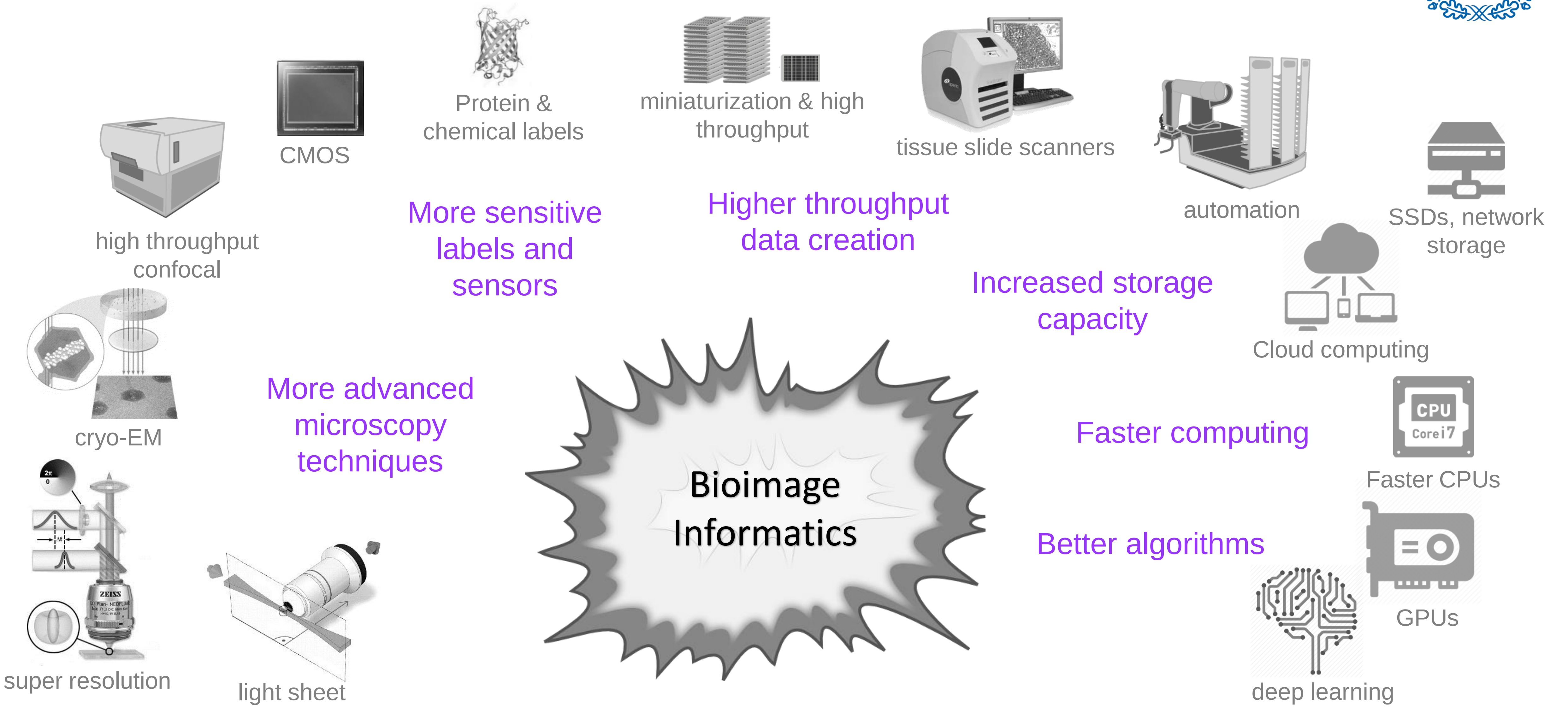
Deep Learning for Life Science

Hossein Azizpour
Assistant Professor in Machine Learning
Robotics, Perception, and Learning (RPL)
KTH

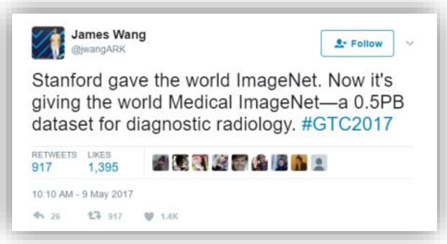
The time is right!

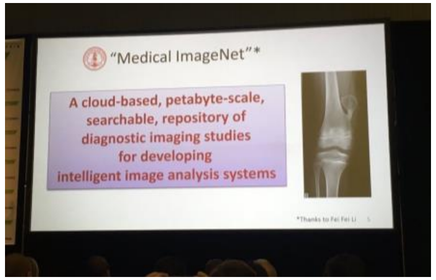
necessary technological developments come together



A New Era

of large publicly available data and benchmarks





DREAM CHALLENGES

powered by Sage Bionetworks

kaggle

CAMELYON17

Year(s)	Challenge	Task
2009-2010	DIADEM	Neuron morphology
2012	PTC	Particle detection in time-lapse microscopy
2012	SNEMI23	Neurite segmentation in 2D
2012,2014	MITOS	Cell mitosis detection in histopathology
2012,2014	DM3D	3D deconvolution microscopy
2013-2015	CTC	Cell segmentation and tracking
2013	SMLM	Single-molecule localization
2013	SNEMI3D	Neurite segmentation in 3D
2013	AMIDA	Cell mitosis detection in histopathology
2014,2015	OCCIS	Overlapping cell segmentation in cancer
2014	MITOS-ATYPIA	Mitosis detection and cancer scoring
2015	GLAS	Gland segmentation in histopathology
2015	NCC	Nucleus counting in multichannel micro.
2015-2016	BigNeuron	Large-scale 3D neuron reconstruction
2016-2017	CAMELYON	Cancer metastasis detection
2017	DREAM	Breast cancer detection
2017	CYTO	Cell atlas protein localization

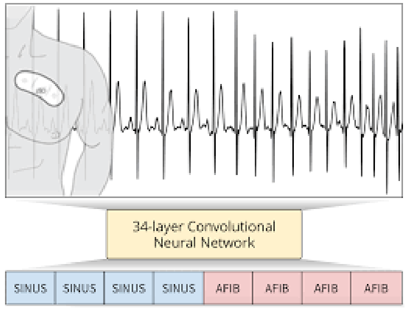
A New Era

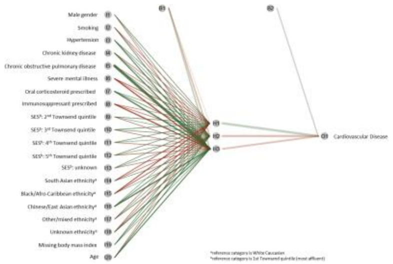
Excitement surrounding bioimage informatics

A New Era

of AI-assisted medical advances

Accuracy

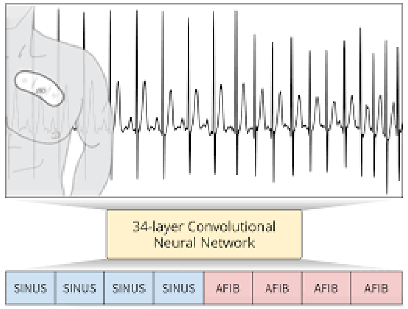


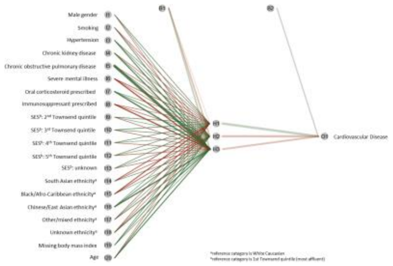


A New Era
of AI-assisted medical advances

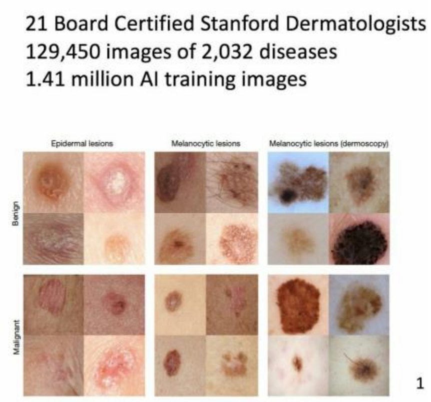
Accuracy

Lack of specialists



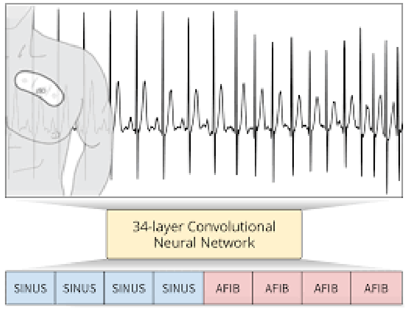


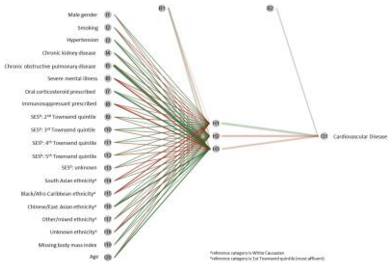
A New Era
of AI-assisted medical advances



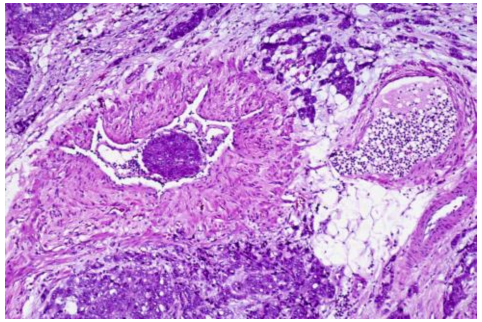
Lack of specialists

Accuracy

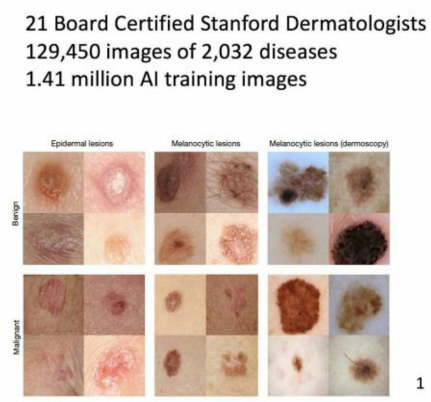




Efficiency



A New Era
of AI-assisted medical advances

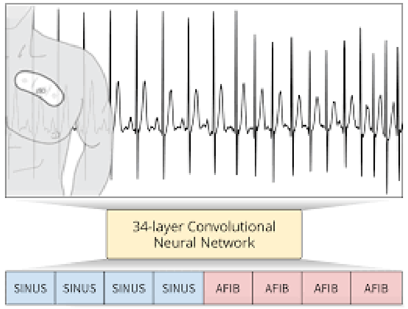


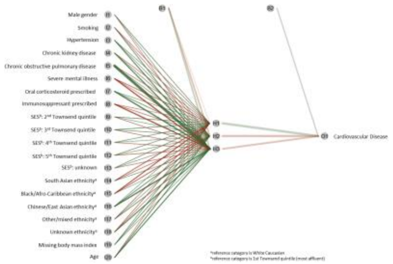
Lack of specialists

Cheaper

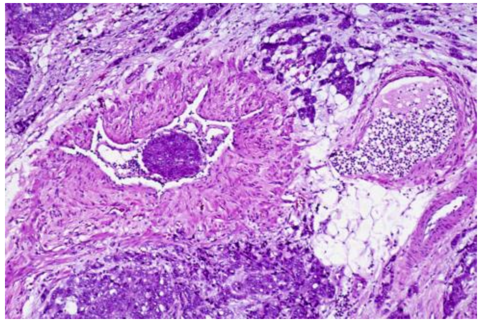
zebra™
MEDICAL.VISION

Accuracy





Efficiency



A New Era

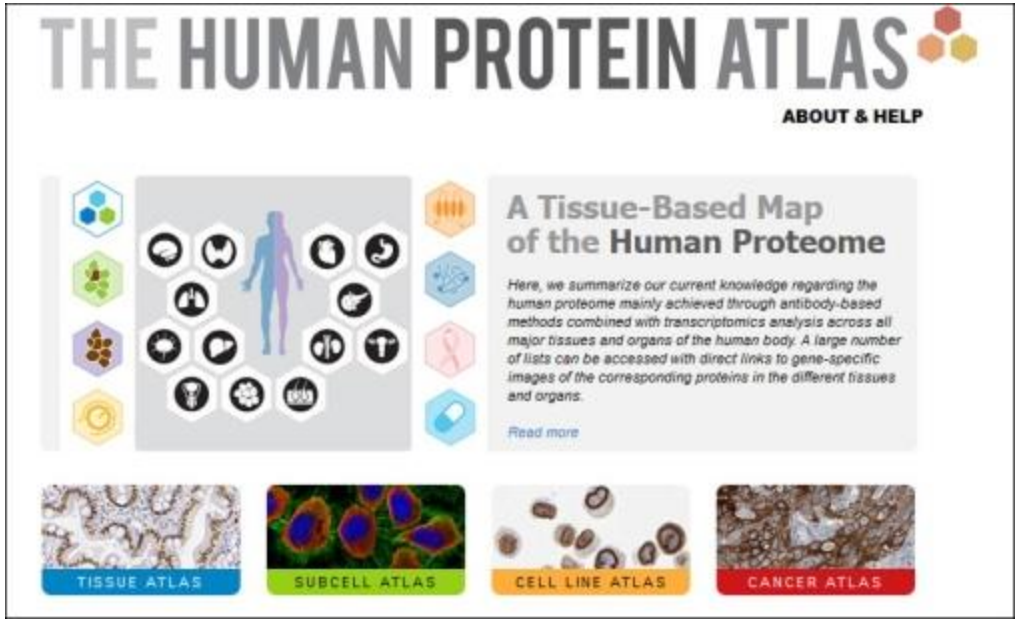
General Life Science

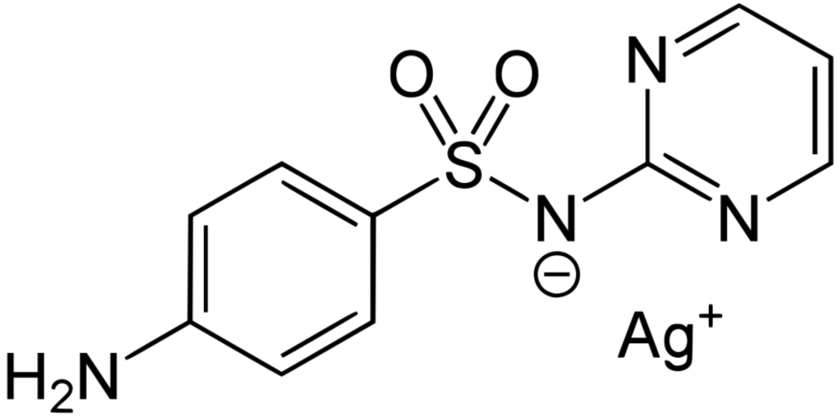
KTH VETENSKAP OCH KONST

verily

SciLifeLab

AstraZeneca



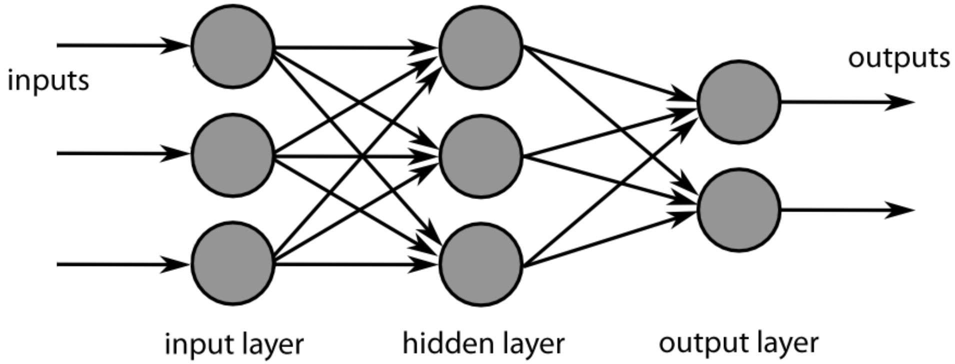


Contents

- Problem Definition
- Opportunities and challenges
- An example pipeline
- Domain Adaptation
- Uncertainty Estimation
- Future Directions

Contents

Deep Learning for Medical Science



Deep Learning for Medical Science

Under-supervised Learning

D ~ P(X,Y) and P(X) P(Y|X) or P(Z|X)

weakly supervised learning

semi-supervised learning

Supervised Learning

D ~ P(X,Y) P(Y|X)

classification

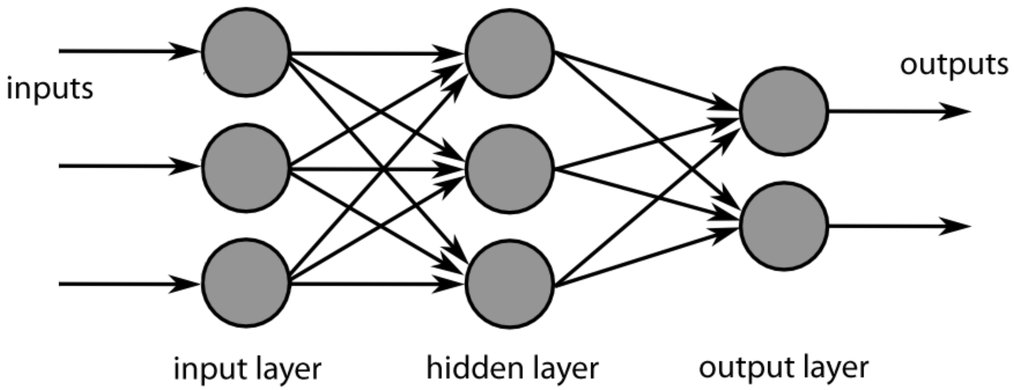
regression

Un/self Supervised Learning

D ~ P(X) P(X), P(Z|X), P(X|Z)

clustering

dimensionality reduction



Discriminative Models

Generative Models

Deep Learning for Medical Science

Prognosis

Risk Estimation

Diagnosis

Novel Biomarker

Patient Care

Personalized Medicine

New Treatments

Contents

- Problem Definition
- Opportunities and challenges
- An example pipeline
- Domain Adaptation
- Uncertainty Estimation
- Future Directions

Opportunities

- Large amount of data is available (stored somewhere, hopefully digitized)
  - High-content screening
  - Digital pathology
  - Radiology
  - Drug discovery
  - Genomics
  - Clinical data
  - …
- Repetitive routine tasks
- Shortage of specialists, outsourcing diagnosis/prognosis to different countries
- Training procedure for new specialists
- Improved and/or consistent accuracy
- Efficient diagnosis process
- Interesting scientific challenges
- Private and public Investments
- Last but not least, direct impact on human lives

Challenges

- Reliability
  - Interpretability
  - Uncertainty
- Recall is extremely important, precision also very important
- Public scrutiny for automated systems and AI in general
- Data standardization
- Different imaging equipment with different internal parameters and operator settings
- Multi-modal data
- Non-Euclidean data
- Bias (regional, temporal, etc.). Fair machine learning.
- Annotation cost (specialists' time are precious/expensive)
- Ambiguous definitions → Disagreement among specialists
- Privacy Issues with data collection
- Reverse engineering networks
- Inherently noisy labels
- Deep learning is data-hungry

