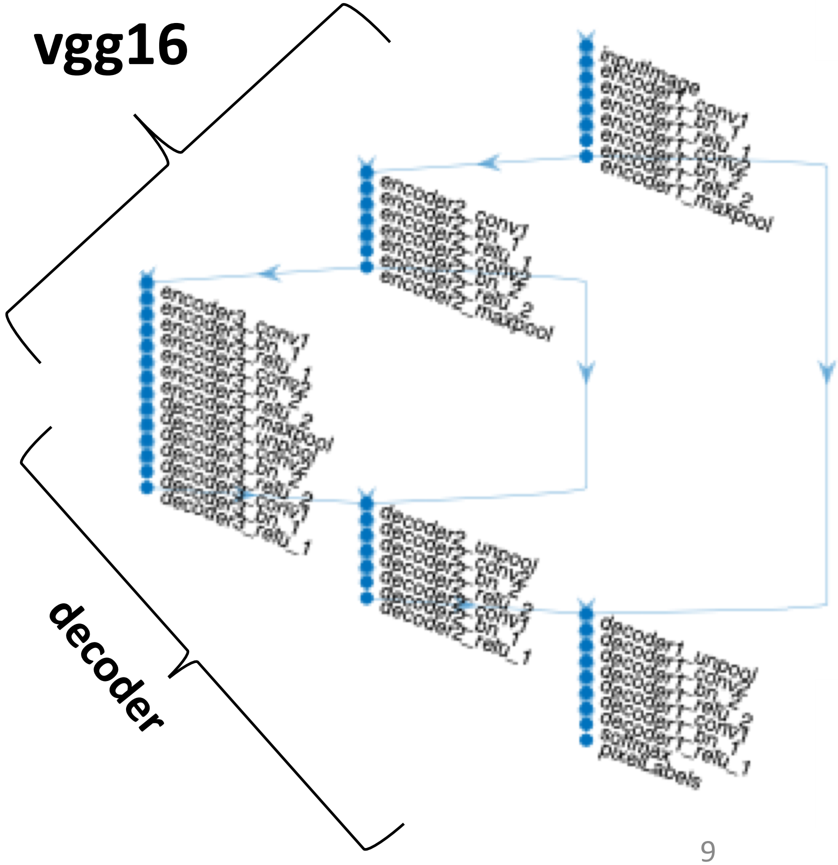
Simple: ~10 layers



FCN: ~50 layers

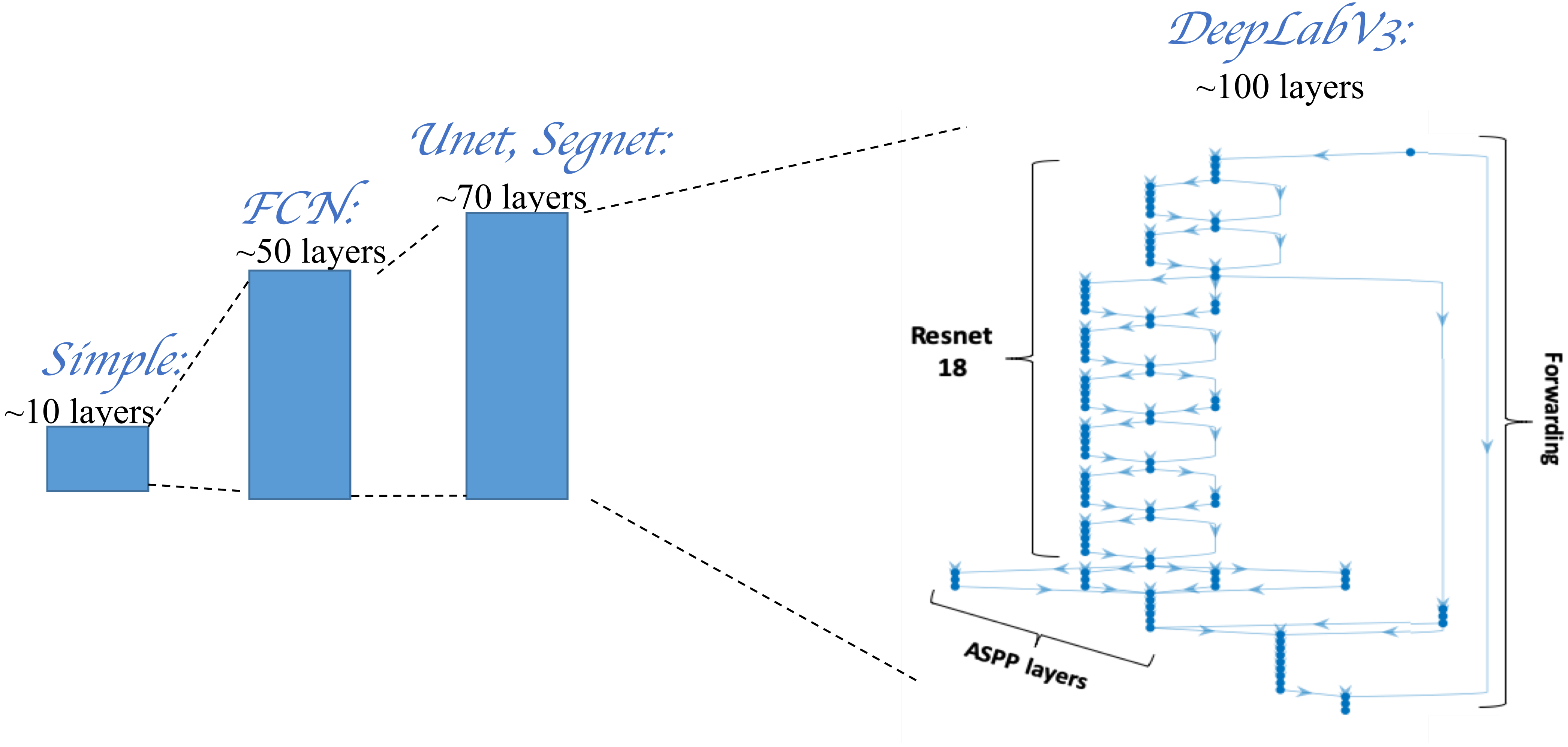


Unet, Segnet: ~70 layers



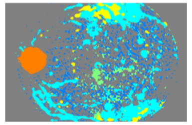
The networks….

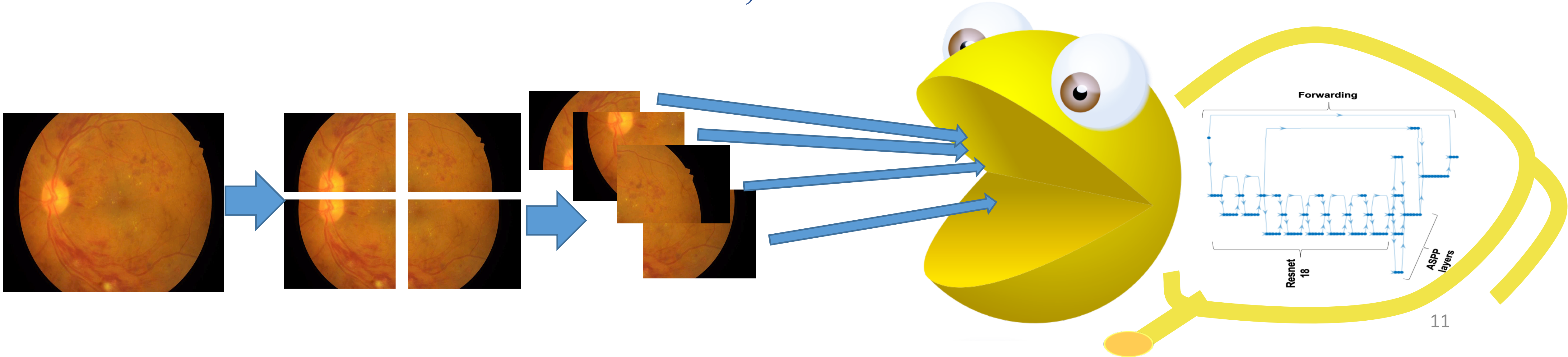
A simple encoder = [conv, relu, maxpool to DNsample]
a simple decoder = [transposed conv to UPsample-2x, relu]



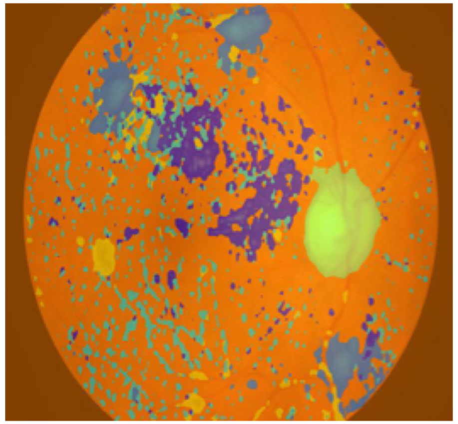
Patching...