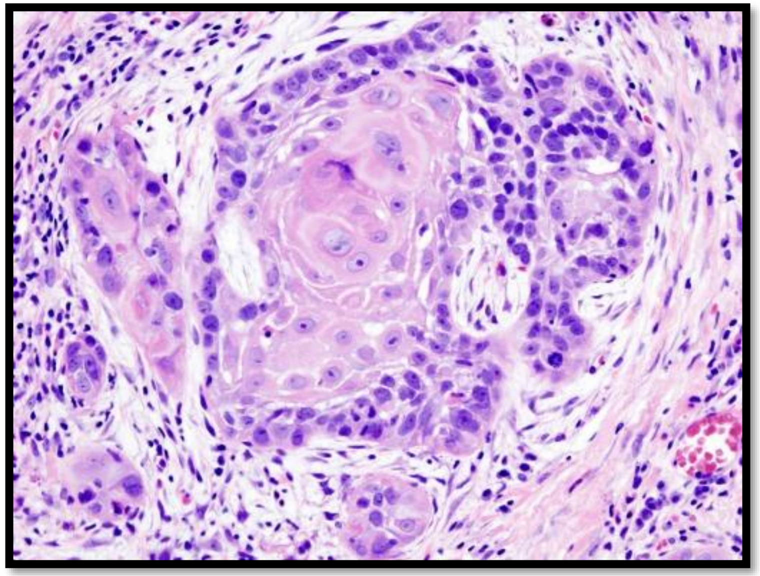
Challenges

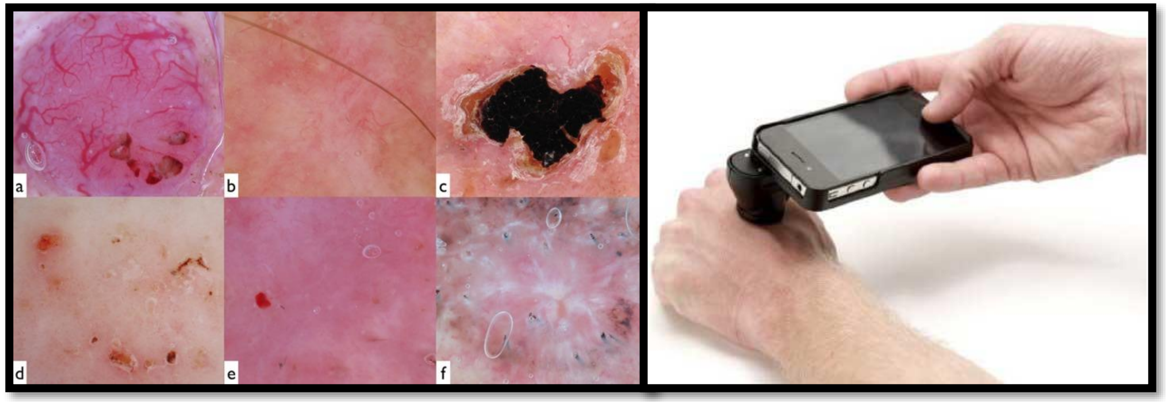
- **Reliability**
  - Interpretability
  - **Uncertainty**
- Recall is extremely important, precision also very important
- Public scrutiny for automated systems and AI in general
- Data standardization
- **Different imaging equipment with different internal parameters and operator settings**
- Multi-modal data
- Non-euclidean data
- Bias (regional, temporal, etc.). Fair machine learning.
- Annotation cost (specialists' time are precious/expensive)
- Ambiguous definitions → Disagreement among specialists
- Privacy Issues with data collection
- Reverse engineering networks
- Inherently noisy labels
- Deep learning is data-hungry

Contents

Medical Projects

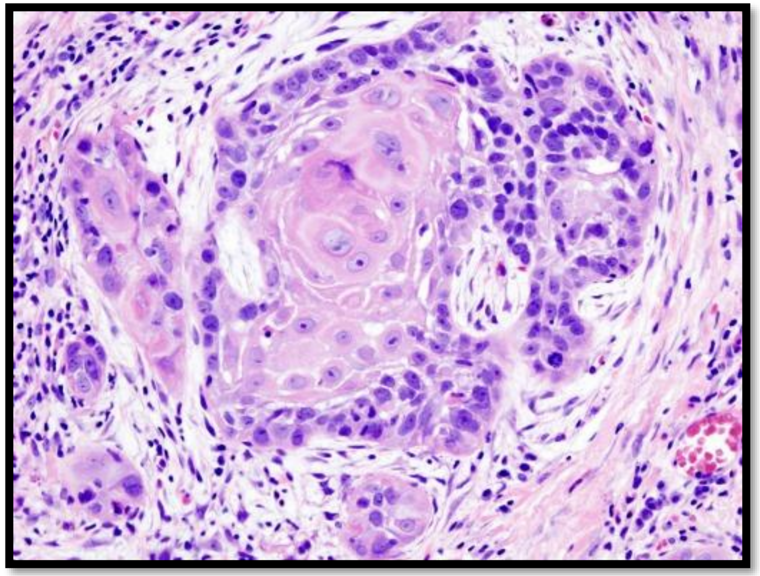
- Histopathology
- Radiology
- Standard camera

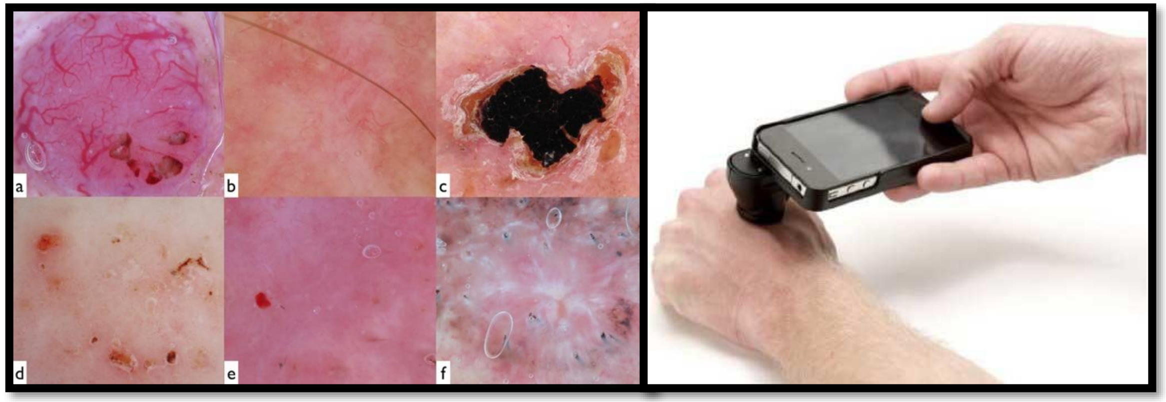




Medical Projects

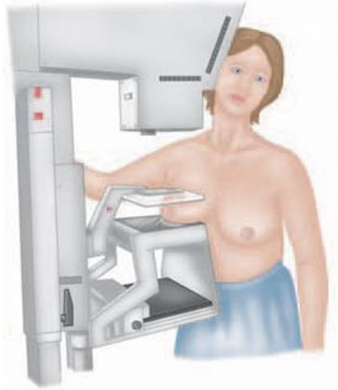
- Histopathology
- **Radiology** ✓
- Standard camera



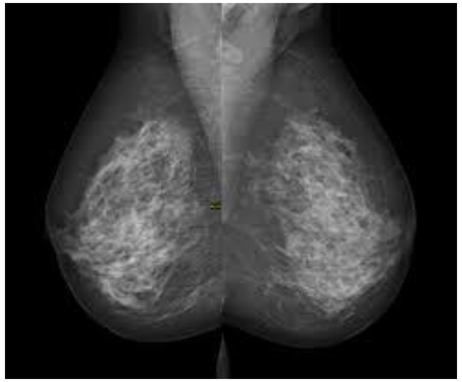


Digital mammographic screening for breast cancer

Current screening in Sweden



Women invited for screening*



Digital mammograms acquired from multiple angles

Two radiologists review case

Discussion and decision

Healthy

Recall

1. Additional imaging
2. Physical exam
3. Tissue sample

Cancer diagnosis

It is estimated that mammography-based screening **reduces breast cancer mortality by around 30%**

*All women in Sweden between 40 and 74 are invited for screening every 18 to 24 months.

Critical shortage of radiologists

SVENSKA DAGBLADET

Start | Näringsliv | Kultur | Ledare | ≡ Meny

Debatt

”Akut personalbrist hotar mammografin”

Landstingen klarar inte själva att lösa dilemmat med att landets mammografienheter dräneras på personal, samtidigt som belastningen ökar. Nu krävs konkret handling för att trygga mammografiverksamheten, skriver Ulrika Årehed Kågström och Jan Zedenius, Cancerfonden.

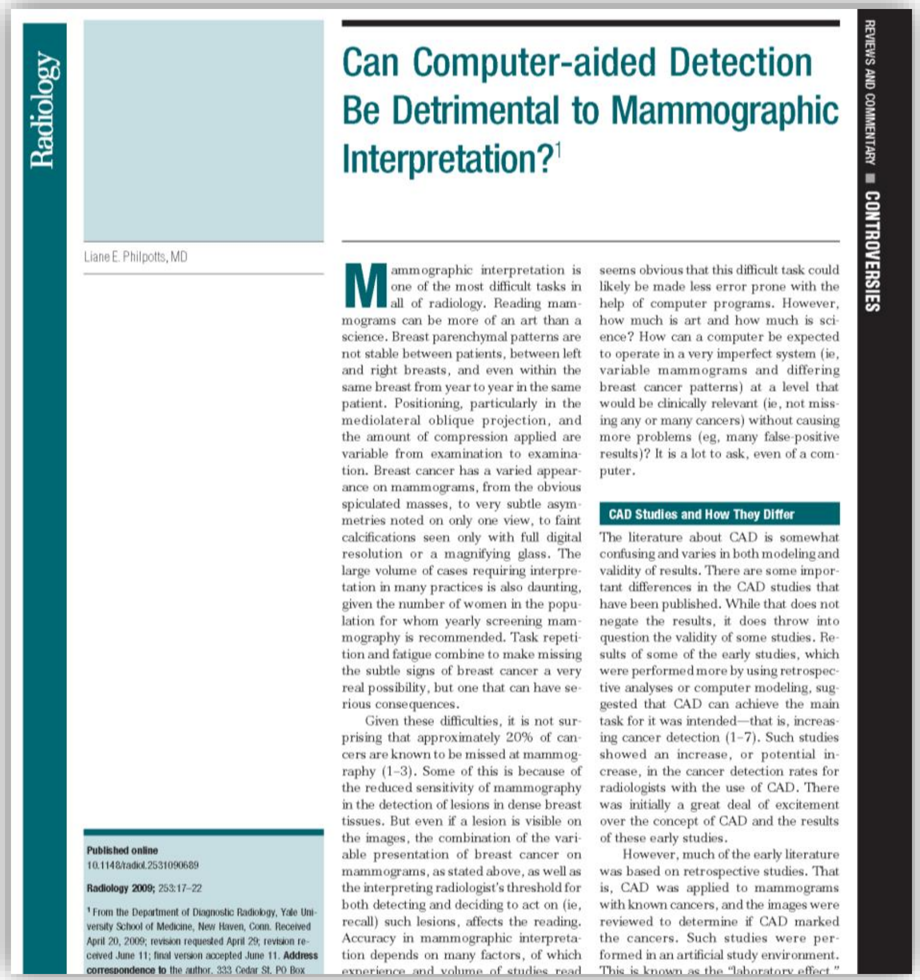
11 okt, 2016

Spara artikel

313 delningar

Foto: Lars Pehrson

Limitations in Computer aided detection (CAD)



Radiology

REVIEWS AND COMMENTARY ■ CONTROVERSIES

Can Computer-aided Detection Be Detrimental to Mammographic Interpretation?[1]

Liane E. Philpotts, MD

Mammographic interpretation is one of the most difficult tasks in all of radiology. Reading mammograms can be more of an art than a science. Breast parenchymal patterns are not stable between patients, between left and right breasts, and even within the same breast from year to year in the same patient. Positioning, particularly in the mediolateral oblique projection, and the amount of compression applied are variable from examination to examination. Breast cancer has a varied appearance on mammograms, from the obvious spiculated masses, to very subtle asymmetries noted on only one view, to faint calcifications seen only with full digital resolution or a magnifying glass. The large volume of cases requiring interpretation in many practices is also daunting, given the number of women in the population for whom yearly screening mammography is recommended. Task repetition and fatigue combine to make missing the subtle signs of breast cancer a very real possibility, but one that can have serious consequences.

Given these difficulties, it is not surprising that approximately 20% of cancers are known to be missed at mammography (1–3). Some of this is because of the reduced sensitivity of mammography in the detection of lesions in dense breast tissues. But even if a lesion is visible on the images, the combination of the variable presentation of breast cancer on mammograms, as stated above, as well as the interpreting radiologist's threshold for both detecting and deciding to act on (ie, recall) such lesions, affects the reading. Accuracy in mammographic interpretation depends on many factors, of which experience and volume of studies read

seems obvious that this difficult task could likely be made less error prone with the help of computer programs. However, how much is art and how much is science? How can a computer be expected to operate in a very imperfect system (ie, variable mammograms and differing breast cancer patterns) at a level that would be clinically relevant (ie, not missing any or many cancers) without causing more problems (eg, many false-positive results)? It is a lot to ask, even of a computer.

CAD Studies and How They Differ

The literature about CAD is somewhat confusing and varies in both modeling and validity of results. There are some important differences in the CAD studies that have been published. While that does not negate the results, it does throw into question the validity of some studies. Results of some of the early studies, which were performed more by using retrospective analyses or computer modeling, suggested that CAD can achieve the main task for it was intended—that is, increasing cancer detection (1–7). Such studies showed an increase, or potential increase, in the cancer detection rates for radiologists with the use of CAD. There was initially a great deal of excitement over the concept of CAD and the results of these early studies.

However, much of the early literature was based on retrospective studies. That is, CAD was applied to mammograms with known cancers, and the images were reviewed to determine if CAD marked the cancers. Such studies were performed in an artificial study environment. This is known as the "laboratory effect."

Published online
10.1148/radiol.2531090689

Radiology 2009; 253:17–22

[1] From the Department of Diagnostic Radiology, Yale University School of Medicine, New Haven, Conn. Received April 20, 2009; revision requested April 29; revision received June 11; final version accepted June 11. **Address correspondence to** the author, 333 Cedar St, PO Box

Limitations of CAD

Clinical success depends on CAD having a high sensitivity, a reasonable specificity, and the reader taking appropriate action when interpreting the CAD prompts. All of these features work in conjunction, and all need to be optimized for the system to be valuable. Because mammography is an imperfect system, particularly in the detection of many of the more subtle cancers in denser breast tissues, to have reasonable sensitivity, the false-positive rate of CAD prompts must be high; thus, the specificity is low. Thus, the balance is not easy to achieve.

Difficult cases result in unacceptably high FP rate

MammoAI

mammoAI

KTH VETENSKAP OCH KONST

BRO

KAROLINSKA INSTITUTET ANNO 1810

UCSF

KAROLINSKA UNIVERSITETSSJUKHUSET

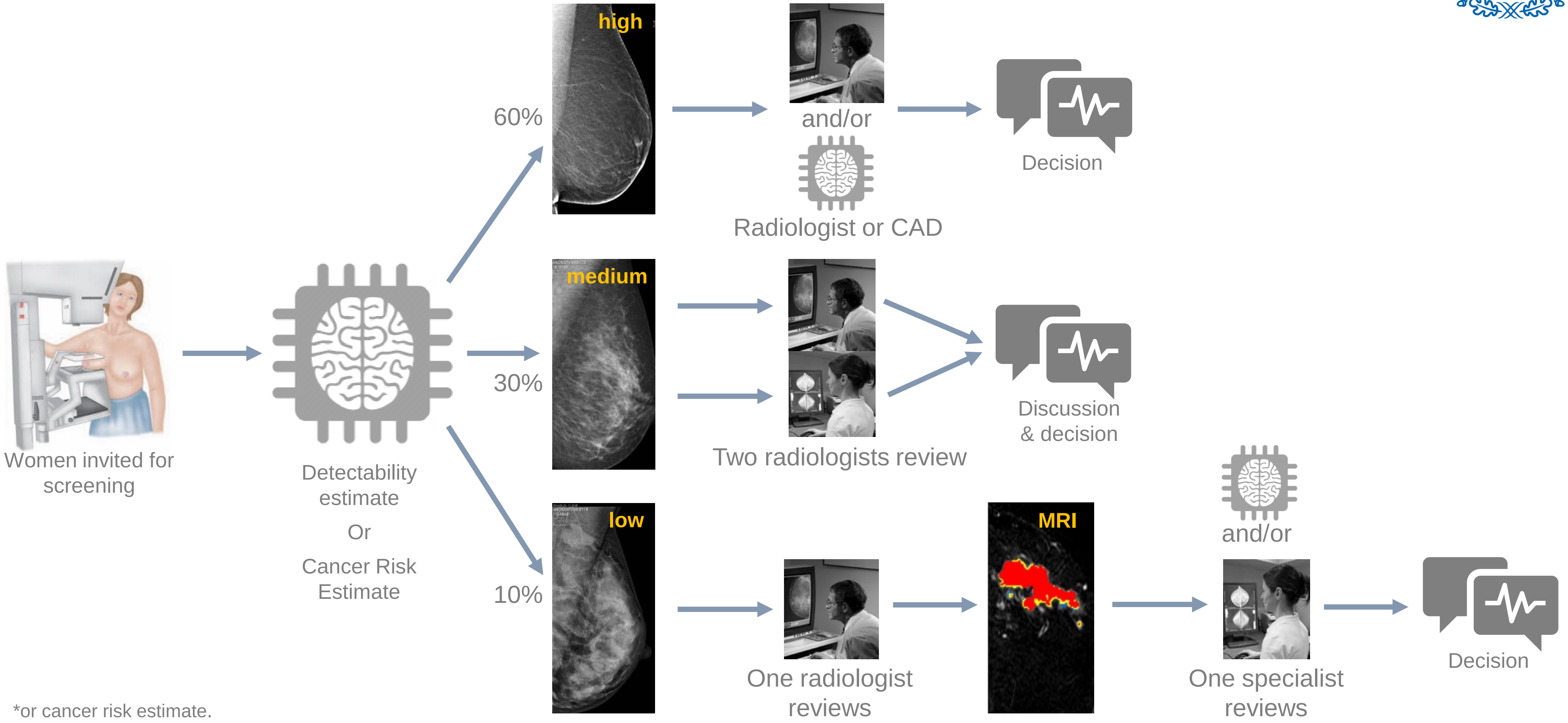
Sage BIONETWORKS

REGIONALT CANCERCENTRUM STOCKHOLM GOTLAND

STOCKHOLM SCIENCE CITY FOUNDATION

SECTRA

Optimized Workflow



*or cancer risk estimate.

Massive dataset:

every mammogram in Stockholms län (2008 – 2015)

KAROLINSKA
UNIVERSITETSSJUKHUSET

Screening: 241,149 exams

Clinical: 57,752 exams

REGIONALT
CANCERCENTRUM
STOCKHOLM GOTLAND

Screening register:

~400,000 women (entire SLL)

REGIONALT
CANCERCENTRUM
STOCKHOLM GOTLAND

Cancer register:

~9,000 women (entire SLL)

mammoAI

Total: ~2,400,000 images

Cancer: ~20,000 images

INbreast

Total: 410 images

Cancer: 360 images

CBIS-DDSM

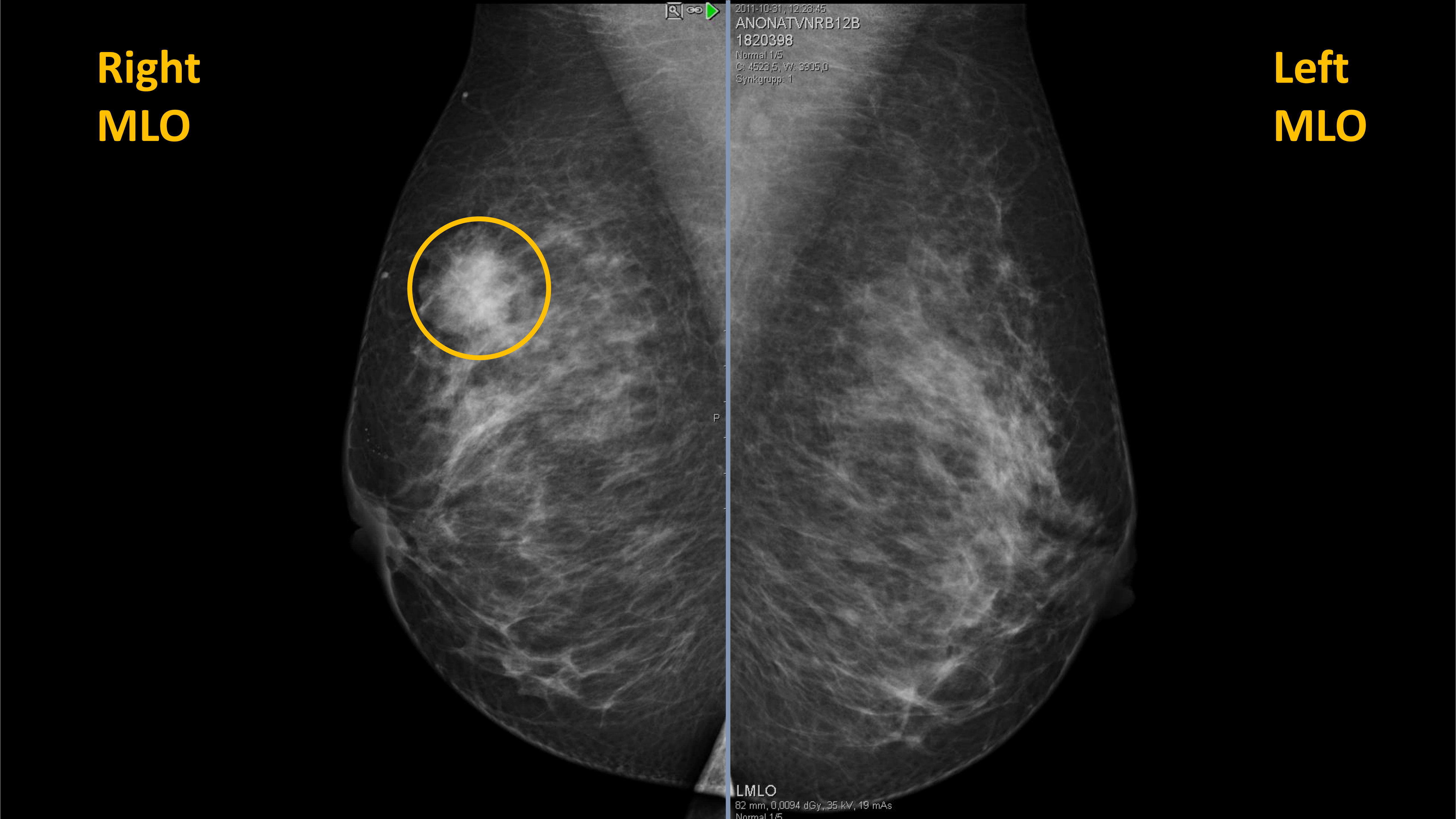
Total: 4,067 images

Cancer: 855 images

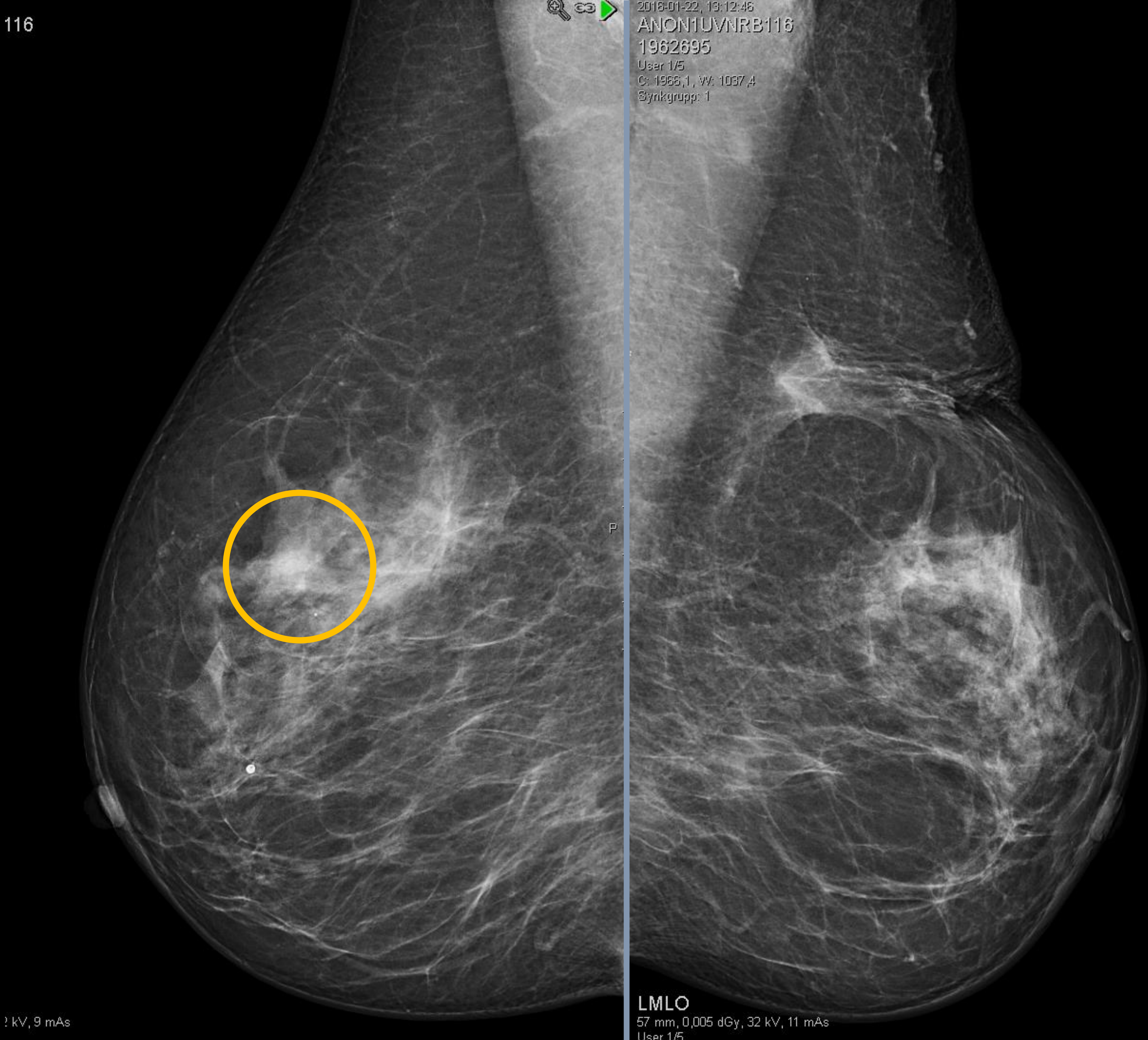
Deep Learning for Mammography

Tumor Detection Network (CAD)

and Risk Estimation


Right
MLO
Left
MLO
2011-10-31, 12:28:45
ANONATVNRB12B
1820398
Normal 1/5
C: 4523,5, W: 3906,0
Synkgrupp: 1
P
LMLO
82 mm, 0,0094 dGy, 35 kV, 19 mAs
Normal 1/5

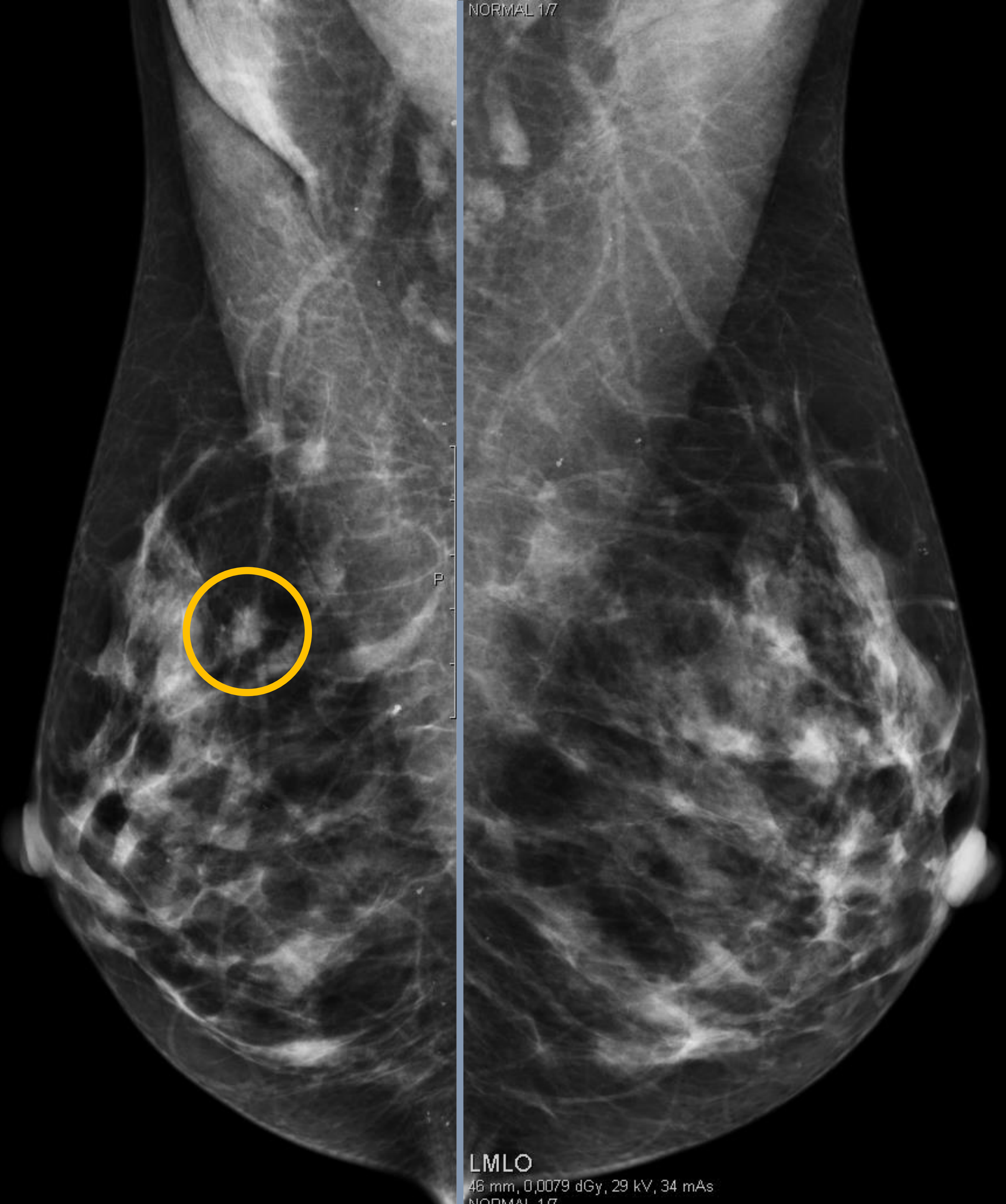
Right
MLO



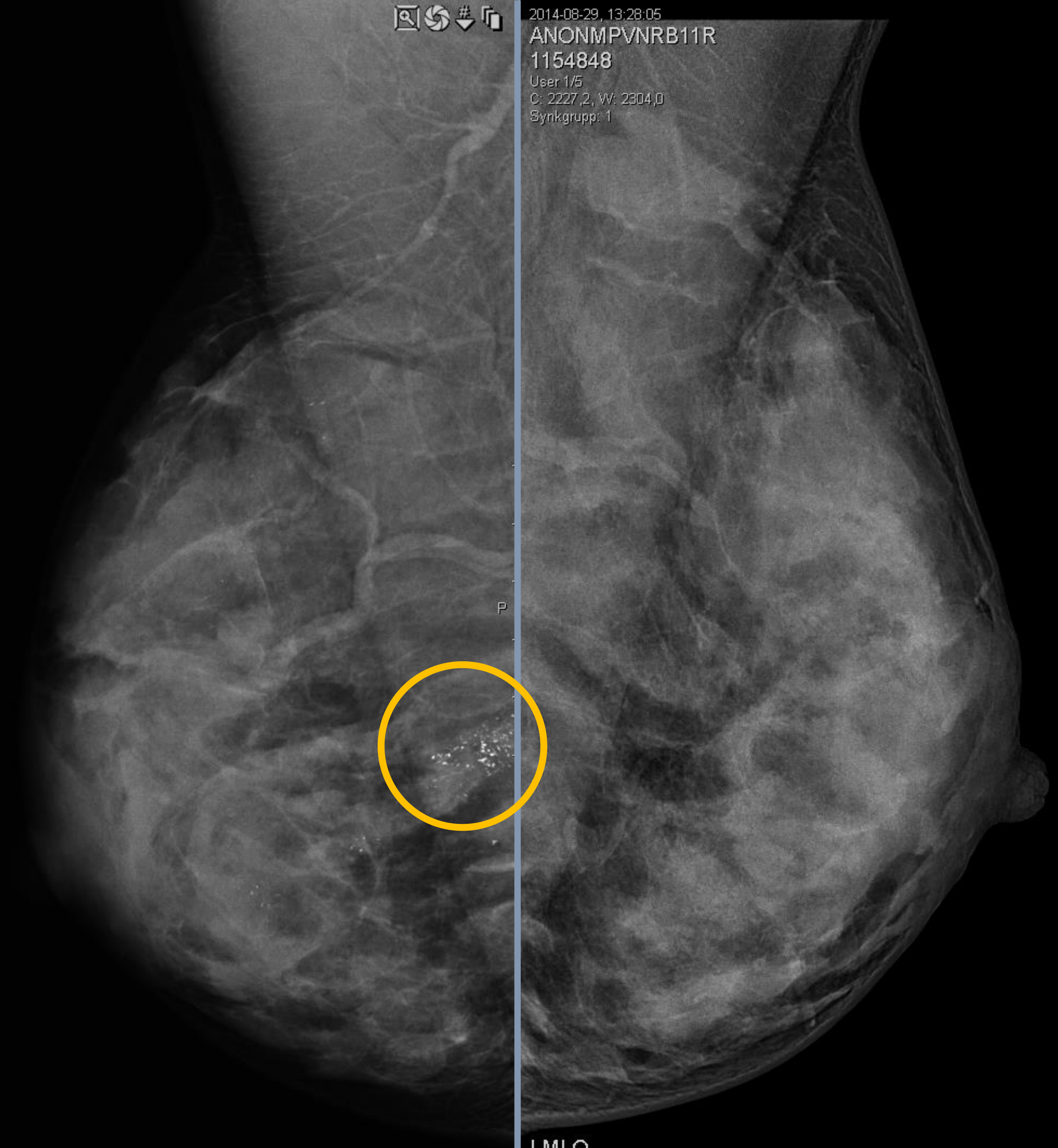
Left
MLO

Right
MLO

Left
MLO



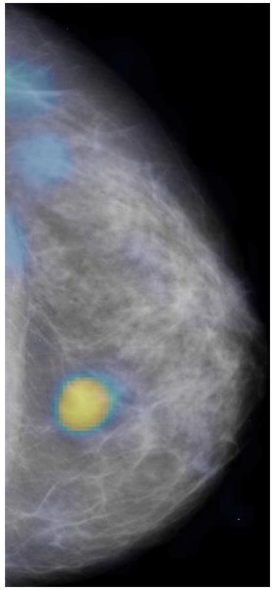
Right
MLO

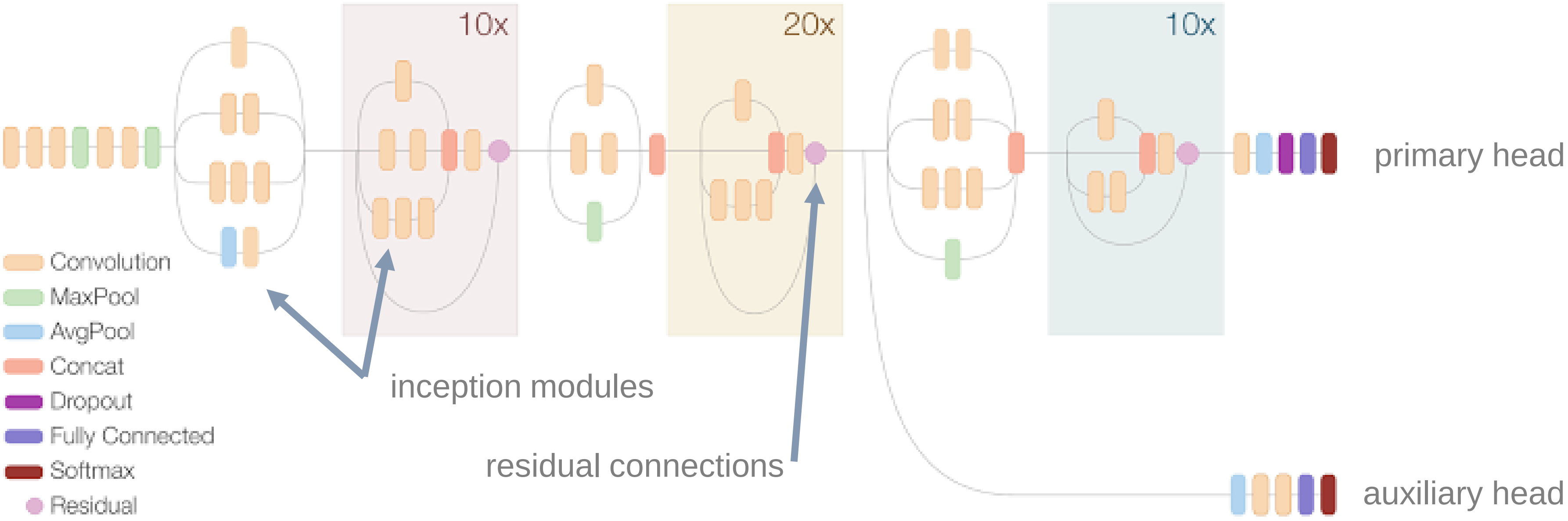


Left
MLO

Tumor detection

Based on state-of-the-art **Inception-ResNet**[1] architecture





[1] Szegedy, Christian, et al. "Inception-v4, inception-resnet and the impact of residual connections on learning." arXiv preprint arXiv:1602.07261 (2016).