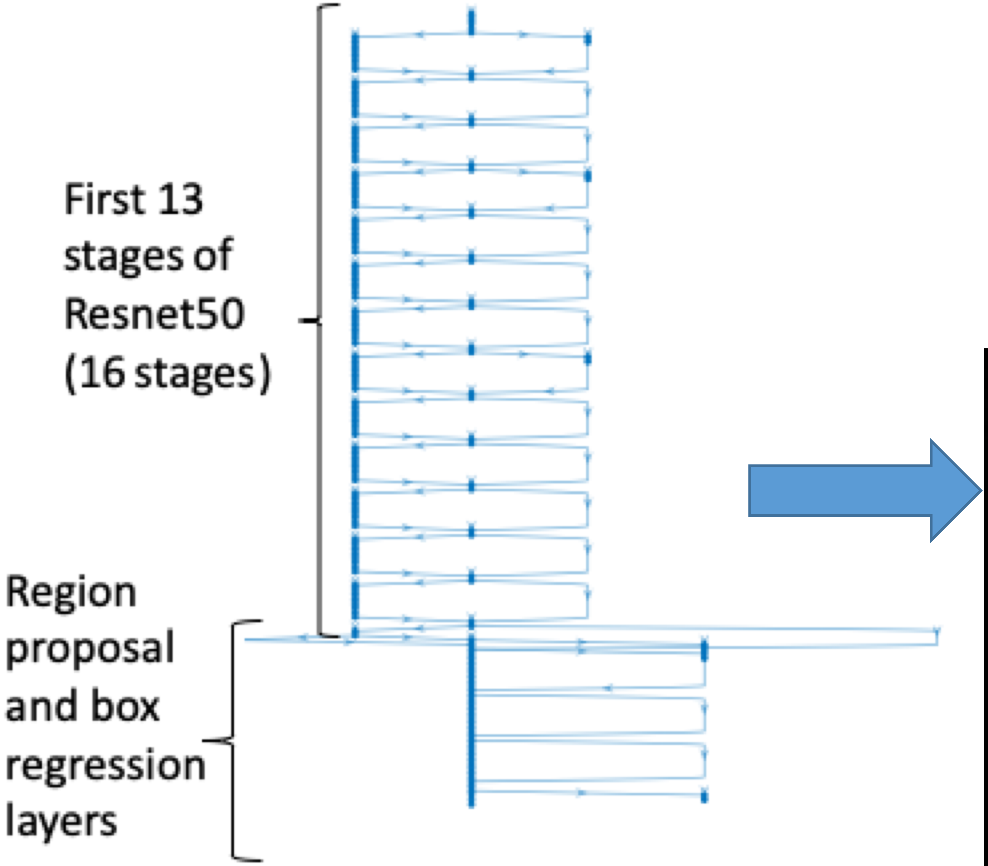
- Original Images are too large to fit a minibatch confortably into GPU memory (4096x2048)
- Solution = they were resized to ~1/4 (2048x1024)
- Do we loose segmentation quality by reducing size so much?
- WE COMPARE WITH no size redux, PATCHING

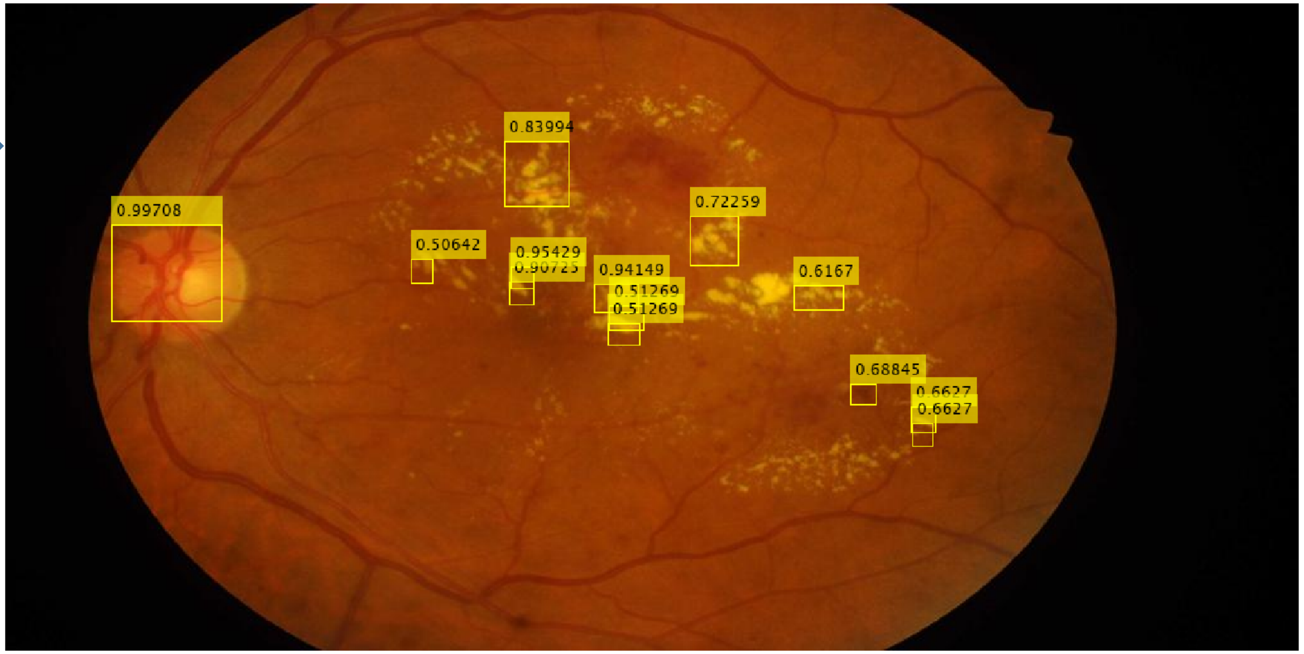




Also test region-based segmentation…





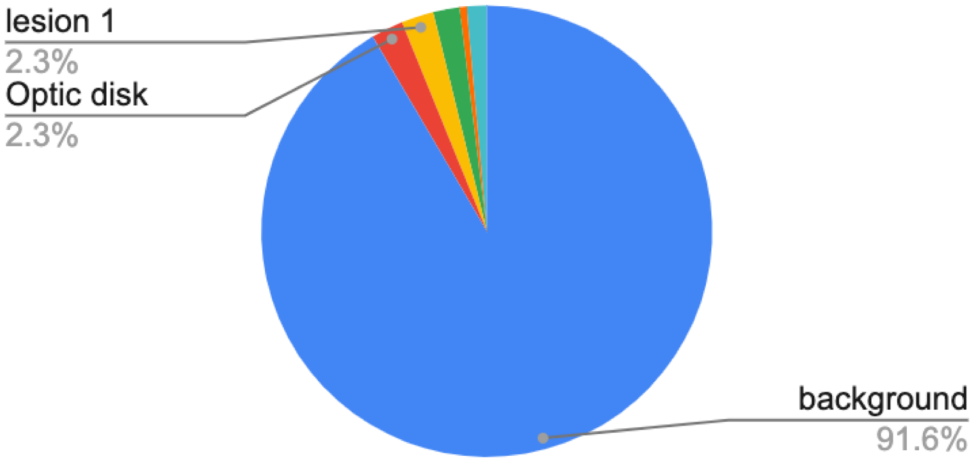


The metrics... And weight balancing

- In the paper we report and analyze all relevant common metrics

- We added weight balancing to all pixel classification layers
  - *To counter class imbalance...*

Most pixels are background...

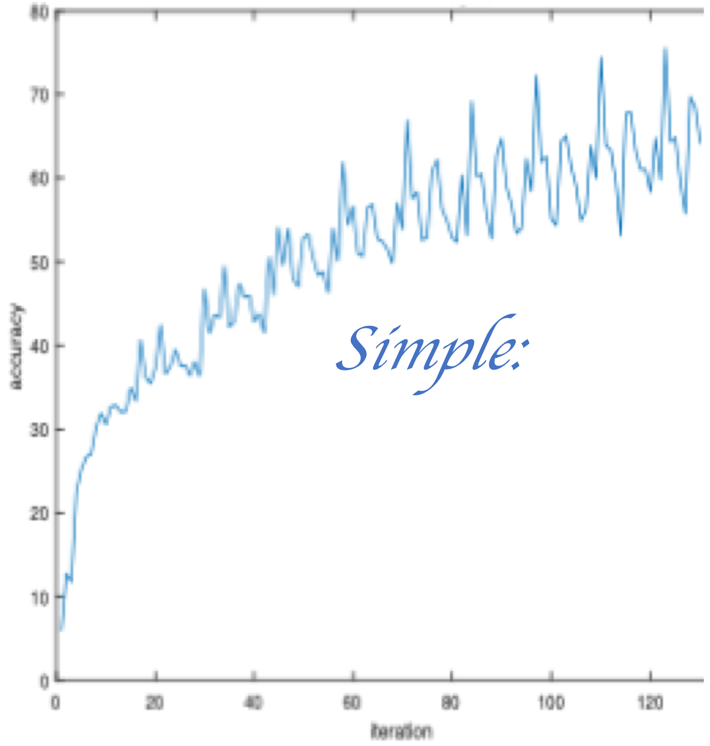


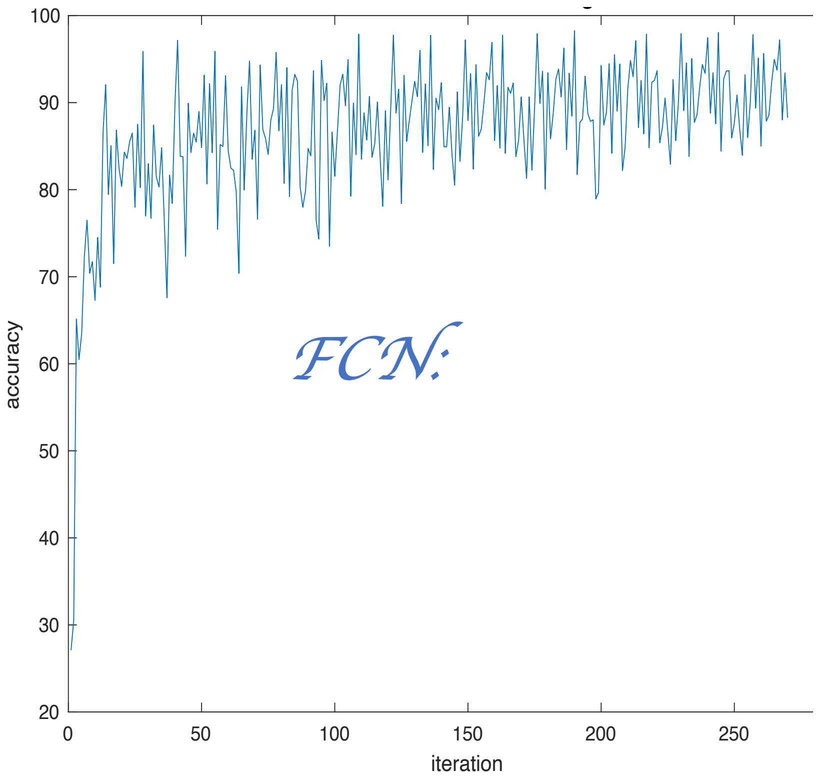
Training accuracy (evolution)....