Training the network

Modified the pretrained network (ImageNet) so it is **fully convolutional**

First, trained the network to localize tumors using existing annotated datasets (semi-supervised learning)

Dream

Total: 500 images

Cancer: 32 images

CBIS-DDSM

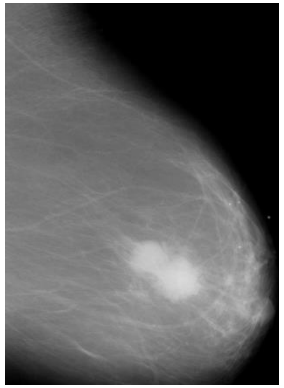
Total: 4,067 images

Cancer: 855 images

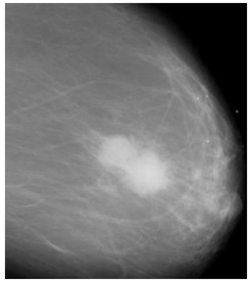
INbreast

Total: 410 images

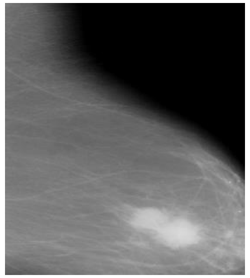
Cancer: 360 images



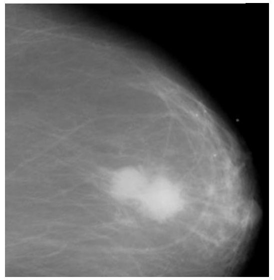
Data augmentation – generated n ing examples from a few thousand



random crops



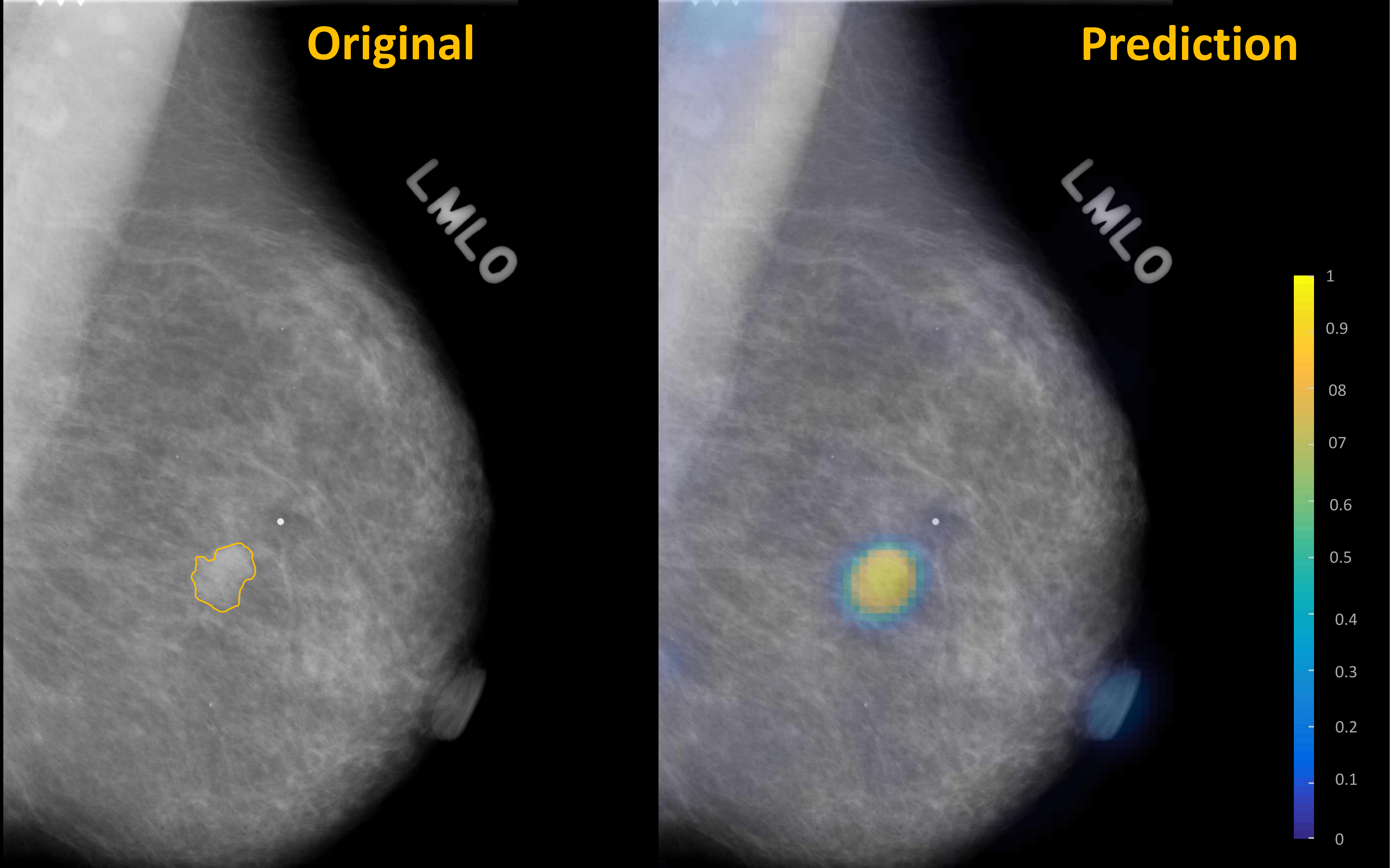
aspect ratio

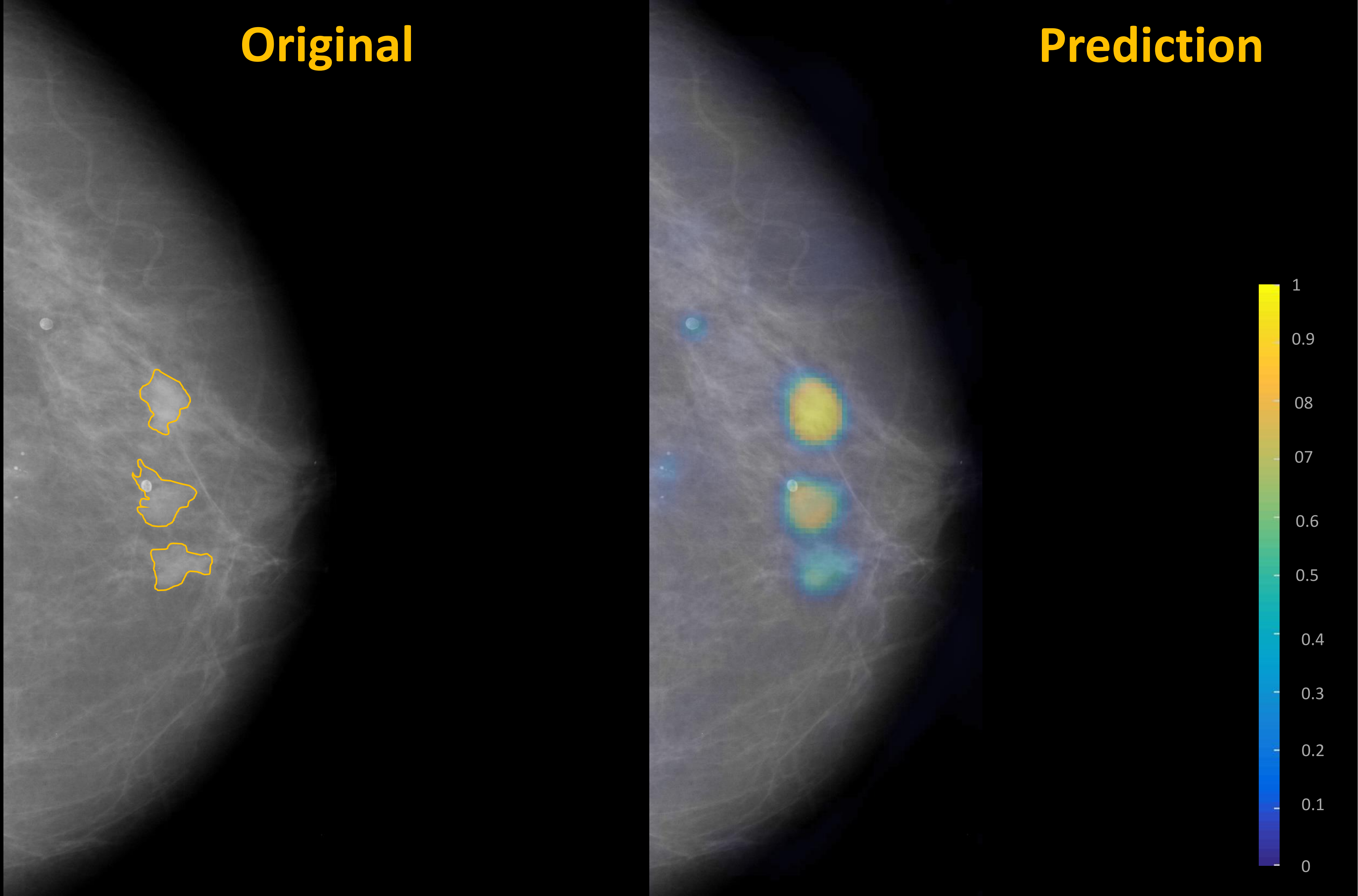


rotation

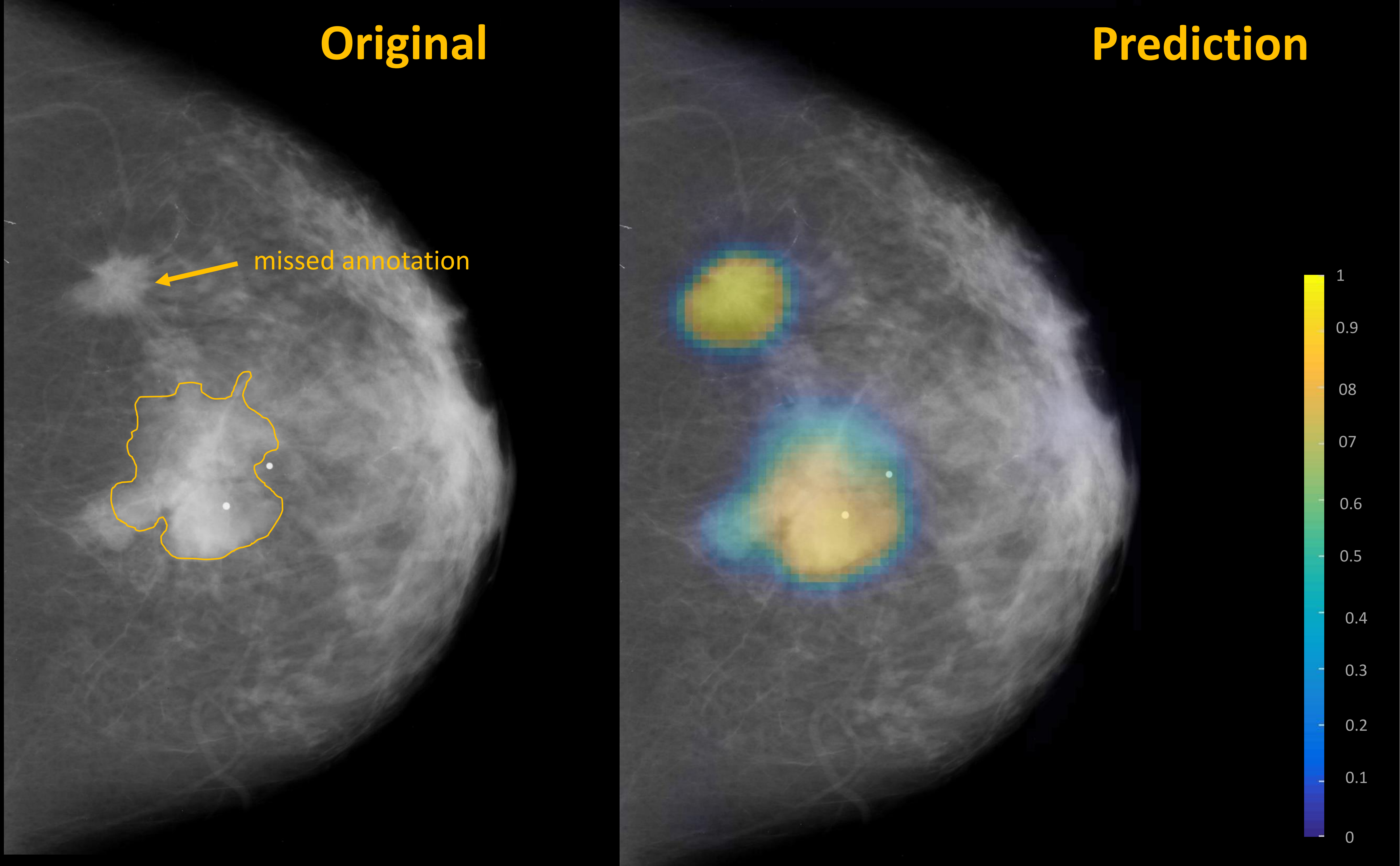
contrast

brightness


Original
LMLO
Prediction
LMLO
1
0.9
08
07
0.6
0.5
0.4
0.3
0.2
0.1
0



Original
Prediction
1
0.9
08
07
0.6
0.5
0.4
0.3
0.2
0.1
0


Original
missed annotation
Prediction
1
0.9
08
07
0.6
0.5
0.4
0.3
0.2
0.1
0

Risk Estimation

- Risk: risk of a woman developing cancer at some point in future
- Positive set: prior and contra-lateral mammograms

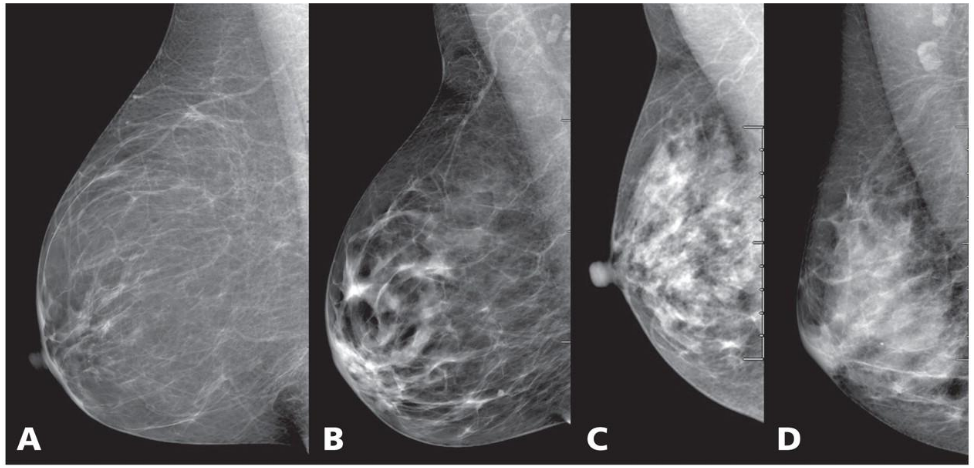


Figure 1 Representations of the 4 Breast Imaging Reporting and Data System (BI-RADS) breast density qualitative and quantitative assessments. A) BI-RADS 1: almost entirely fat; B) BI-RADS 2: scattered fibroglandular densities; C) BI-RADS 3: heterogenously dense; and D) BI-RADS 4: extremely dense.

Method	AUC	Odds Ratio (95% CI)
Deep Learning Risk	64%	5.32 (2.39-9.69)
Mammographic Density Risk	57%	1.96 (1.23-3.11)

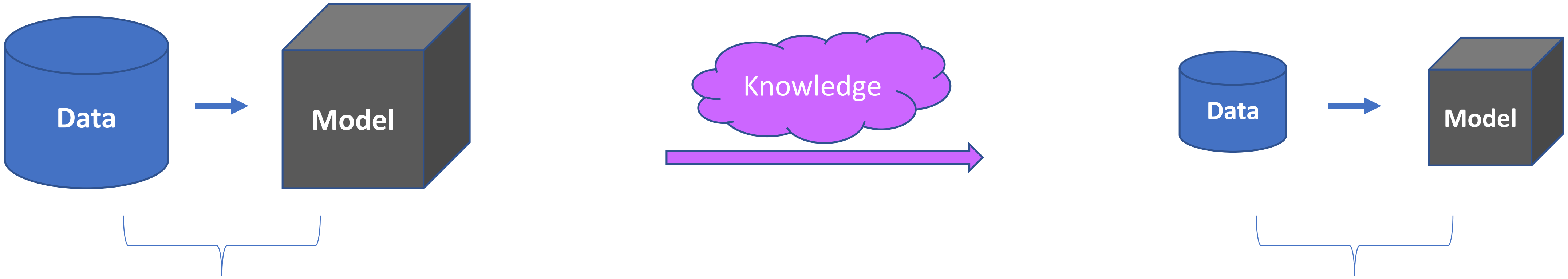
[Dembrower, Azizpour, Smith, Konuk, Strand “TRAINING A DEEP LEARNING NETWORK TO ASSESS BREAST CANCER RISK”. , CARS 2018]

Contents

Knowledge Transfer

SOURCE **TARGET**

$D_s \sim P_s(X,Y)$ $M_s \equiv P_s(Y_s|X)\ or\ P(X,Y_s)$ $D_t \sim P_t(X,Y)$ $M_t \equiv P_t(Y_t|X)\ or\ P(X,Y_t)$



Rich/large **small/noisy**

Transfer Learning

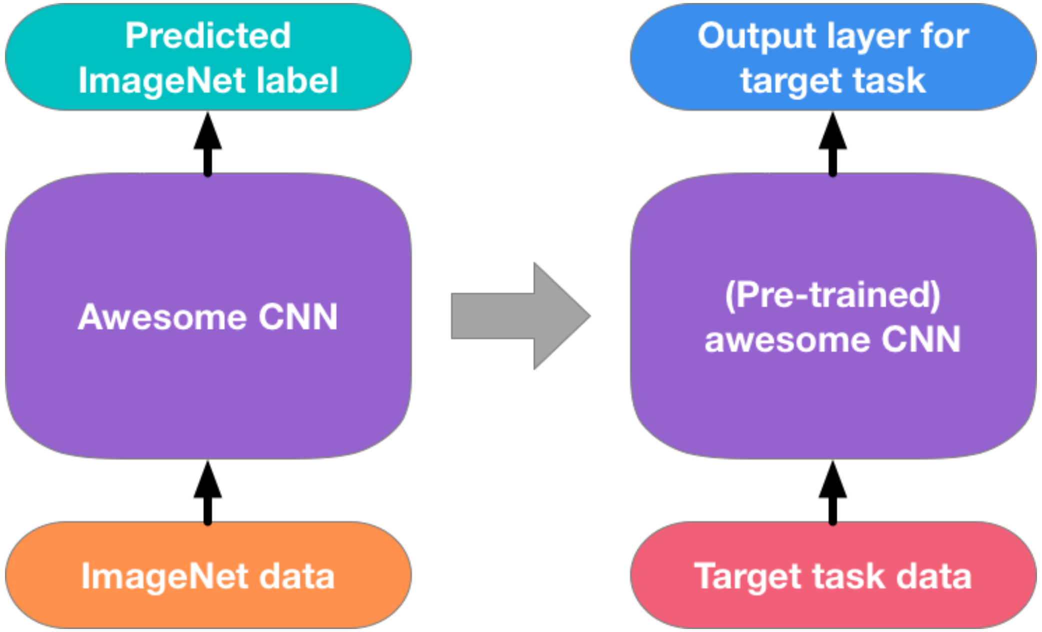
- when $Y_s \neq Y_t$ or $P_s(Y_s|X)$

Human detection

Horse detection

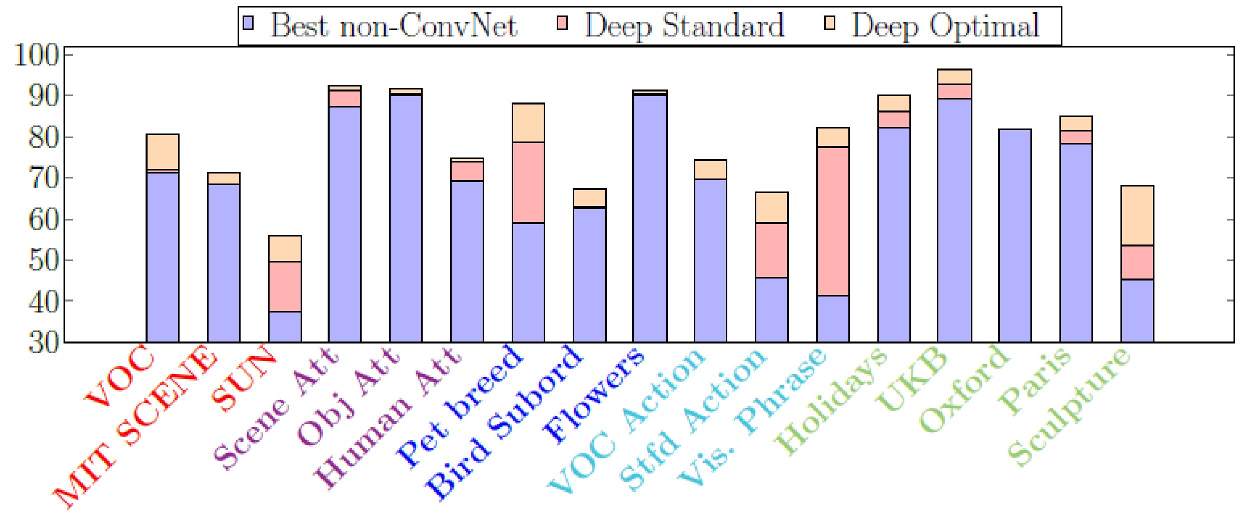
Deep Transfer Learning

- Fine tuning is the most common way



[source: gluon.mxnet.io]

Deep Transfer Learning



	Target task				
Factor	Source task ImageNet	···	FineGrained recognition	···	Instance retrieval
Early stopping	Don't do it				
Network depth	As deep as possible				
Network width	Wider				Moderately wide
Diversity/Density	More classes better than more images per class				
Fine-tuning	Yes, more improvement with more labelled data				
Dim. reduction	Original dim				Reduced dim
Rep. layer	Later layers				Earlier layers

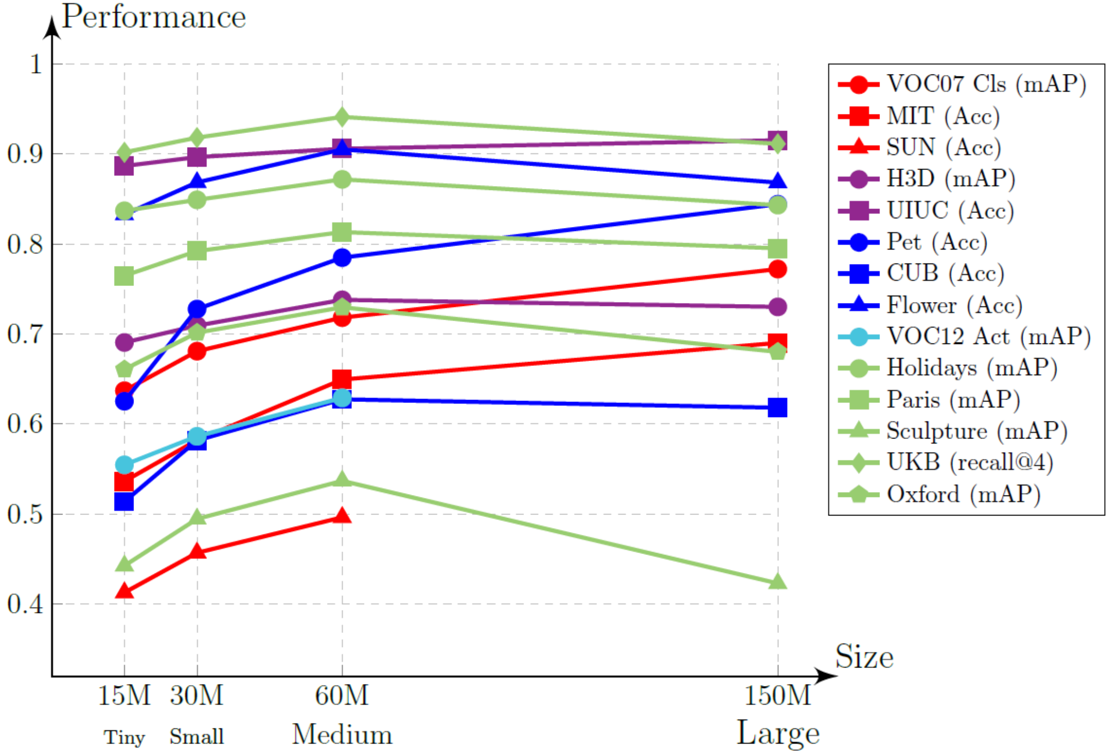
Increasing distance from ImageNet

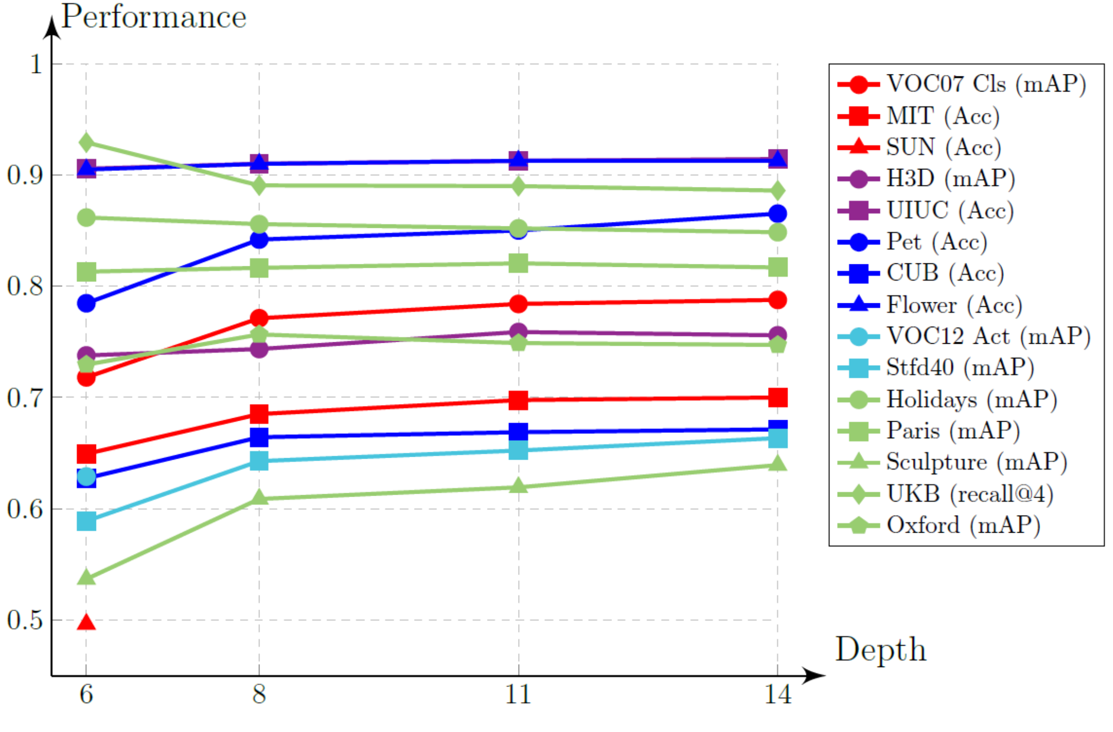
Image Classification	Attribute Detection	Fine-grained Recognition	Compositional	Instance Retrieval
PASCAL VOC Object [9]	H3D human attributes [6]	Cat&Dog breeds [29]	VOC Human Action [9]	Holiday scenes [17]
MIT 67 Indoor Scenes [33]	Object attributes [10]	Bird subordinate [43]	Stanford 40 Actions [46]	Paris buildings [31]
SUN 397 Scene [45]	SUN scene attributes [30]	102 Flowers [27]	Visual Phrases [34]	Sculptures [4]

[Azizpour et al. Factors of Transferability for a Generic ConvNet Representation, PAMI 2016]

Deep Transfer Learning

Width vs Depth

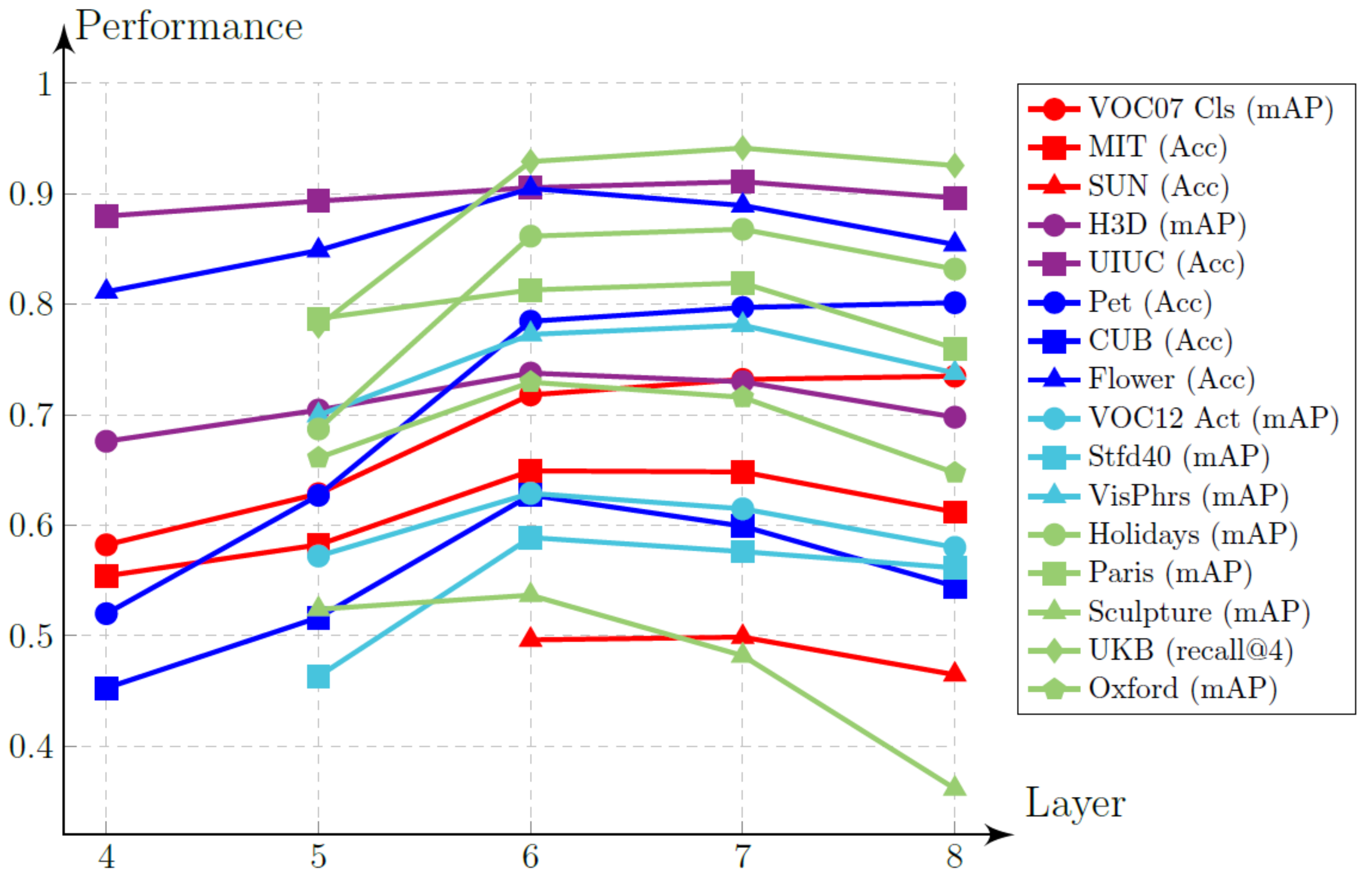




[Azizpour et al. Factors of Transferability for a Generic ConvNet Representation, PAMI 2016]

Deep Transfer Learning

Which Layer?



[Azizpour et al. Factors of Transferability for a Generic ConvNet Representation, PAMI 2016]

Deep Transfer Learning

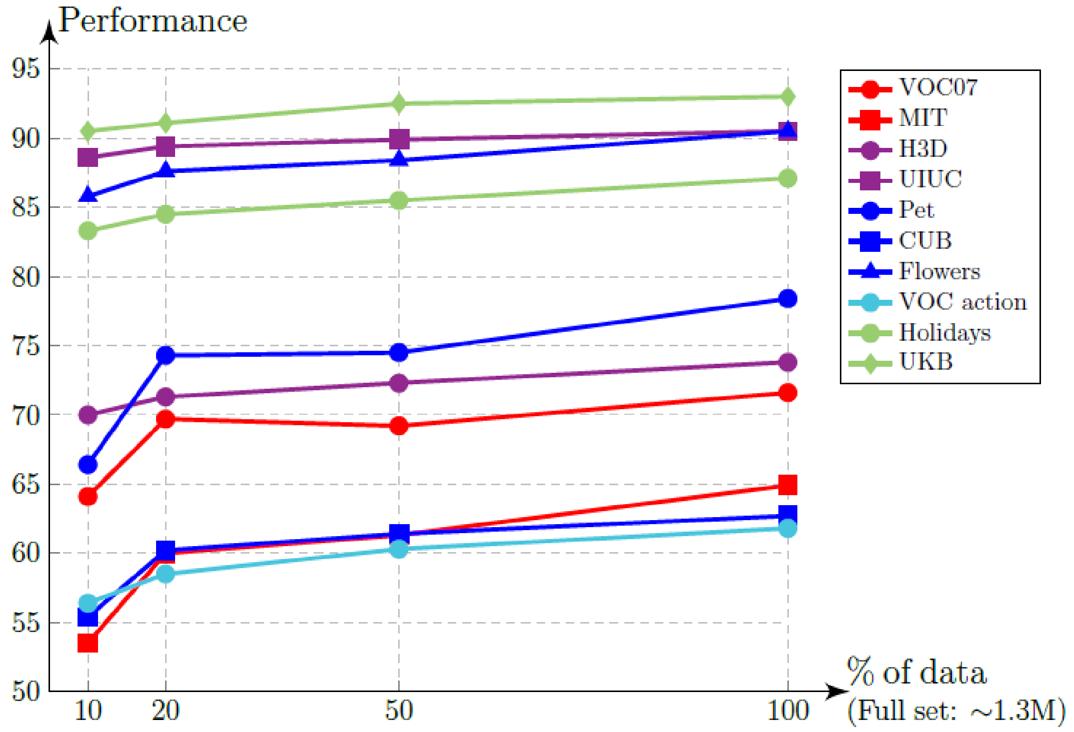
Multiple Representations

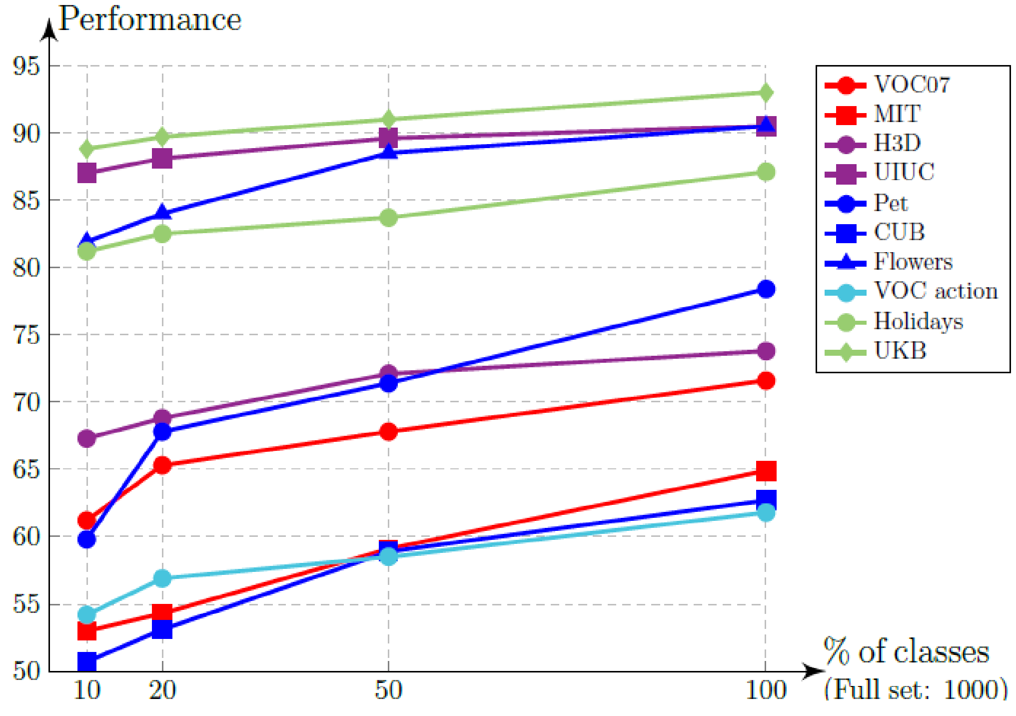
	Image Classification			Attribute Detection		Fine-grained Recognition			Compositional	Instance Retrieval		
Source task	VOC07	MIT	SUN	H3D	UIUC	Pet	CUB	Flower	Stanf. Act40	Oxf.	Scul.	UKB
ImageNet	71.6	64.9	49.6	73.8	**90.4**	**78.4**	**62.7**	**90.5**	58.9	71.2	52.0	93.0
Places	68.5	69.3	55.7	68.0	88.8	49.9	42.2	82.4	53.0	70.0	44.2	88.7
Hybrid	72.7	69.6	56.0	72.6	90.2	72.4	58.3	89.4	58.2	**72.3**	52.3	92.2
Concat	**73.8**	**70.8**	**56.2**	**74.2**	**90.4**	75.6	60.3	90.2	**59.6**	72.1	**54.0**	**93.2**

[Azizpour et al. Factors of Transferability for a Generic ConvNet Representation, PAMI 2016]

Deep Transfer Learning

Diversity vs Density





[Azizpour et al. Factors of Transferability for a Generic ConvNet Representation, PAMI 2016]

Deep Transfer Learning

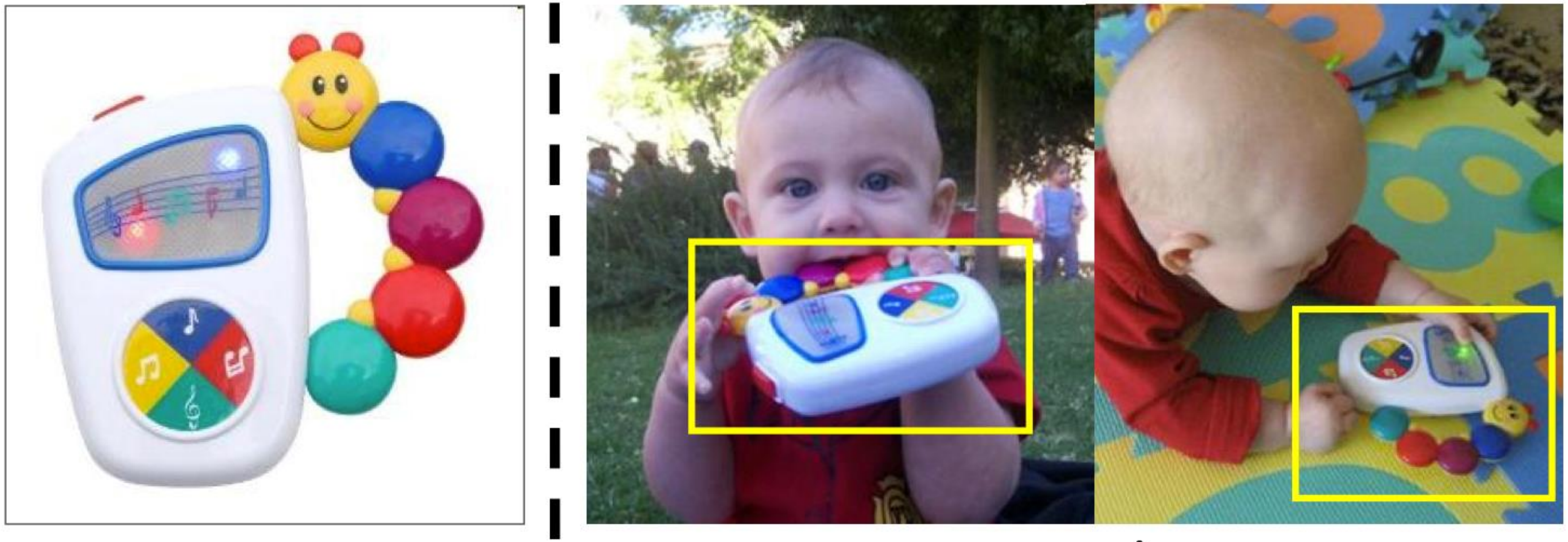
Optimizing Transferability Factors are Important!