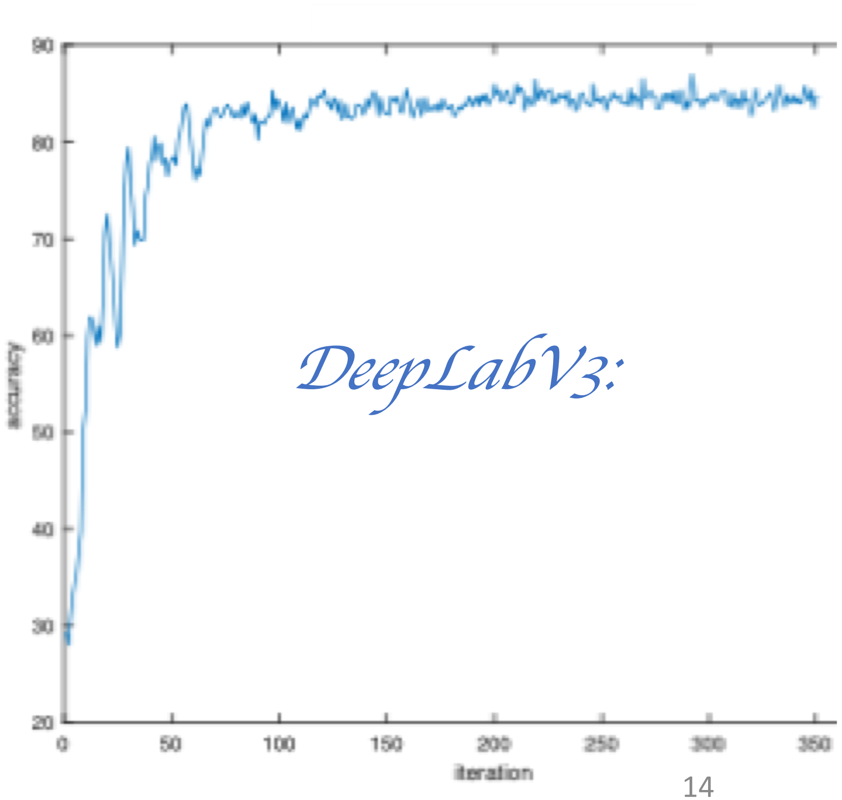
- Simple had more difficulties converging to a high accuracy...

  FCN and DeepLab converged better to high accuracy...

  had to adjust FCN learning rate



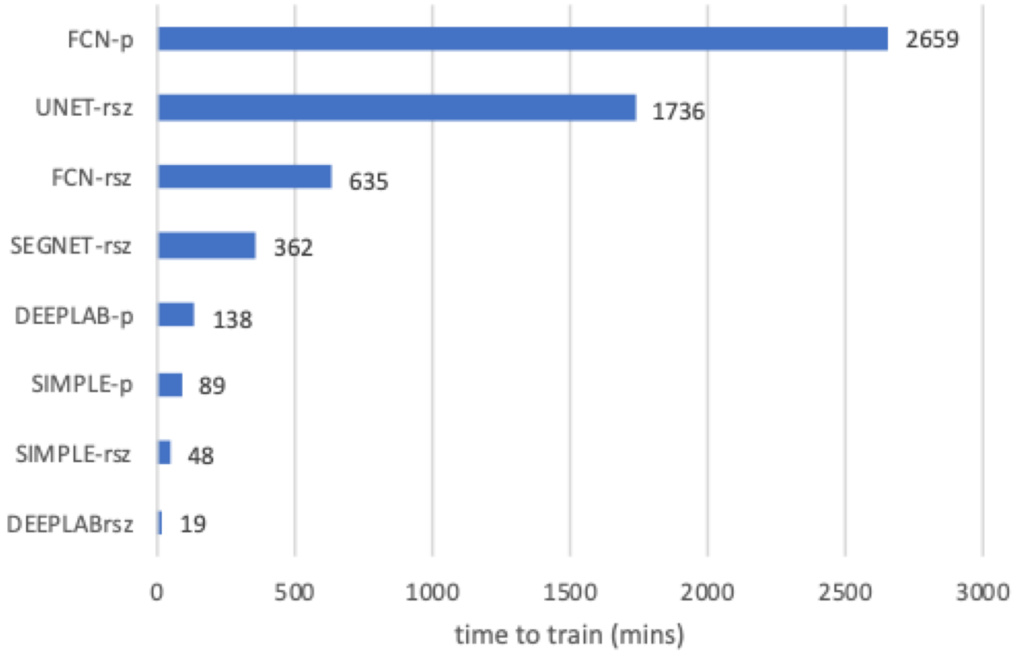


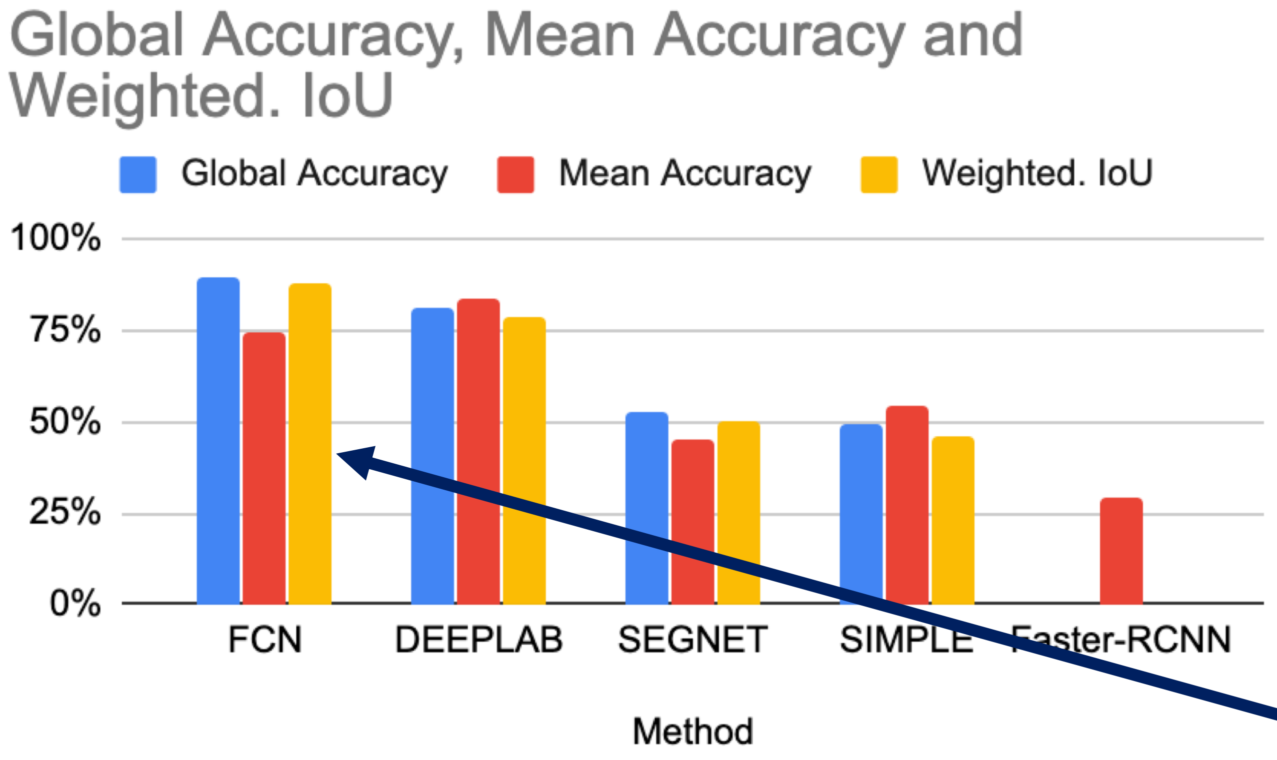


Training times….