	Image Classification			Attribute Detection			Fine-grained Recognition			Compositional			Instance Retrieval				
	VOC07	MIT	SUN	SunAtt	UIUC	H3D	Pet	CUB	Flower	VOCa.	Act40	Phrase	Holid.	UKB	Oxf.	Paris	Scul.
non-ConvNet	[38] 71.1	[25] 68.5	[44] 37.5	[30] 87.5	[42] 90.2	[50] 69.1	[29] 59.2	[12] 62.7	[20] 90.2	[28] 69.6	[46] 45.7	[34] 41.5	[40] 82.2	[51] 89.4	[40] **81.7**	[40] 78.2	[4] 45.4
Deep Standard	71.8	64.9	49.6	91.4	90.6	73.8	78.5	62.8	90.5	69.2	58.9	77.3	86.2	93.0	73.0	81.3	53.7
Deep Optimized[4]	**80.7**	**71.3**	**56.0**	**92.5**	**91.5**	**74.6**	**88.1**	**67.1**	**91.3**	**74.3**	**66.4**	**82.3**	**90.0**	**96.3**	79.0	**85.1**	**67.9**
Err. Reduction	32%	18%	13%	13%	10%	4%	45%	12%	8%	17%	18%	22%	28%	47%	22%	20%	31%
Source Task	ImgNet	Hybrid	Hybrid	Hybrid	ImgNet	ImgNet	ImgNet	ImgNet	ImgNet	ImgNet	ImgNet	ImgNet	Hybrid	ImgNet	ImgNet	ImgNet	ImgNet
Network Width	Medium	Medium	Medium	Medium	Large	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium
Network Depth	16	8	8	8	8	16	16	16	16	16	16	16	8	8	16	16	16
Rep. Layer	last	last	last	last	2nd last	2nd last	2nd last	3rd last	3rd last	3rd last	3rd last	3rd last	4th last	4th last	4th last	4th last	4th last
PCA	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓
Pooling	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	1×1	1×1	2×2	2×2	3×3

[Azizpour et al. Factors of Transferability for a Generic ConvNet Representation, PAMI 2016]

Domain Adaptation

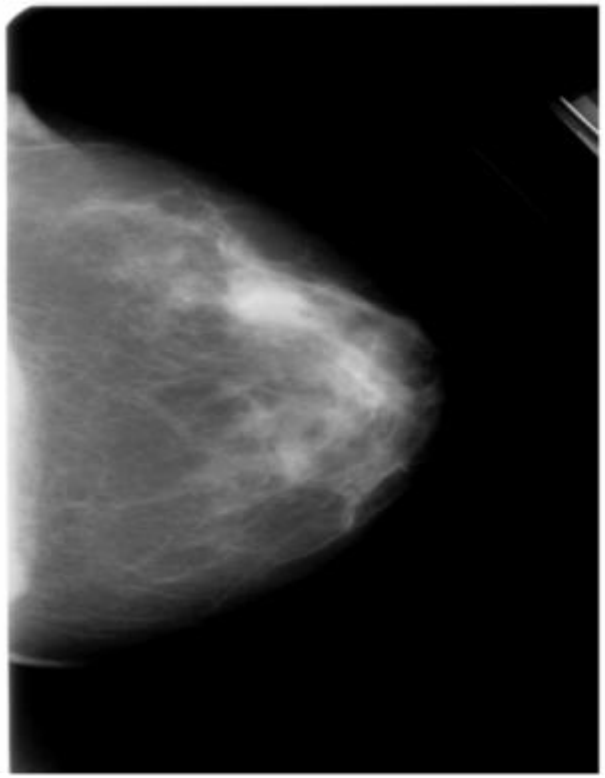
- when $P_s(X) \neq P_t(X)$



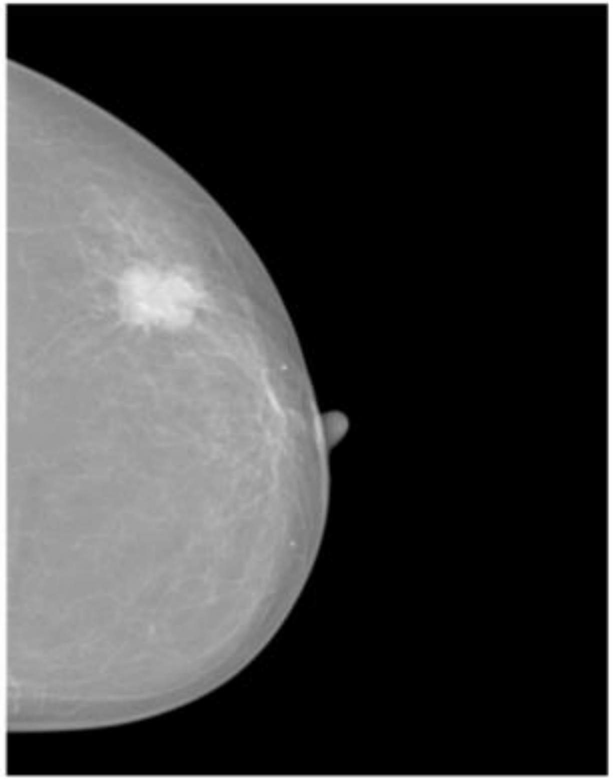
amazon.com consumer images

Domain Adaptation

DDSM



INBREAST



DREAM

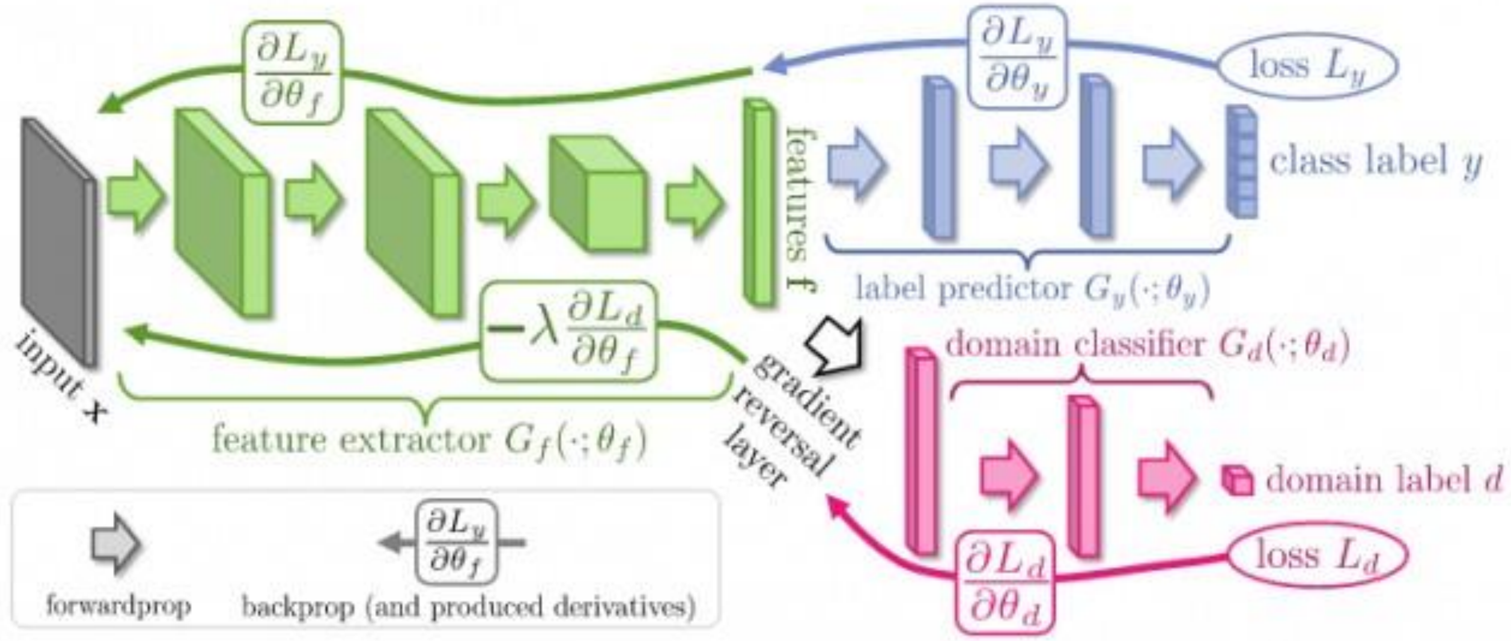
Domain Adaptation

There is usually the assumption that samples drawn from $P_s(X)$ have a corresponding sample in $P_t(X)$.

- Supervised Domain Adaptation
- Unsupervised Domain Adaptation

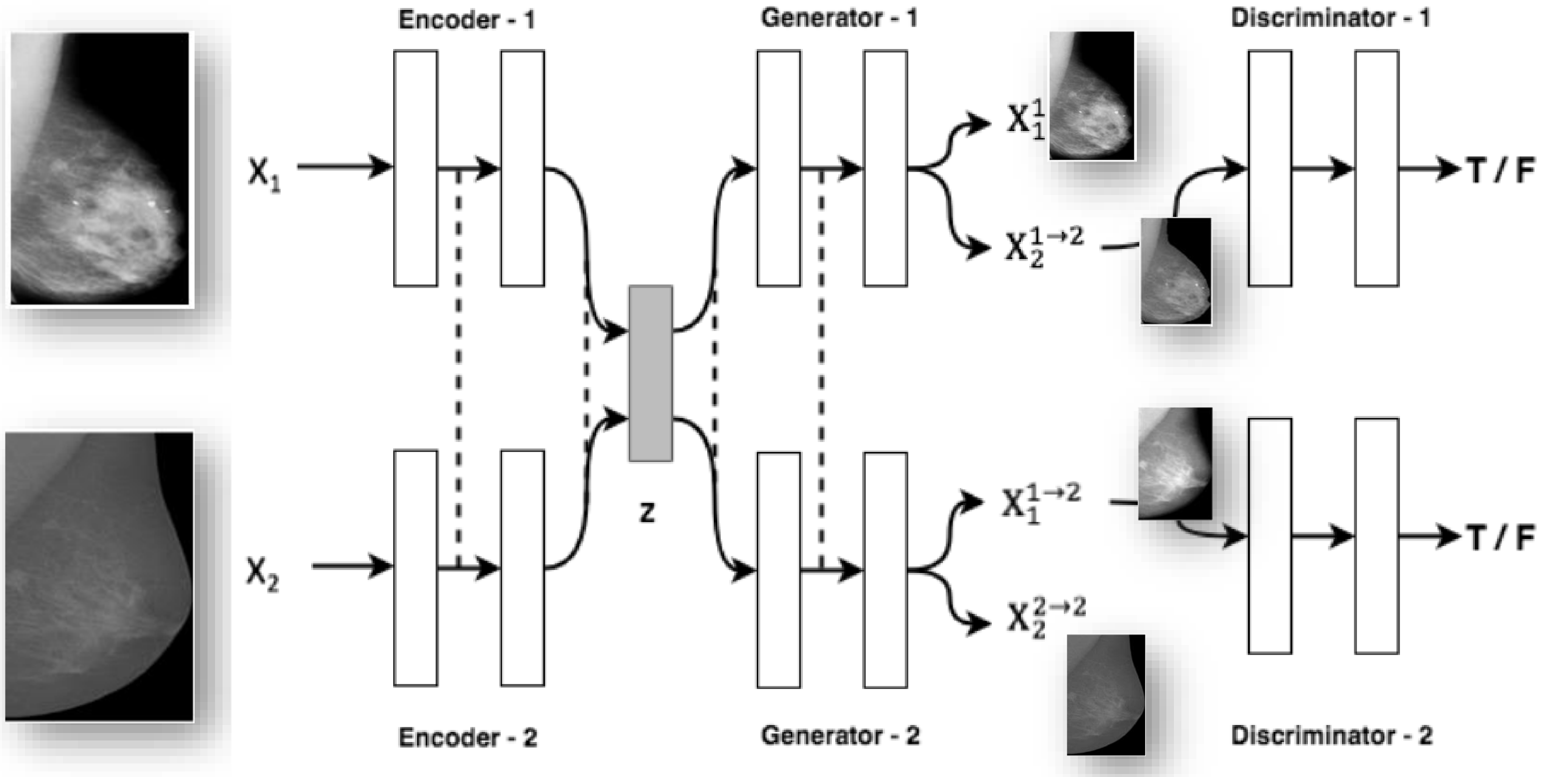
- Directly adapting the model parameters [Rozantsev et al. 2017]
- Learning a common embedding space [Ganin et al 2015, 2016]
- Adapting in the input domain [Bousmalis et al. 2016 Liu et al 2017]

Domain Adaptation



[Ganin et al. Domain-adversarial training of neural networks, JMLR 2016]

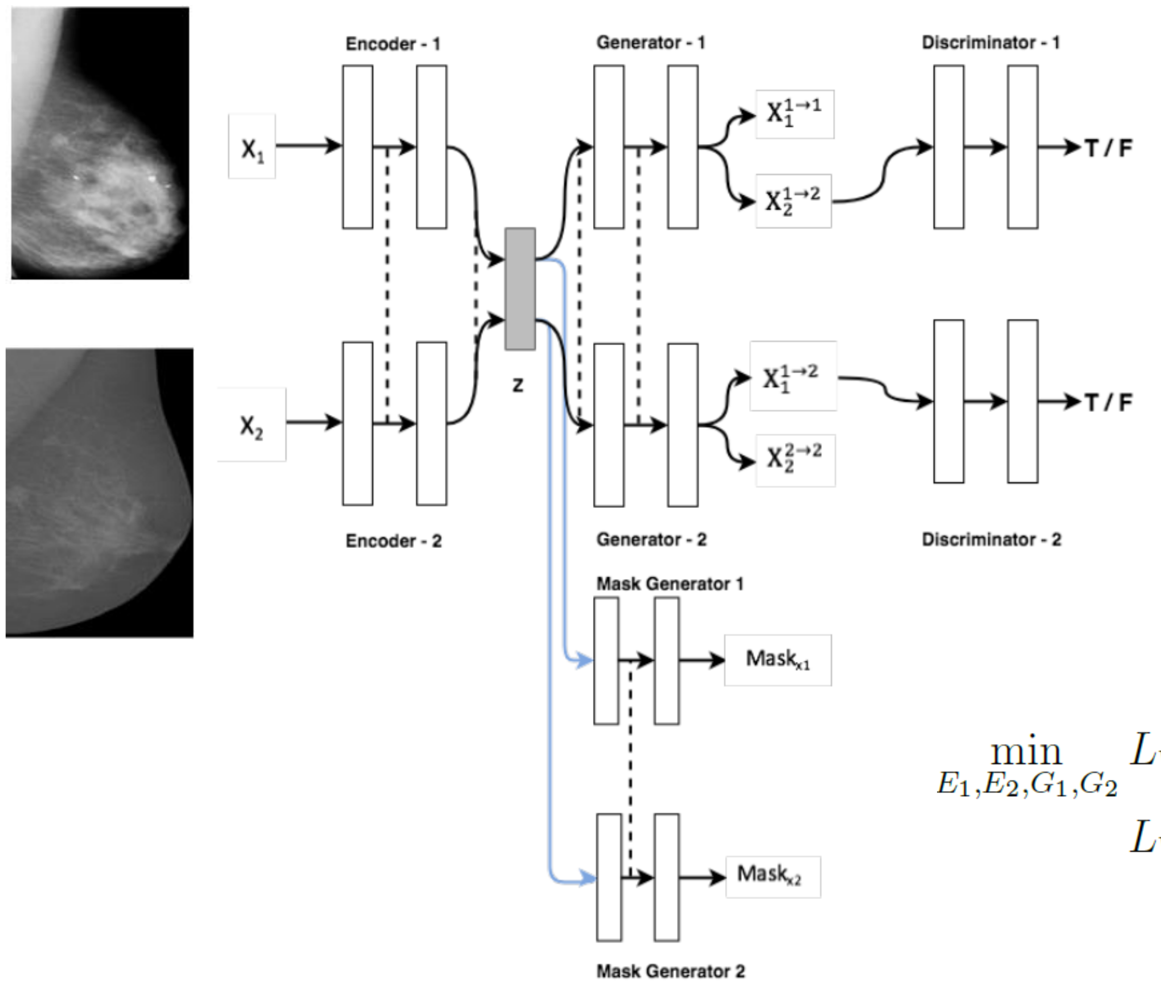
Domain Adaptation



$$\begin{aligned}\min_{E_1,E_2,G_1,G_2} \; & L_{VAE_1}(E_1,G_1) \;+\; L_{CC_1}(E_1,G_1,E_2,G_2) + V_{LSGAN}(E_1,G_1) \\ & L_{VAE_2}(E_2,G_2) \;+\; L_{CC_2}(E_2,G_2,E_1,G_1) + V_{LSGAN}(E_2,G_2)\end{aligned}$$

[Liu et al. Unsupervised Image-to-Image Translation Networks, NIPS 2017]

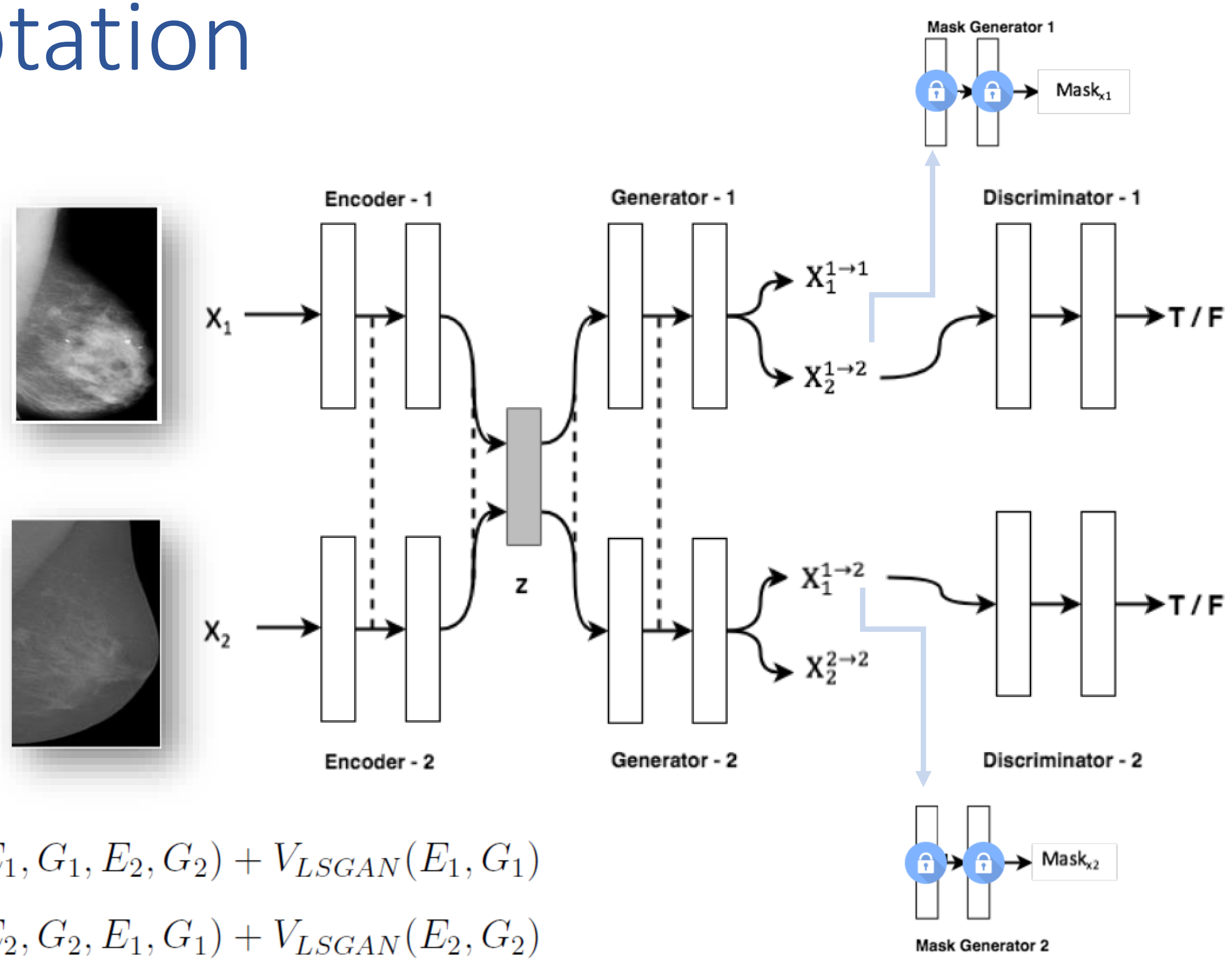
Domain Adaptation



$$\min_{E_1,E_2,G_1,G_2} L_{VAE_1}(E_1,G_1) + L_{CC_1}(E_1,G_1,E_2,G_2) + V_{LSGAN}(G_1)$$
$$L_{VAE_2}(E_2,G_2) + L_{CC_2}(E_2,G_2,E_1,G_1) + V_{LSGAN}(G_2)$$
$$+ L_{label}(E_1, Mask_{G_1}) + L_{label}(E_2, Mask_{G_2})$$

[Tao Wang, Adapting multiple datasets for better mammography tumor detection, KTH 2018]

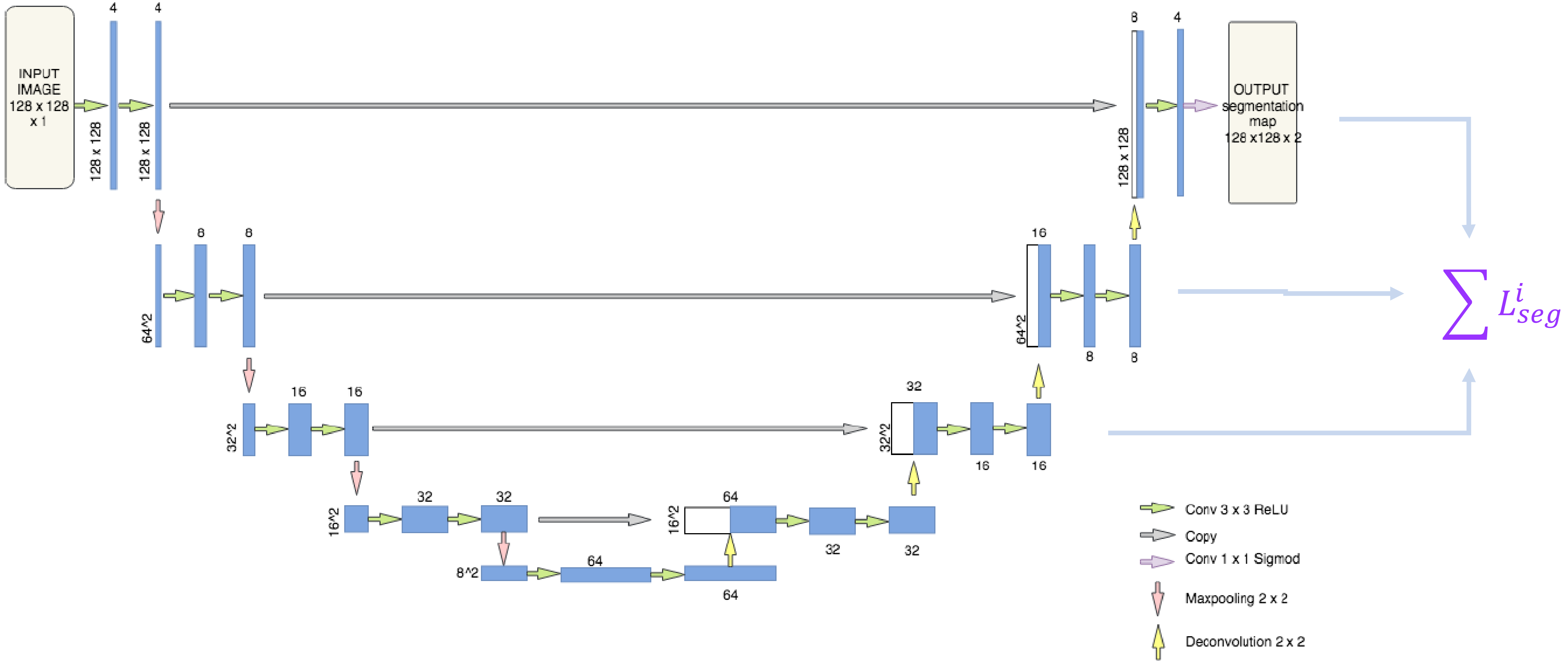
Domain Adaptation



$$
\begin{aligned}
\min_{E_1,E_2,G_1,G_2} \; & L_{VAE_1}(E_1,G_1) &+& \quad L_{CC_1}(E_1,G_1,E_2,G_2) + V_{LSGAN}(E_1,G_1) \\
& L_{VAE_2}(E_2,G_2) &+& \quad L_{CC_2}(E_2,G_2,E_1,G_1) + V_{LSGAN}(E_2,G_2) \\
& &+& \quad L_{seg}(E_2,G_1) + L_{seg}(E_1,G_2)
\end{aligned}
$$

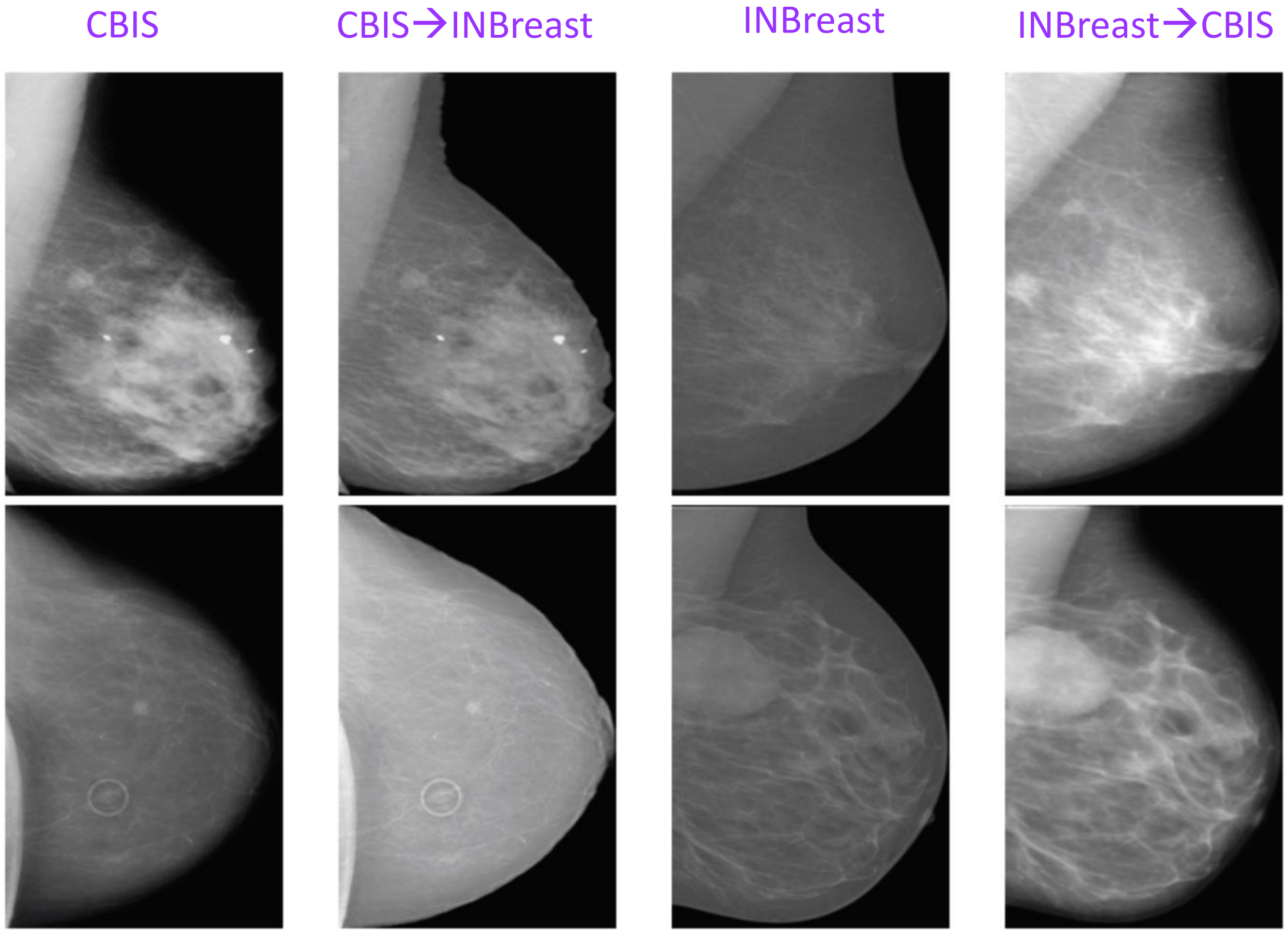
[Tao Wang, Adapting multiple datasets for better mammography tumor detection, KTH 2018]

Domain Adaptation



[Tao Wang, Adapting multiple datasets for better mammography tumor detection, KTH 2018]
[Ronneberger et al. "U-Net: Convolutional Networks for Biomedical Image Segmentation" MICCAI 2015

Domain Adaptation



[Tao Wang, Adapting multiple datasets for better mammography tumor detection, KTH 2018]

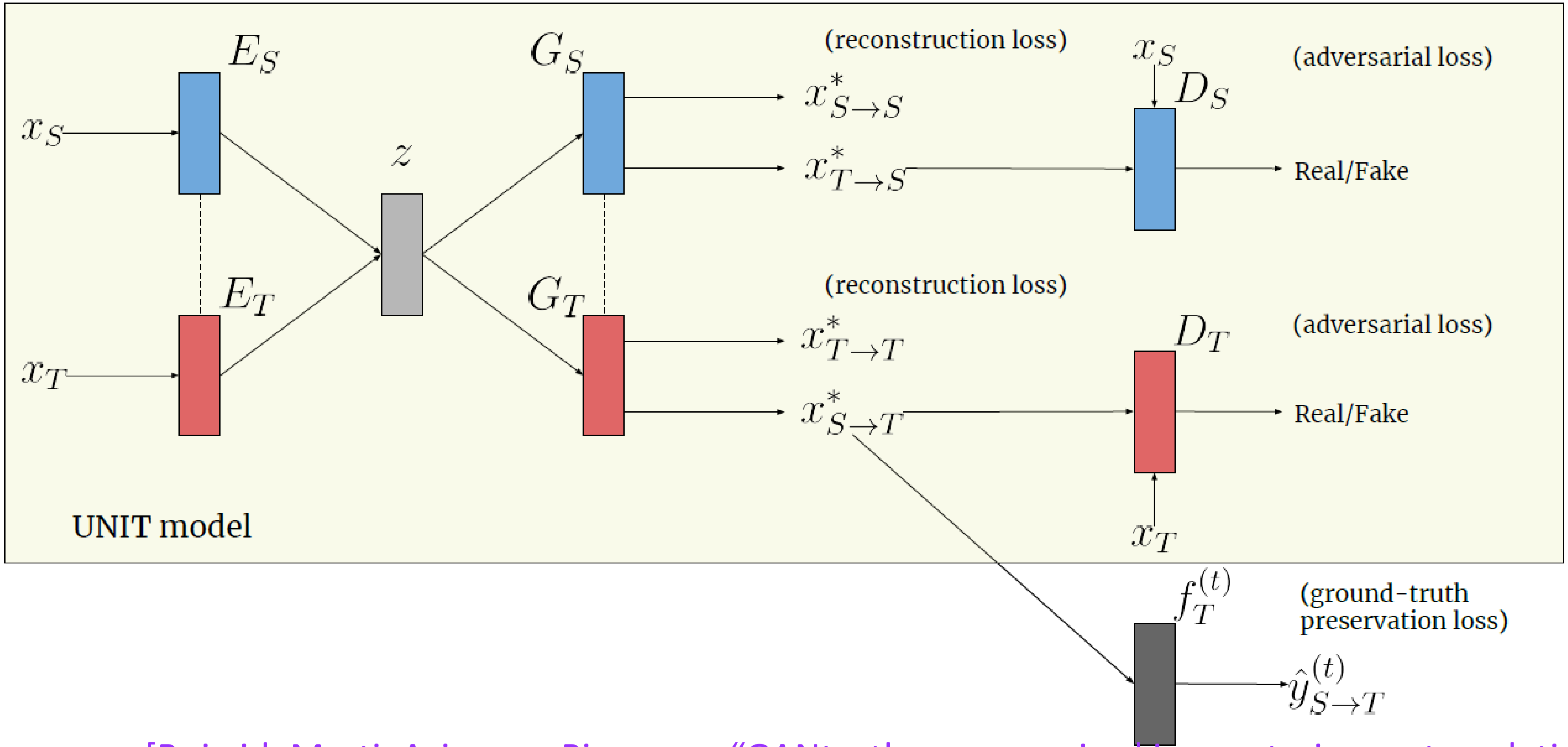
Domain Adaptation

		Pixel Level Metrics					Instance Level Metrics					
Training dataset	Testing dataset	Dice	Precision	Recall	MCC	IoU	$P_{0.25}$	$R_{0.25}$	$P_{0.5}$	$R_{0.5}$	AP	AR
CBIS	INBreast	0.1170	0.1965	0.1648	0.2029	0.1188	0.16	0.32	0.12	0.24	0.1072	0.2145
CBIS	Transferred INBreast	0.1412	0.1544	0.2934	0.2334	0.2108	0.1467	0.44	0.133	0.4	0.1151	0.3455
Transferred CBIS	INBreast	**0.1934**	**0.2521**	**0.3076**	**0.3377**	**0.2992**	**0.25**	**0.56**	**0.2321**	**0.52**	**0.1996**	**0.4472**

[Tao Wang, Adapting multiple datasets for better mammography tumor detection, KTH 2018]

Domain Adaptation

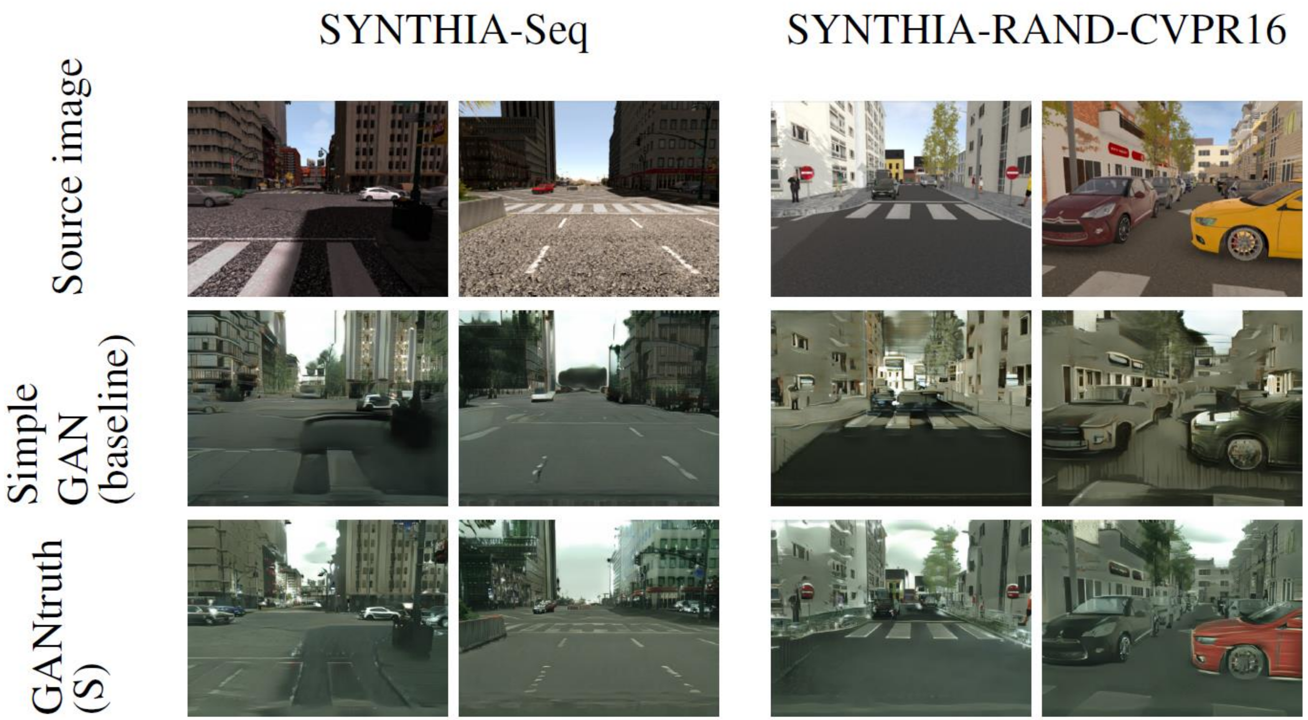
Synthetic-to-Real Adaptation



[Bujwid, Marti, Azizpour, Pieropan, “GANtruth – an unpaired image-to-image translation method for driving scenarios”, NeurIPS MLIT workshop 2018]

Domain Adaptation

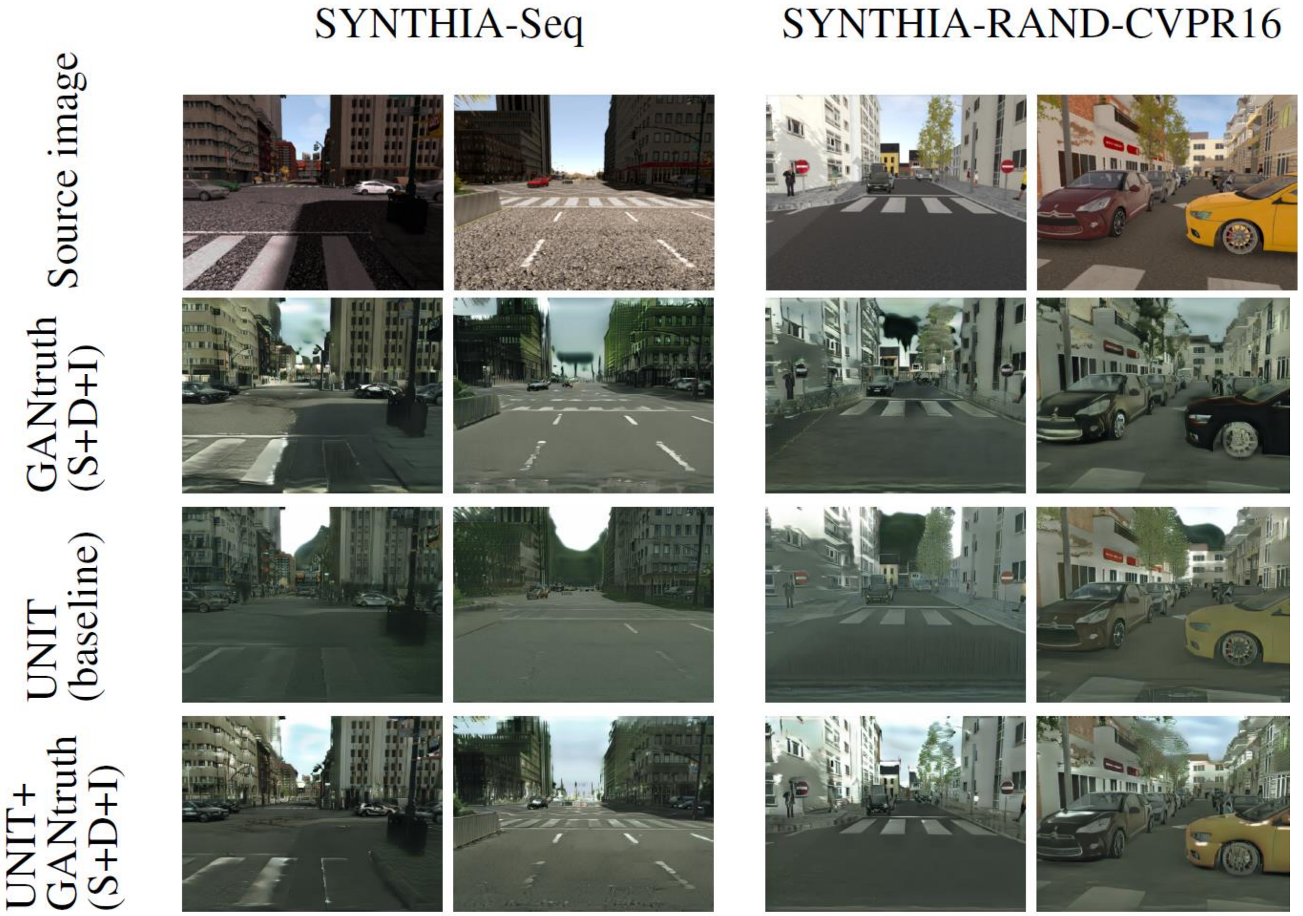
Synthetic-to-Real Adaptation



[Bujwid, Marti, Azizpour, Pieropan, “GANtruth – an unpaired image-to-image translation method for driving scenarios”, NeurIPS MLIT workshop 2018]

Domain Adaptation

Synthetic-to-Real Adaptation



[Bujwid, Marti, Azizpour, Pieropan, “GANtruth – an unpaired image-to-image translation method for driving scenarios”, NeurIPS MLIT workshop 2018]

Contents

Why do we care about uncertainty?