P=patching
Rsz= resize

- DeepLabV3 and Simple fastest converging (19 mins, 48 mins)
- FCN and UNET are slowest, 80x slower than deeplabV3
- Train times with Patching are 2 to 5 times larger (more data)





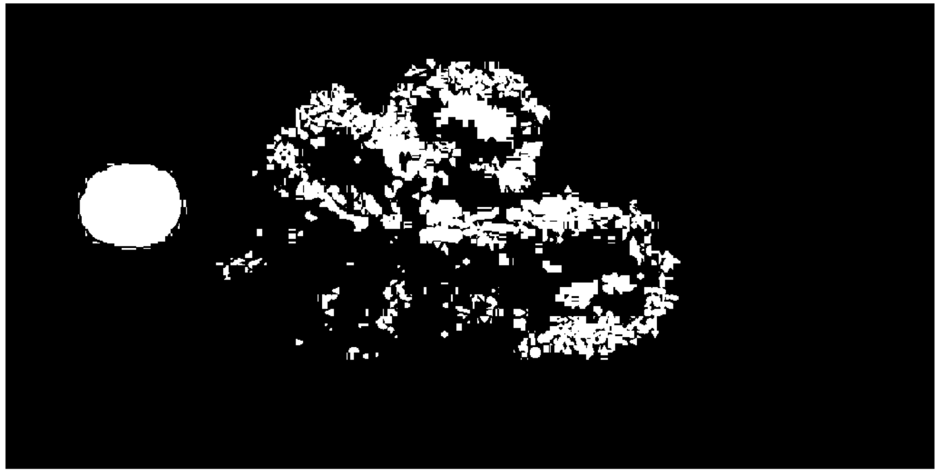
Method	Global Accuracy	Mean Accuracy	Weighted. IoU
FCN	**90%**	**75%**	88%
DEEPLAB	**81%**	**84%**	79%
SEGNET	53%	45%	50%
SIMPLE	49%	55%	46%
Faster-RCNN	-	30%	

IoU

- **FCN** very good **accuracy and IoU (90%, 88%)**
- DeepLabV3 quite good, always > 75%
- **Huge improvement over SIMPLE, Segnet (25 to 40% better)**
- R-CNN seems much worse = 30%

Pictorially, FCN case...

- **Actual GND** pixels of lesions & OD



- **Lesions (and OD) NOT Detected=2.2%**

Per-class Toolbox output conf matrices (~ 60 to 97%)

DeepLabV3

Confusion Matrix (%):

True Class \ Predicted Class	Background	Microaneurysms	Haemorrhages	HardExudates	SoftExudates	OpticDisc
Background	81.11	7.077	4.882	5.311	1.009	0.6128
Microaneurysms	5.116	82.99	10.04	1.822	0.03504	0
Haemorrhages	21.37	11.08	64.71	1.369	1.242	0.2356
HardExudates	1.386	0.6664	0.03913	97.31	0.5835	0.01726
SoftExudates	9.815	3.143	3.153	12.29	70.33	1.268
OpticDisc	2.268	0.09411	0.0289	0.3757	0.1927	97.04

FCN

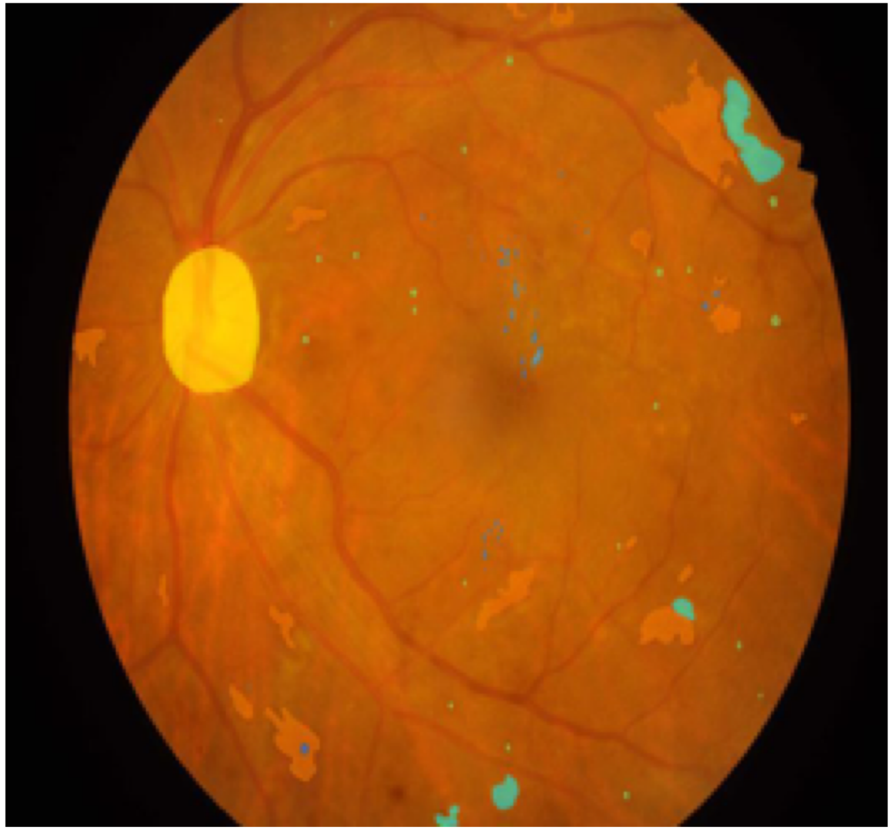
Confusion Matrix (%):

True Class \ Predicted Class	Background	Microaneurysms	Haemorrhages	HardExudates	SoftExudates	OpticDisc
Background	89.88	3.487	1.828	4.15	0.2712	0.3861
Microaneurysms	16.52	62.95	17.72	2.466	0.1772	0.1624
Haemorrhages	26.7	10	58.01	4.31	0.6087	0.3708
HardExudates	3.39	0.9354	6.272	80.42	6.15	2.834
SoftExudates	14.94	2.61	4.103	10.26	61.01	7.073
OpticDisc	3.325	0.1127	0.6245	0.4694	0.1571	95.31

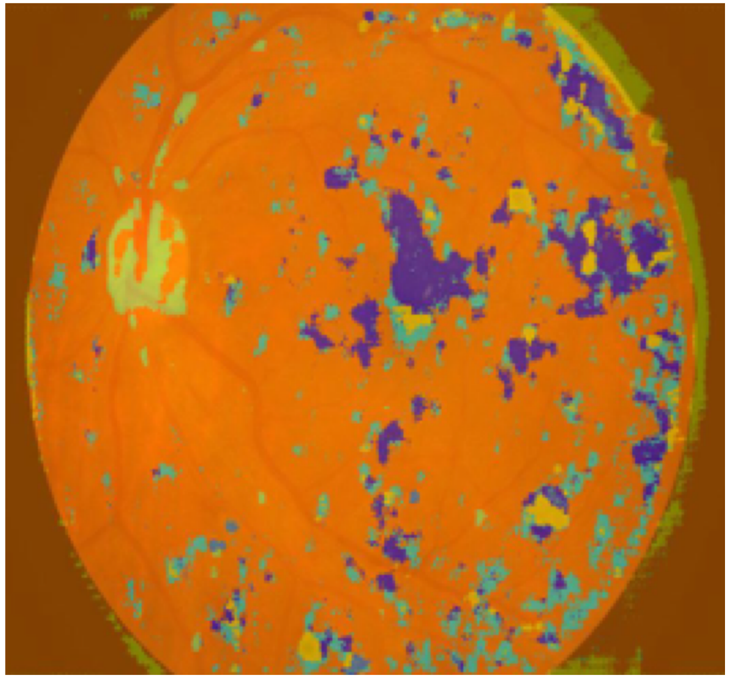
But, visually….

Evidence 1: there seem to be some problems...

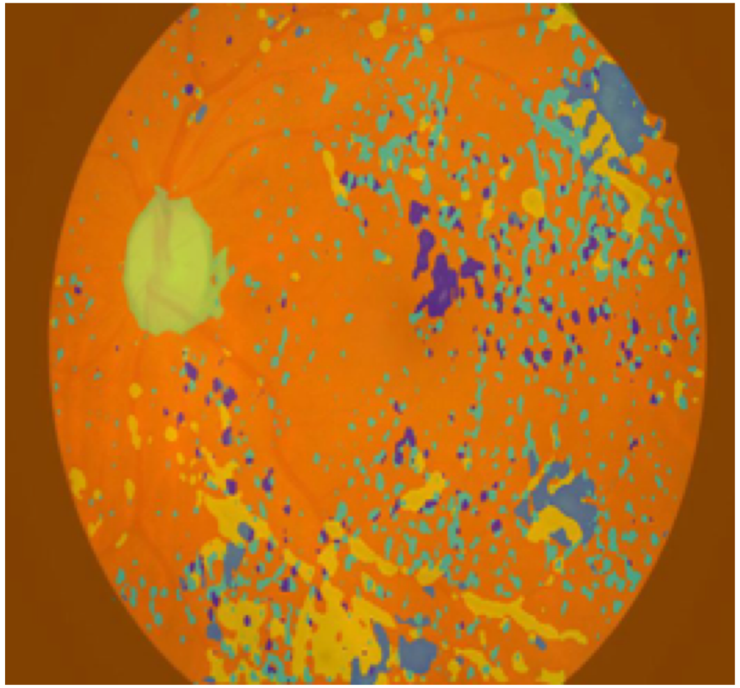
Groundtruth



FCN

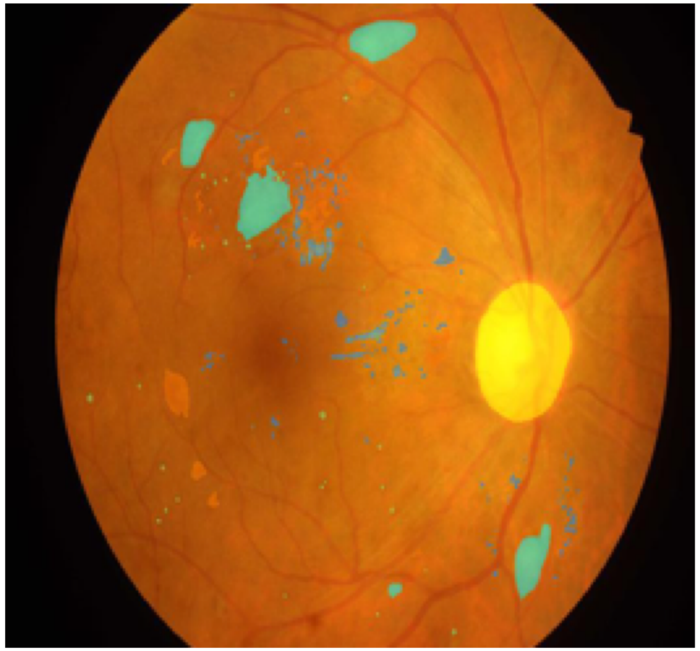


DeepLabV3

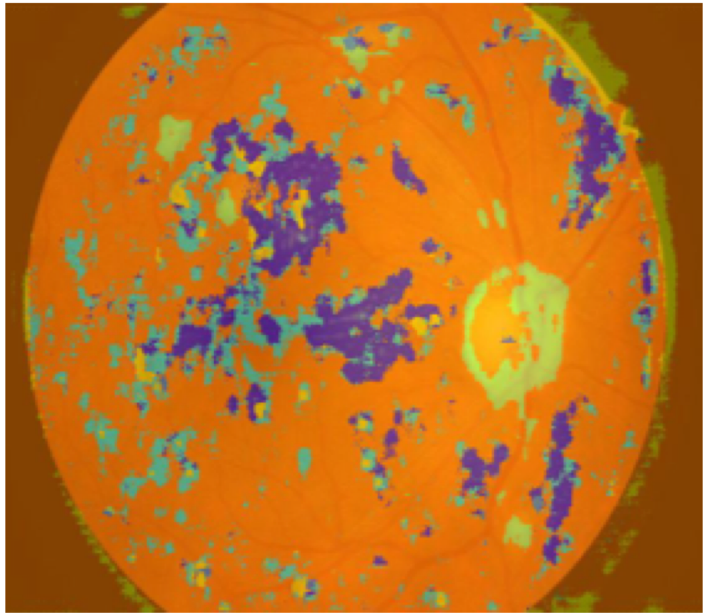


Evidence 2: same problem…

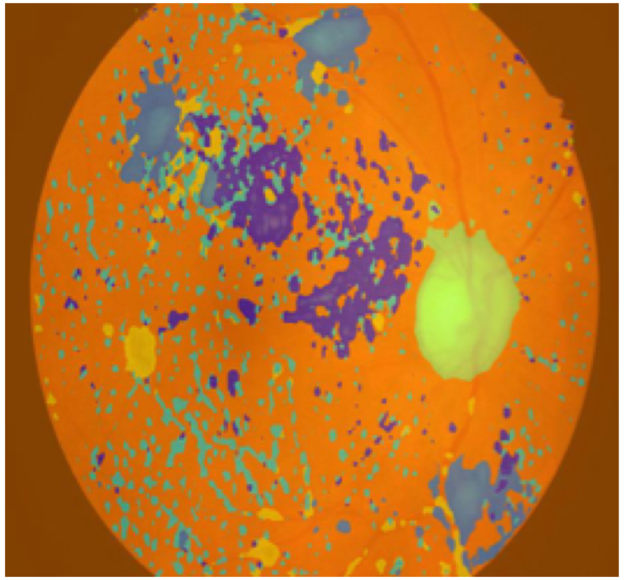
Groundtruth



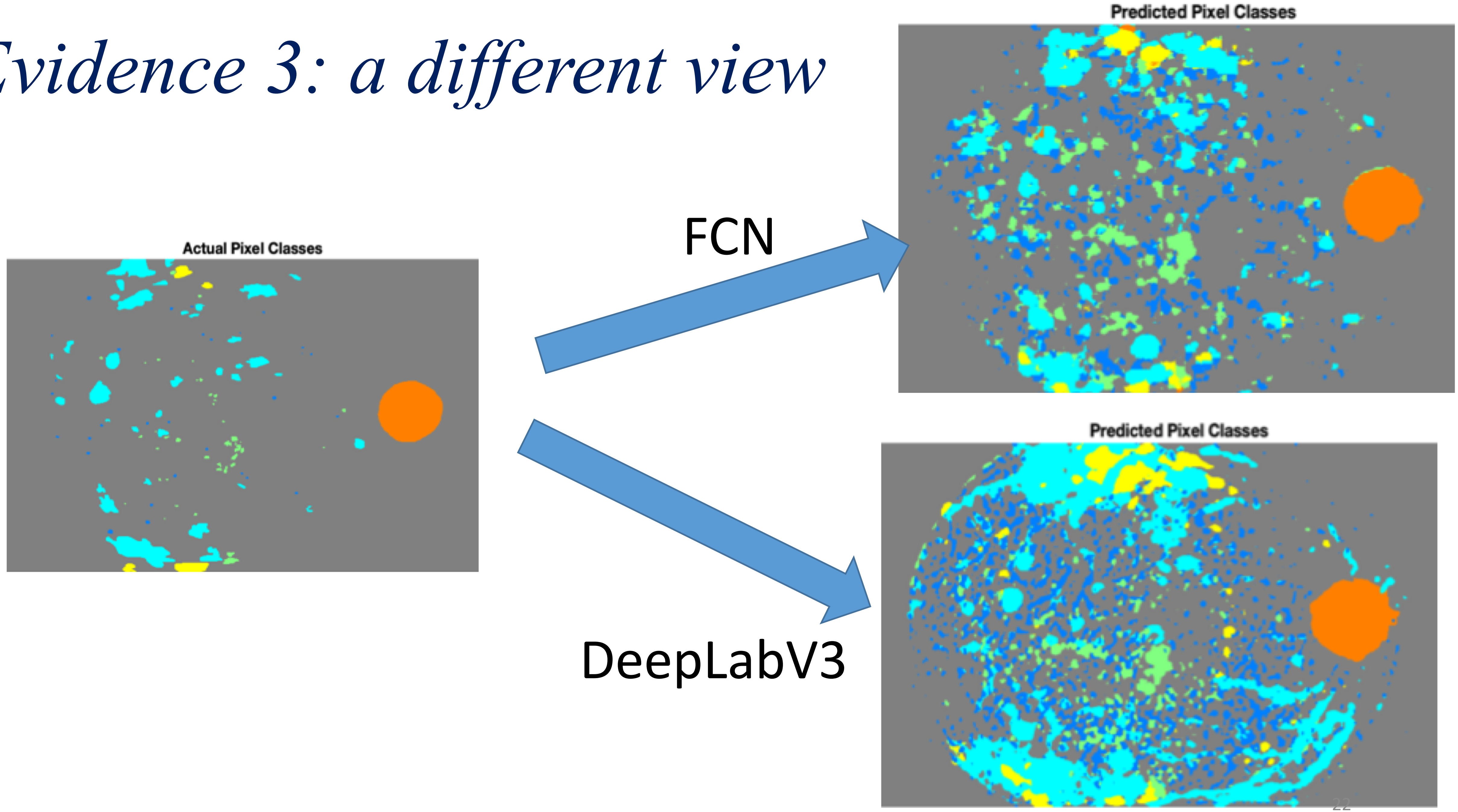
FCN



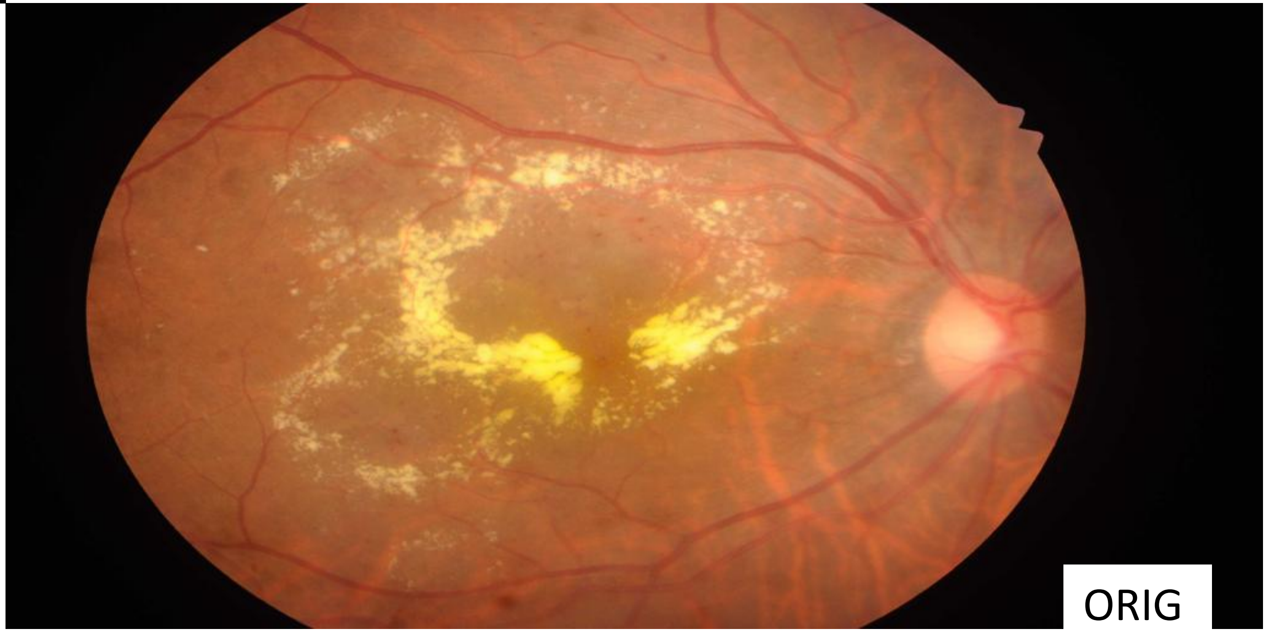
DeepLabV3



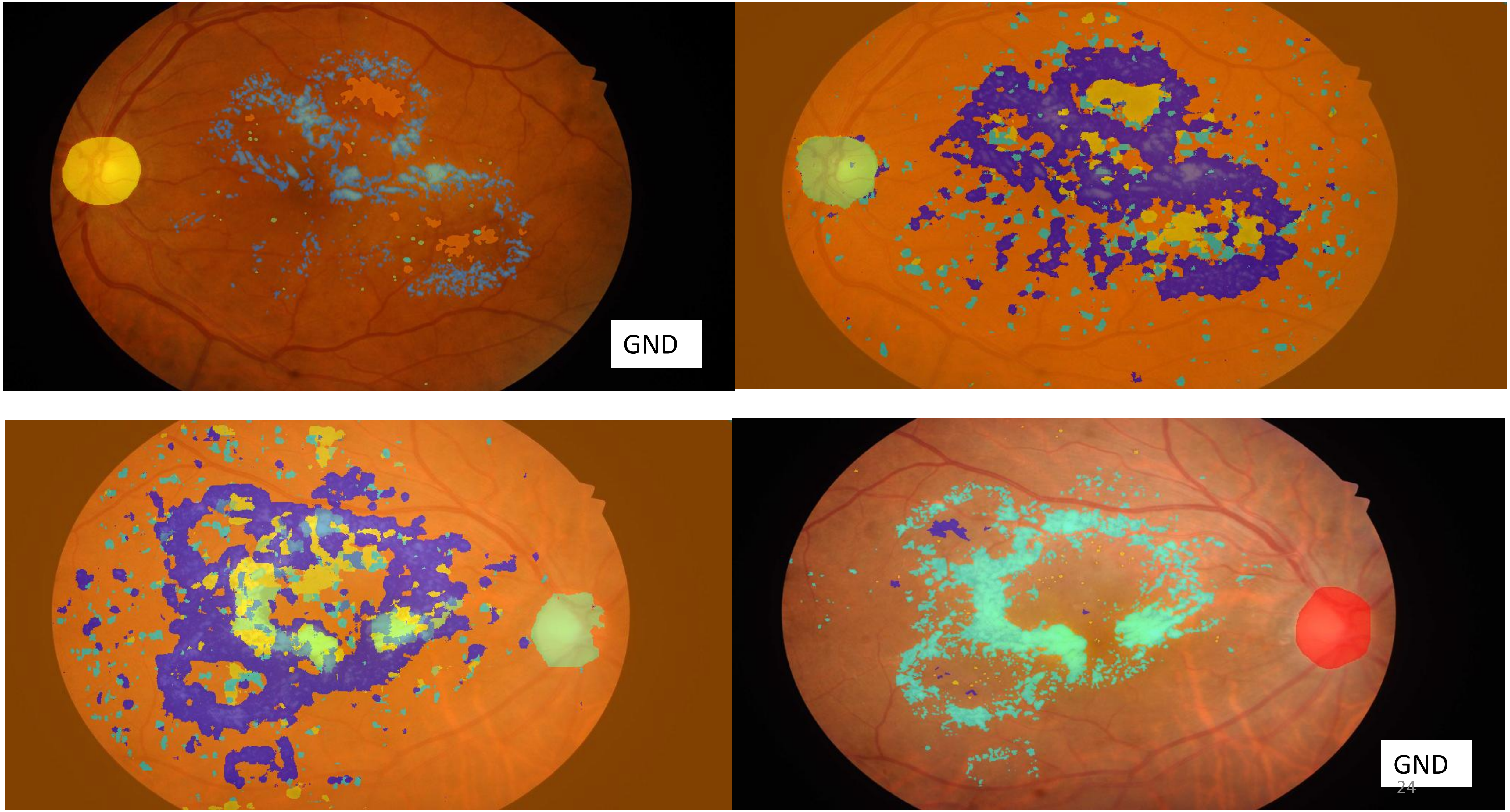
Evidence 3: a different view



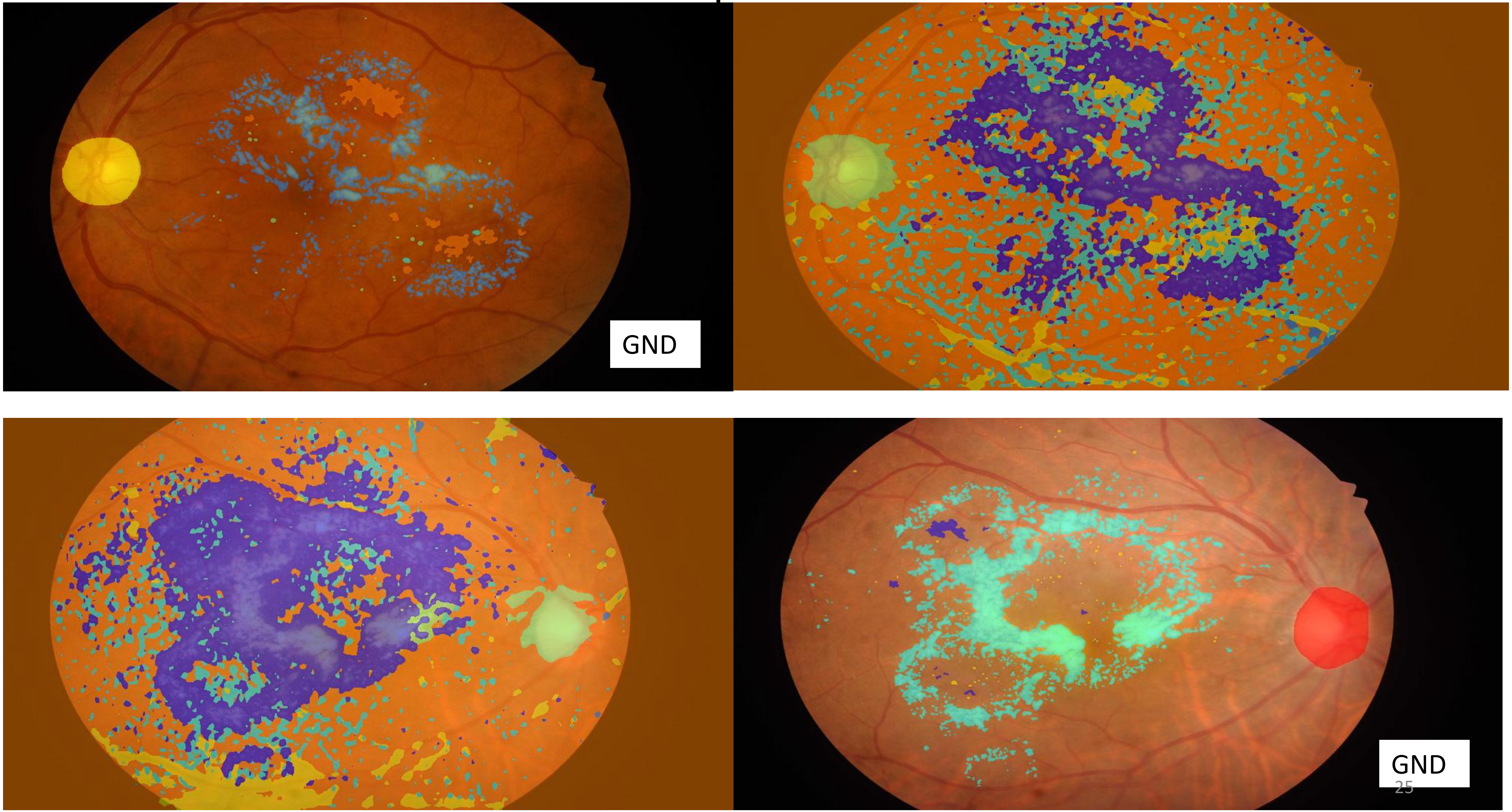

ORIG


ORIG

FCN



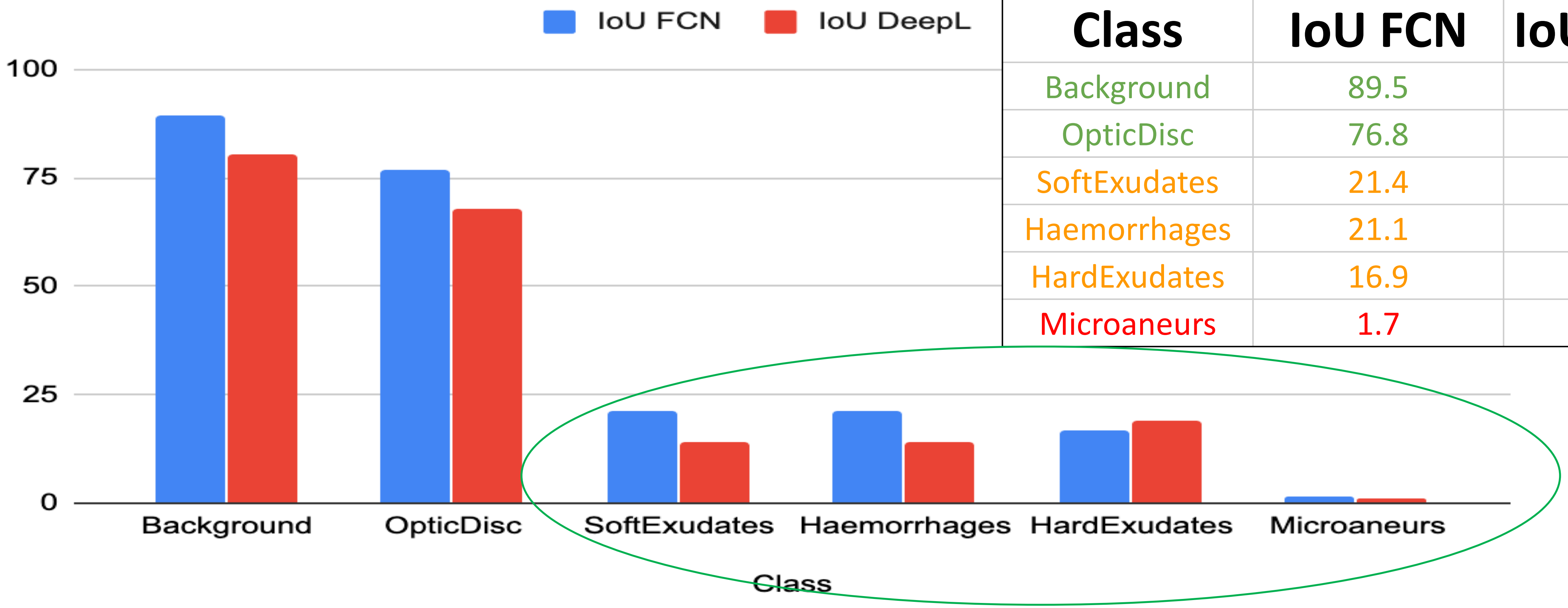