Sources

Model Uncertainty

- Small Dataset
- Covariate shift

Input Uncertainty

- Noisy measurements
- Incomplete input

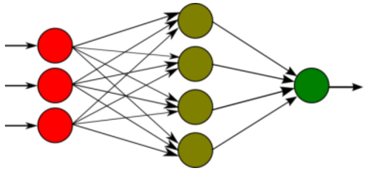
Applications

- Active Learning
- Exploration/ Exploitation
- Risk associated with certain decisions
- Model Pathology
- Forecasting
- Non-Stationary Environment

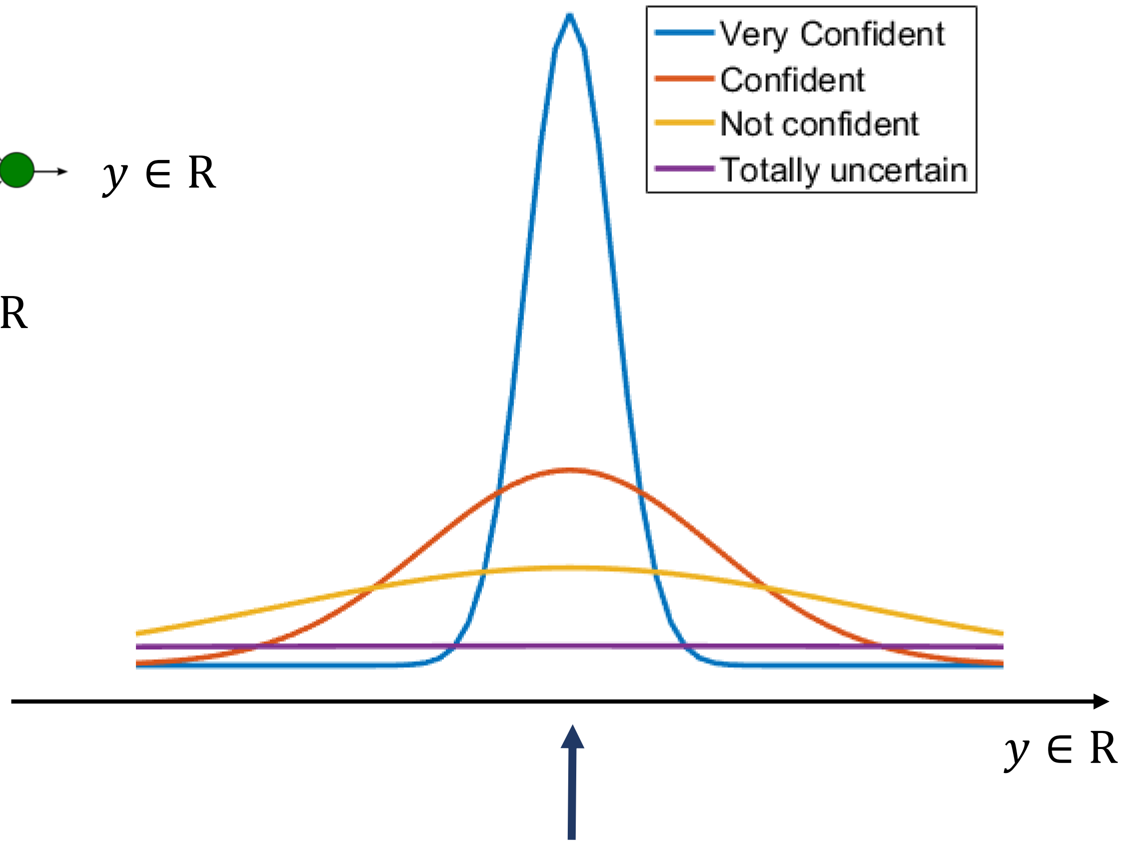
And many more…

What do we mean by modelling uncertainty?

A regression example

$x \in \mathrm{R}^d$  $y \in \mathrm{R}$

$f_\theta(x) \in \mathrm{R}^d \times \mathrm{R}$



point estimate $\hat{y}$

Multi-modal distributions!

Bayesian Modeling
for epistemic uncertainty

$$P(y|x,D) = \int P(y|x,D,\theta)\, P(\theta|D)\, d\theta$$

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

Uncertainty with Deep Networks

These approaches require redesigning the architecture and training procedure and are usually computationally expensive to train

- Directly output probability distributions [Le et al. 2005, Huang et al. 2016, Kendall&Gal 2017]
- Place distribution over model parameters [Denker&LeCun 1991, Neal 1995]
- Variational Approximation [Hinton&Camp 1993, Graves 2011, Blundell et al. 2015]
- Expectation Propagation [Jylanki et al. 2014, Soudry et al. 2014]
- Probabilistic Backpropagation [Rezenede et al. 2014, Lobato&Adams 2015]
- ….

Uncertainty with Deep Networks
No change in the training procedure

- Monte Carlo Dropout [Gal&Ghahramani 2016]
- Batch Normalization Dropout [Azizpour, Teye, Smith 2018]
- Frequentist Uncertainty [Lakshminarayanan 2017]
- Other approaches (e.g. entropy of the softmax distribution)

Stochastic Regularization as VI

Stochastic Regularization Techniques (SRT)

as approximate
Bayesian inference

[Gal&Ghahramani, Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning , JMLR 2016]

DropOut

$$\theta_i = \begin{cases} w_i \;\; with \;\; prob\, p \\ 0 \;\; with \;\; prob\, 1-p \end{cases}$$

[Gal&Ghahramani, Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning , JMLR 2016]

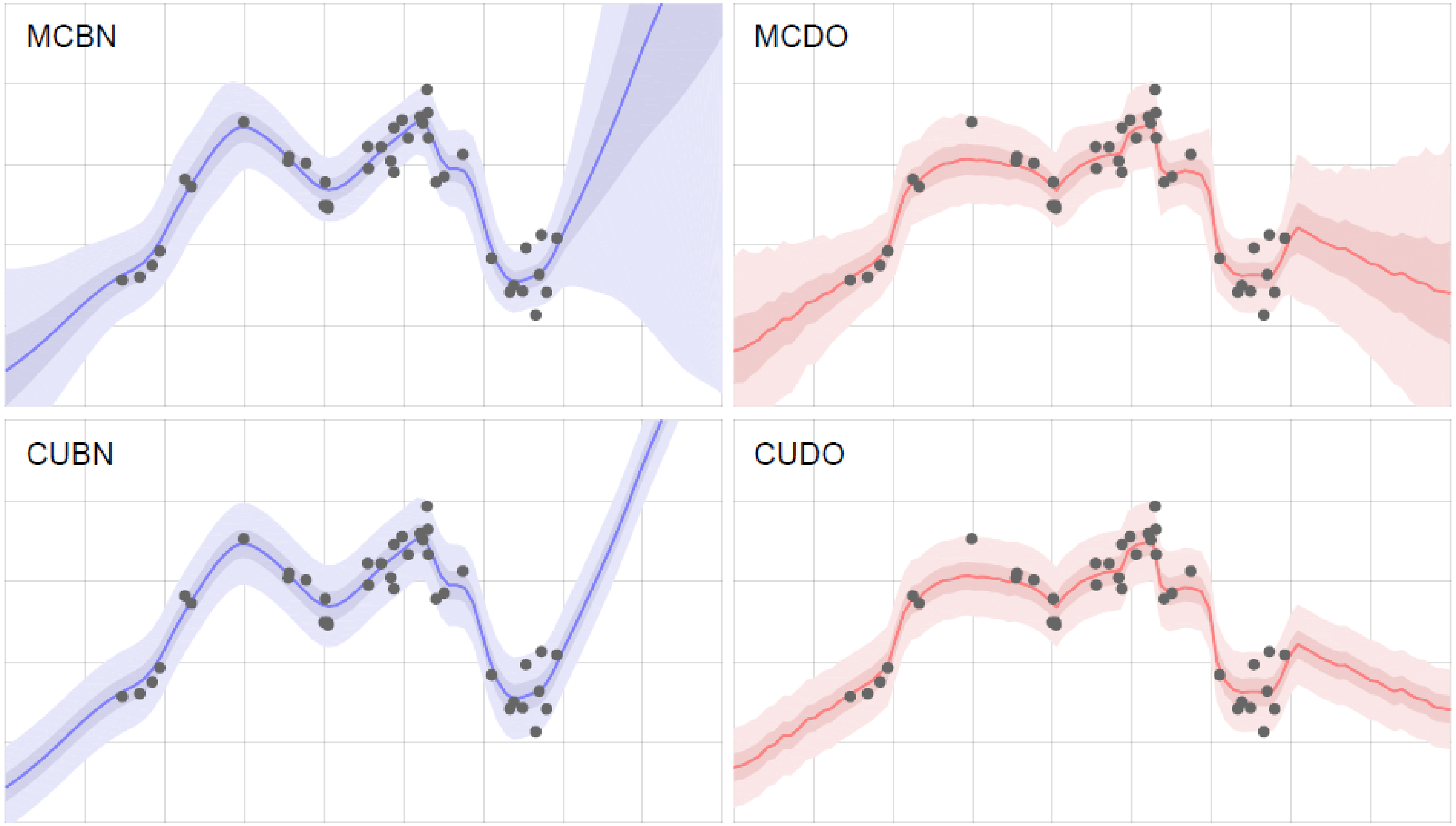
Batch Normalization

$$\overline{w_i} = \frac{w_i - \mu_B}{\sigma_B}$$

mini-batch
statistics

[Teye, Azizpour, Smith, "Bayesian Uncertainty Estimation For Batch Normalized Deep Networks" arXiv 2018]

Toy Dataset



[Teye, Azizpour, Smith, “Bayesian Uncertainty Estimation For Batch Normalized Deep Networks” arXiv 2018]

Quantitative Results

Regression - Normalized Measure

$$\sum_{i=1}^{N} \log P(\hat{y}_i = y_i)$$

Lower bound variance: τ^{-1}

Upper bound variance: σ^*

Dataset	PLL MCBN		PLL MCDO	
Boston Housing	10.49	****	5.51	****
Concrete	-36.36	**	10.92	****
Energy Efficiency	10.89	****	-14.28	*
Kinematics 8nm	1.68	***	-0.26	ns
Power Plant	0.33	**	3.52	****
Protein Tertiary Structure	2.56	****	6.23	****
Wine Quality (Red)	0.19	*	2.91	****
Yacht Hydrodynamics	45.58	****	-41.54	ns

[Teye, Azizpour, Smith, “Bayesian Uncertainty Estimation For Batch Normalized Deep Networks” arXiv 2018]

Quantitative Results

Classification - CIFAR

$$\sum_{i=1}^{N} \log P(\hat{y}_i = y_i)$$

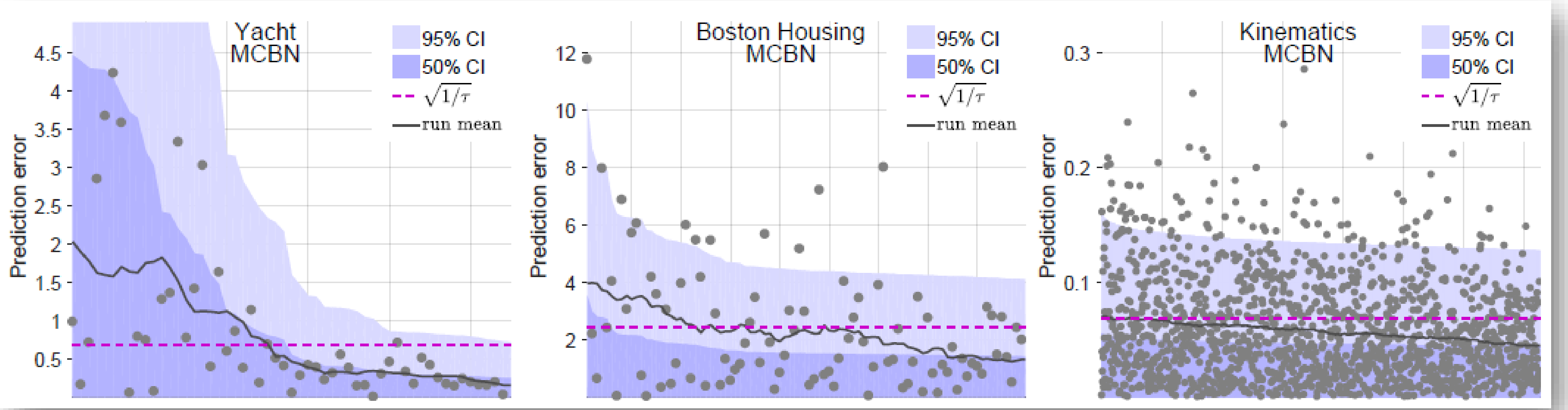
Baseline (standard network with dataset-average BN): **-0.32**

	Number of stochastic forward passes							
	1	2	4	8	16	32	64	128
PLL	-.36	-.32	-.30	-.29	-.29	-.28	-.28	**-.28**

[Teye, Azizpour, Smith, “Bayesian Uncertainty Estimation For Batch Normalized Deep Networks” arXiv 2018]

Qualitative Results

Error plot



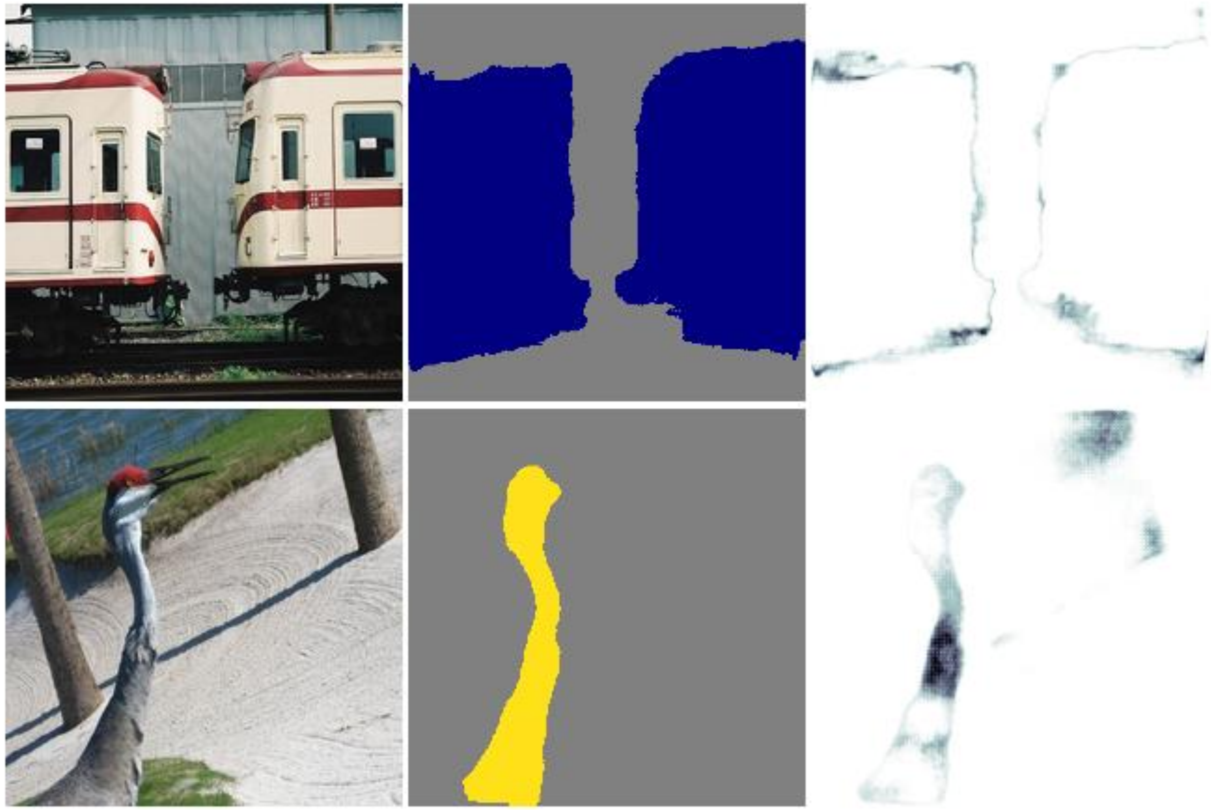
[Teye, Azizpour, Smith, "Bayesian Uncertainty Estimation For Batch Normalized Deep Networks" arXiv 2018]

Qualitative Results
Semantic Segmentation - CamVid

[Teye, Azizpour, Smith, “Bayesian Uncertainty Estimation For Batch Normalized Deep Networks” arXiv 2018]

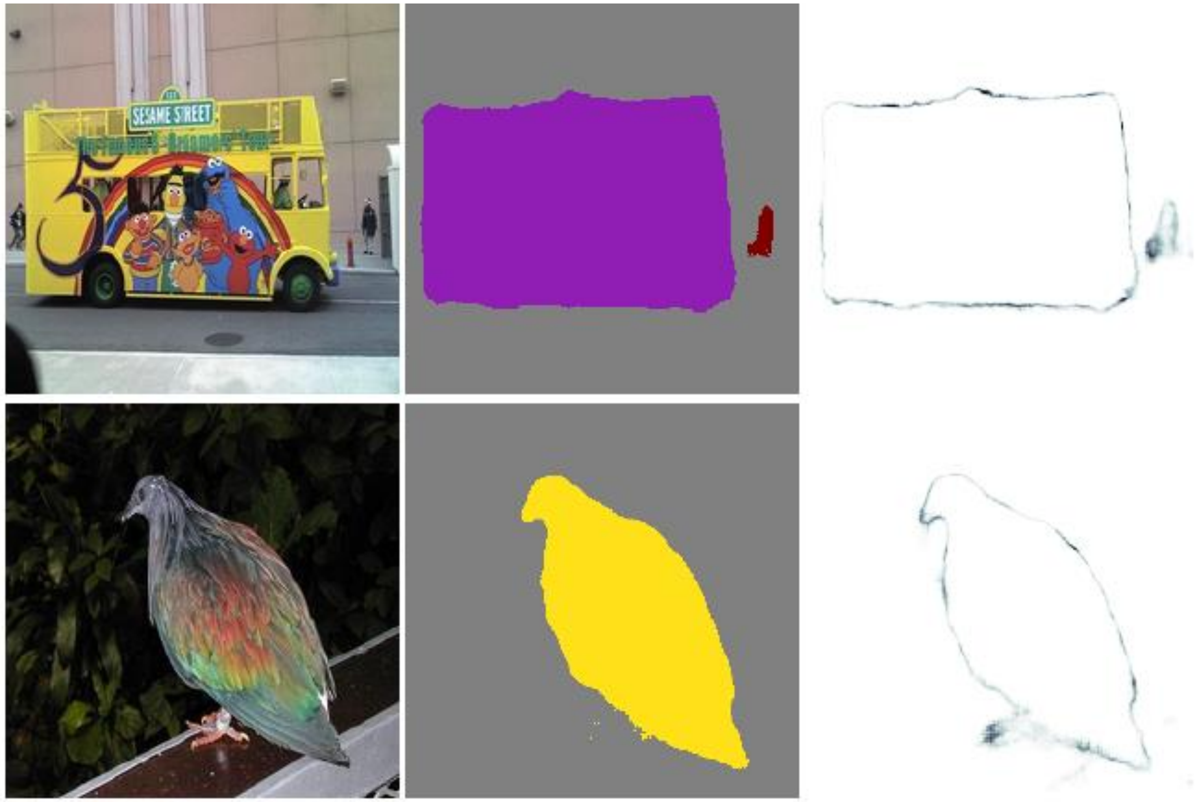
Qualitative Results

Semantic Segmentation – Pascal VOC



[Teye, Azizpour, Smith, “Bayesian Uncertainty Estimation For Batch Normalized Deep Networks” arXiv 2018]

Qualitative Results
Semantic Segmentation – Pascal VOC



[Teye, Azizpour, Smith, “Bayesian Uncertainty Estimation For Batch Normalized Deep Networks” arXiv 2018]

Simple Recipe

- Take any pretrained network with batch normalization and/or dropout layers
- Calculate constant observation noise on training/validation set, call it τ^{-1}
- At test time, sample different batches and/or dropout masks, get the predictions set
- Calculate the mean and standard deviation of the predictions, μ, σ
- The new point estimate of our prediction is: μ
- The associated uncertainty to it is: $\tau^{-1} + \sigma^2$

[Teye, Azizpour, Smith, “Bayesian Uncertainty Estimation For Batch Normalized Deep Networks” arXiv 2018]

Conclusion

- Positive points
  - Standard training
  - Simple algorithm
  - Vast applicability
- Negative points
  - Lots of assumptions
  - Under/over estimating the uncertainty
  - Considerable computation at test time

[Teye, Azizpour, Smith, “Bayesian Uncertainty Estimation For Batch Normalized Deep Networks” arXiv 2018]

Contents

Future Directions

- Geometric Deep Learning
- Side information (clinical information) as well as privileged information
- In a general scale: Multi-modal learning (unformatted textual description, images, clinical information, sequencing data)
- Probabilistic Uncertainty
- Weakly supervised learning (finding patterns that specialists are not aware of) new biomarkers, would make big leaps in life science
- Causality
- With missing data
- Encrypted networks
- Learning with noisy labels

Questions