deeplabV3



So far, I WOULD SAY quantitative results do not match Visualizations

So, let’s analyze quantitatively in some more detail…

Per-class IoU FCN and IoU DeepL



Class	IoU FCN	IoU DeepL
Background	89.5	80.6
OpticDisc	76.8	68
SoftExudates	21.4	14
Haemorrhages	21.1	14.3
HardExudates	16.9	19.1
Microaneurs	1.7	1

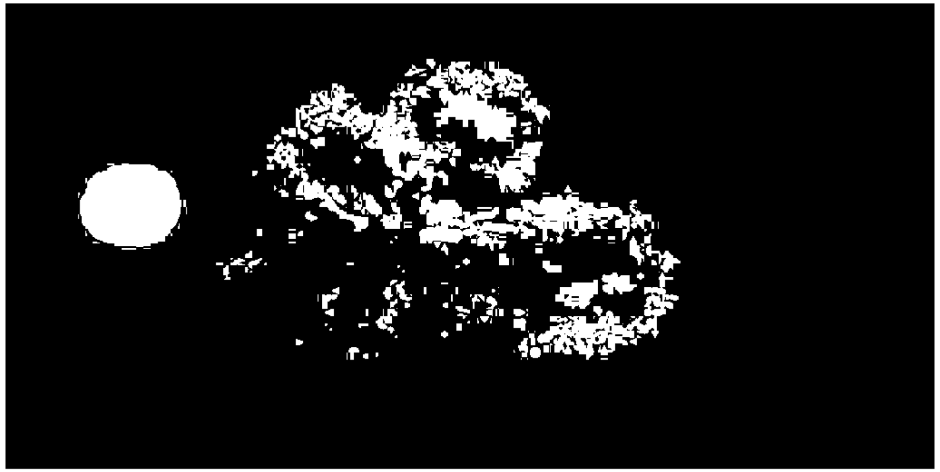
- Per-class IoU reveals the deficiencies...

  e.g. FCN weighted IoU 88%, BUT IoU of individual lesions only 1 to 21%

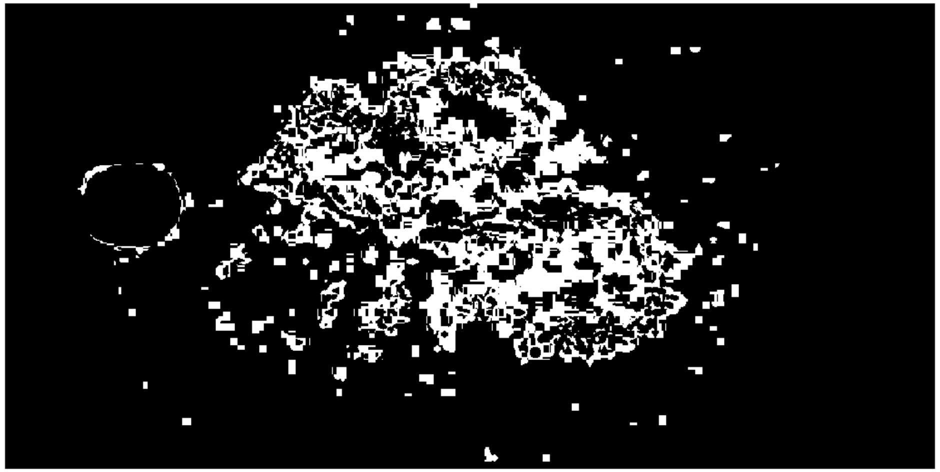
- CONCLUSION: Only the background and the optic disk are well segmented

2. Lesions false positives

- GND Pixels of lesions and



- **Bkground pixels wrongly** classified as lesions/OD ...



= ~11% of all pixels

=136% of all lesion pixels

Finally, we changed loss function of DeepLabV3

- From crossentropy
- To...

IoU

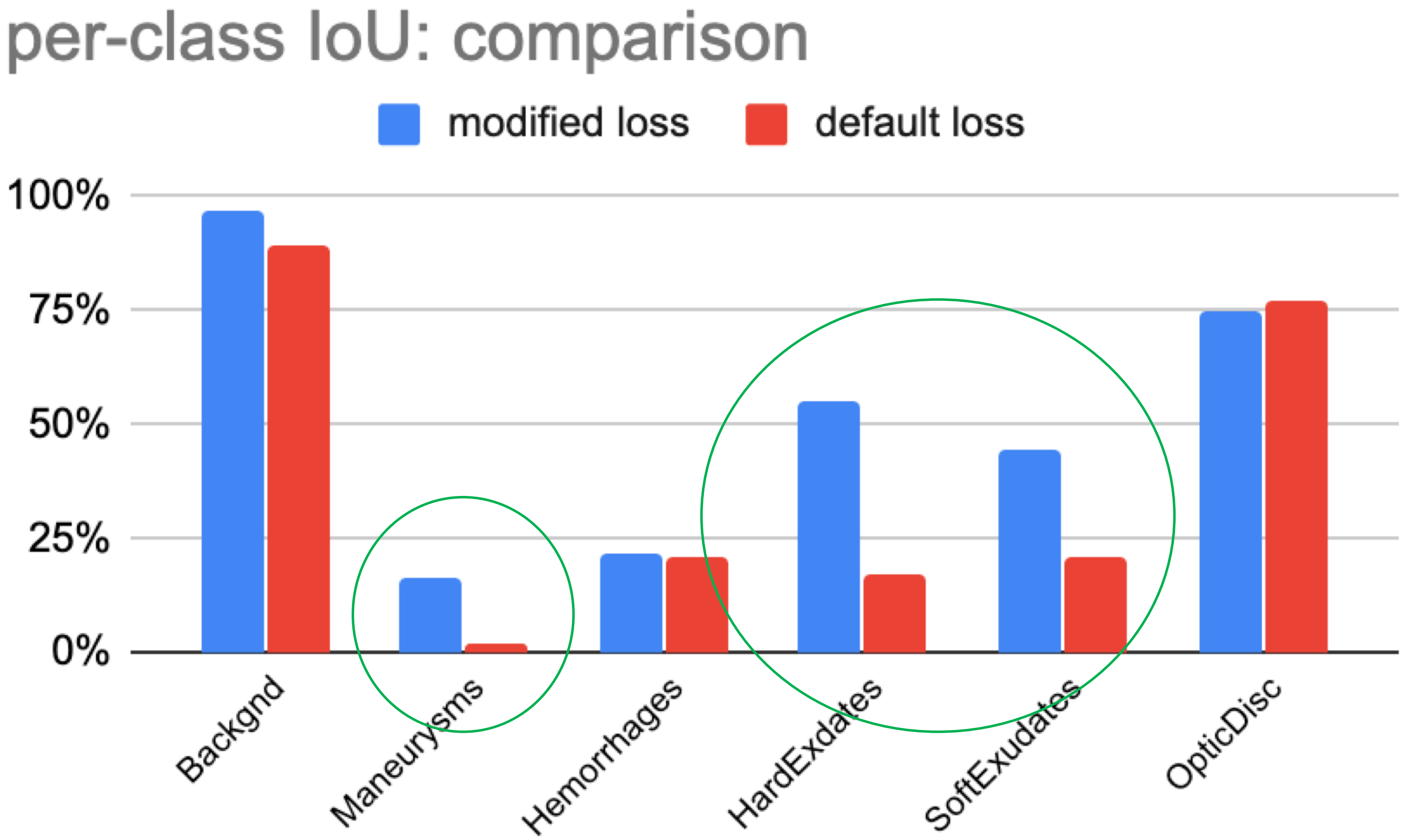
- **IoU of class** = degree of “exact matching” of regions

  IoU(c)= TPc / (TPc + FNc + **FPc**)

- **Loss function = IoU weighted on inverse class frequencies**

Per-class IoU

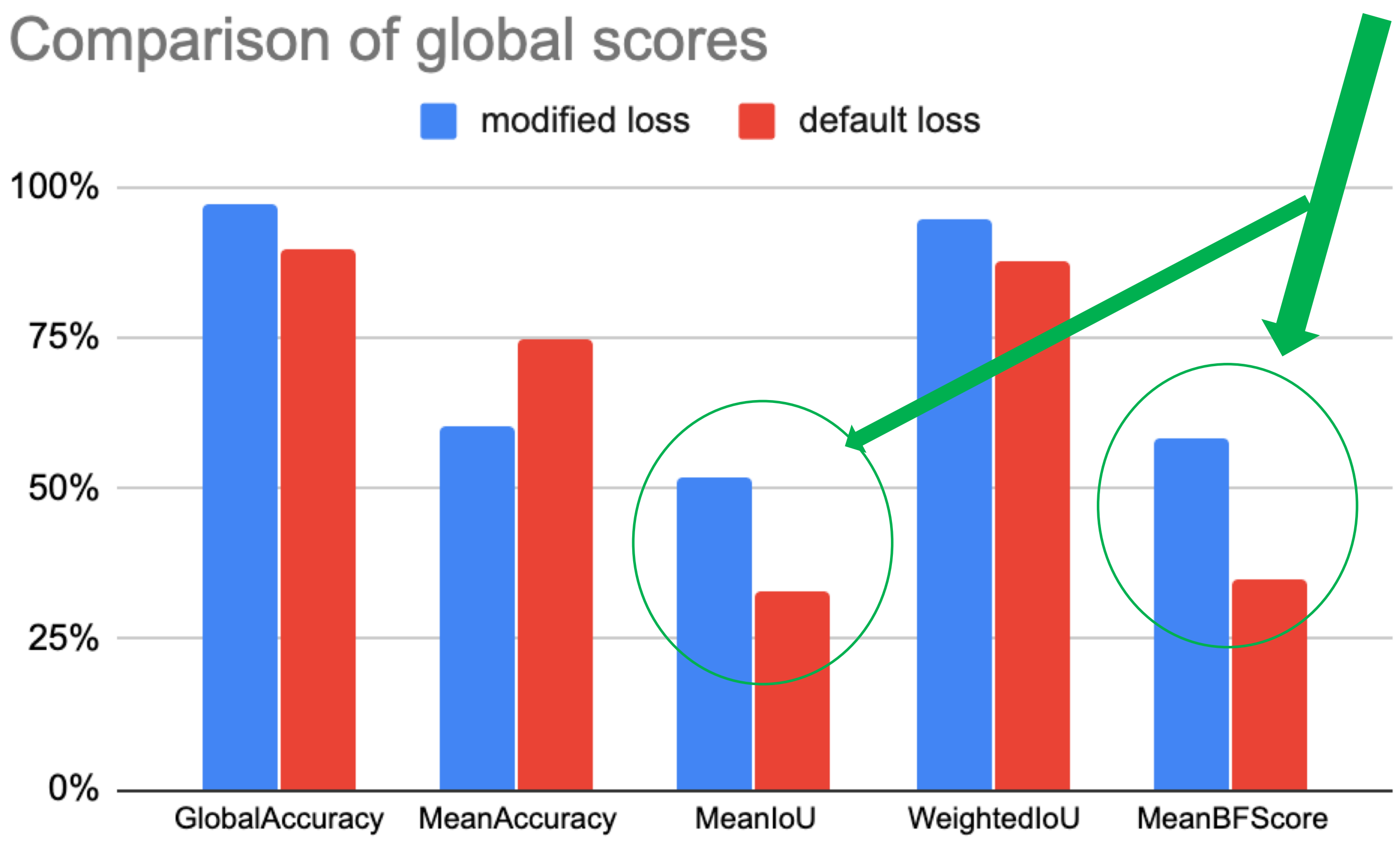
Modified loss (IoU) vs default (crossentropy)



Global scores

Modified loss (IoU) vs default (crossentropy)

Very relevant improvements



Conclusions

- Deep segmentation networks are amazing, they can learn to segment...
- FCN and DeepLabV3 seemed quite accurate (IoU,acc) (88 to 95%), but...
- Significant number of BKGROUND pixels were classified as lesions
  - *Quality of segmentation of Micro-aneurisms given by IoU is 1 to 2%*
  - *Quality of segmentation of other lesions given by IoU is 14 to 21%*

- Using IoU as loss function improved significantly...
- But we still need further improvements
  - *Quality of segmentation of Micro-aneurisms and Haemorrhages given by IoU is ~20%*
  - *Quality of segmentation of other lesions given by IoU is 45 to 60%*

Future work

- **Can we successfully add/modify details in deep segmentation networks for better results?**
  - ***Specific new architectural features***
  - ***Further experiments with modification of loss functions***
  - ***More data? already tried augmentation, loss function seems better try***
- **Can we add post-processing to filter false positive lesions (bkgrnd?)**
  - ***Traditional machine learning pipeline together with deep learning***

Final acknowledgments

Acknowledgments:

We would also like to acknowledge the collaboration of the Endocrinology Diabetes and Metabolism Depart of Coimbra University Hospital Centre.

Acknowledgments:

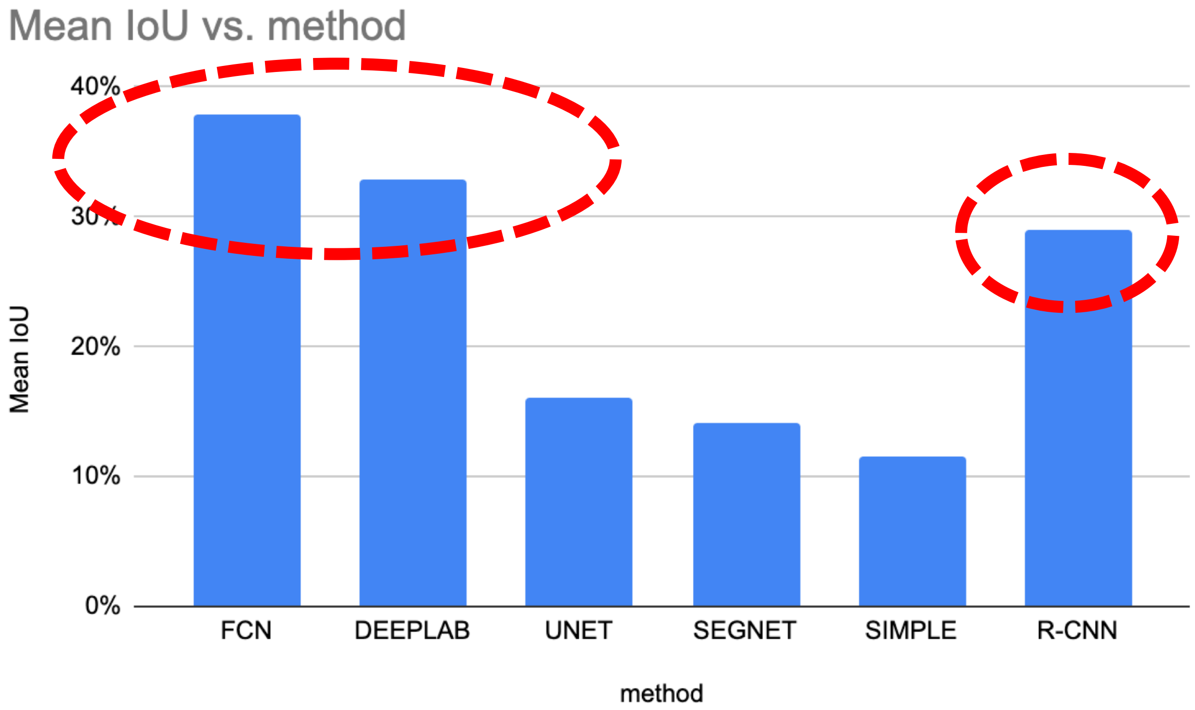
IDRiD challenge dataset for this work [3]. We would like therefore to thank the IDRiD challenge organizers for sharing the dataset and making this work possible.

- Pedro Furtado, University of Coimbra, Portugal

pnf@dei.uc.pt, eden.dei.uc.pt/~pnf

Region-based (R-CNN) is not "much worse" if we do not take BKGND into the equation...

- If we invest more in R-CNN, I think we can get similar to SS
- **Conclusion:** not better, but deserves another look



method	Mean IoU	Mean BF Score
FCN	38%	49%
DEEPLAB	33%	34%
UNET	16%	20%
SEGNET	14%	18%
SIMPLE	12%	19%
R-CNN	29%	-

Patching vs resizing to 1/4

- Patching was worse for DeepLabV3, similar for FCN...
- ... In the BF-Score patching was 5 to 10% better
- **Conclusion:** also deserves another look

	mean IoU		mean Acc		mean BF
DeepLab	**33%**		**84%**		34%
DeepLabPATCH	24%		70%		**44%**
FCN	38%		**75%**		49%
FCN-patch	39%		72%		**53%**

Some hints on formulas…

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

  acc= (TP+TN)/ALL

  Background is BIG

- **Accuracy of object** = recall = fraction of correct classifs of pixels of object

  acc(c) = recall(c) = TPc/(TPc+FNc) **I (lesion) segment well my pixels, but FPc?**

- **IoU** = degree of "exact matching" of regions = ratio of pixels of object well classified by all pixels of object + pixels of other objects also classified as this object

  IoU(c)= TPc / (TPc + FNc + **FPc**) **Adding importante measure (FPc)**

- **BF-Score** = degree of matching of boundaries (within a defined threshold)

  Fair enough, if boundaries are short distance, its ok, but what dist?

Loss as IoU

- Loss metric is now very diferent from accuracy... E.g. acc 97% with loss 60%
- But the results did not improve...
- And, with validation dataset, noted overfitting... More data also needed?