Imperial College London

Stochastic Segmentation Networks:

Modelling Spatially Correlated Aleatoric Uncertainty

Miguel Monteiro, Loïc Le Folgoc, Daniel Coelho de Castro, Nick Pawlowski, Bernardo Marques, Konstantinos Kamnitsas, Mark van der Wilk, Ben Glocker

BioMedIA Imperial College London

Methods & Background

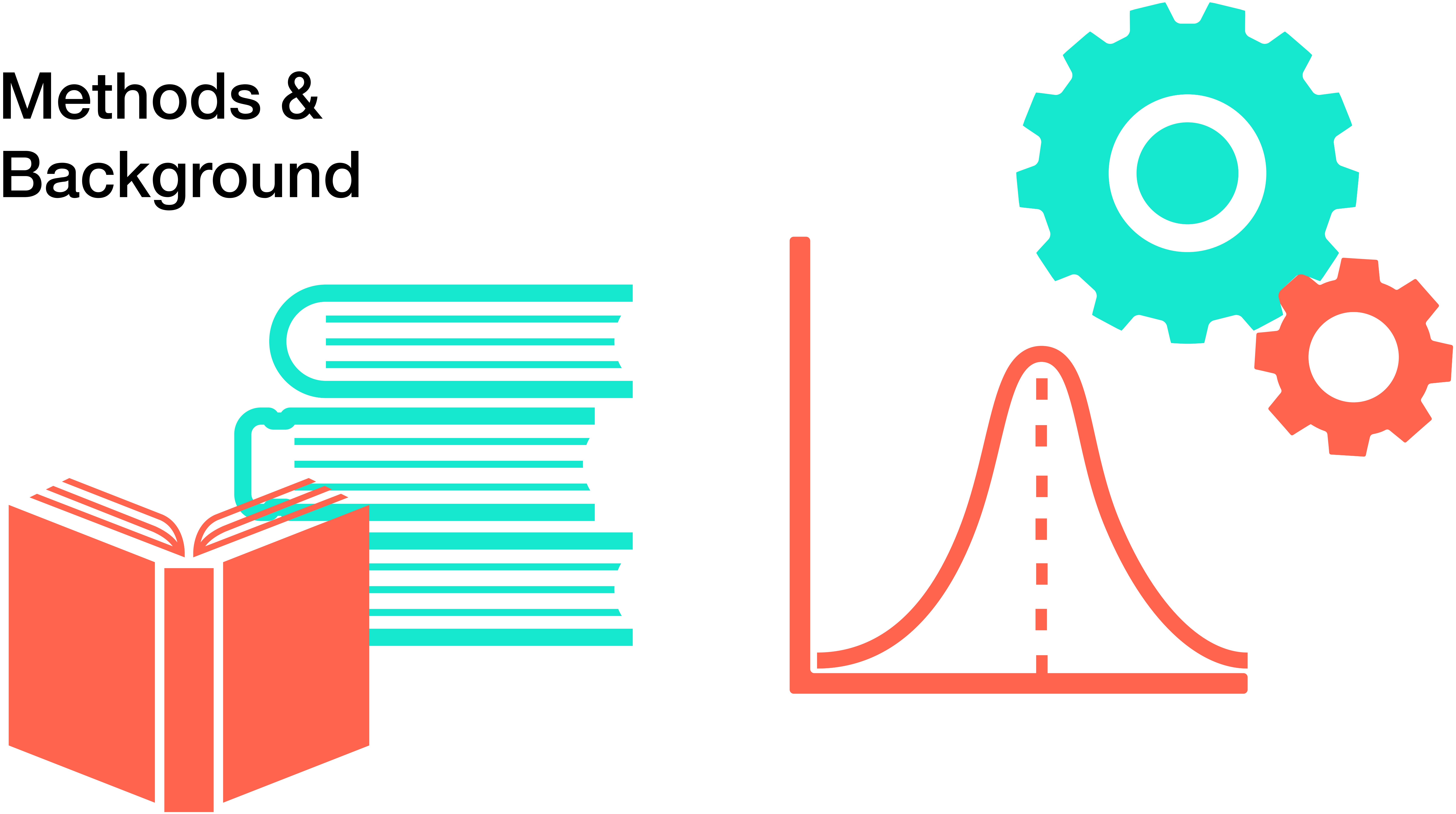
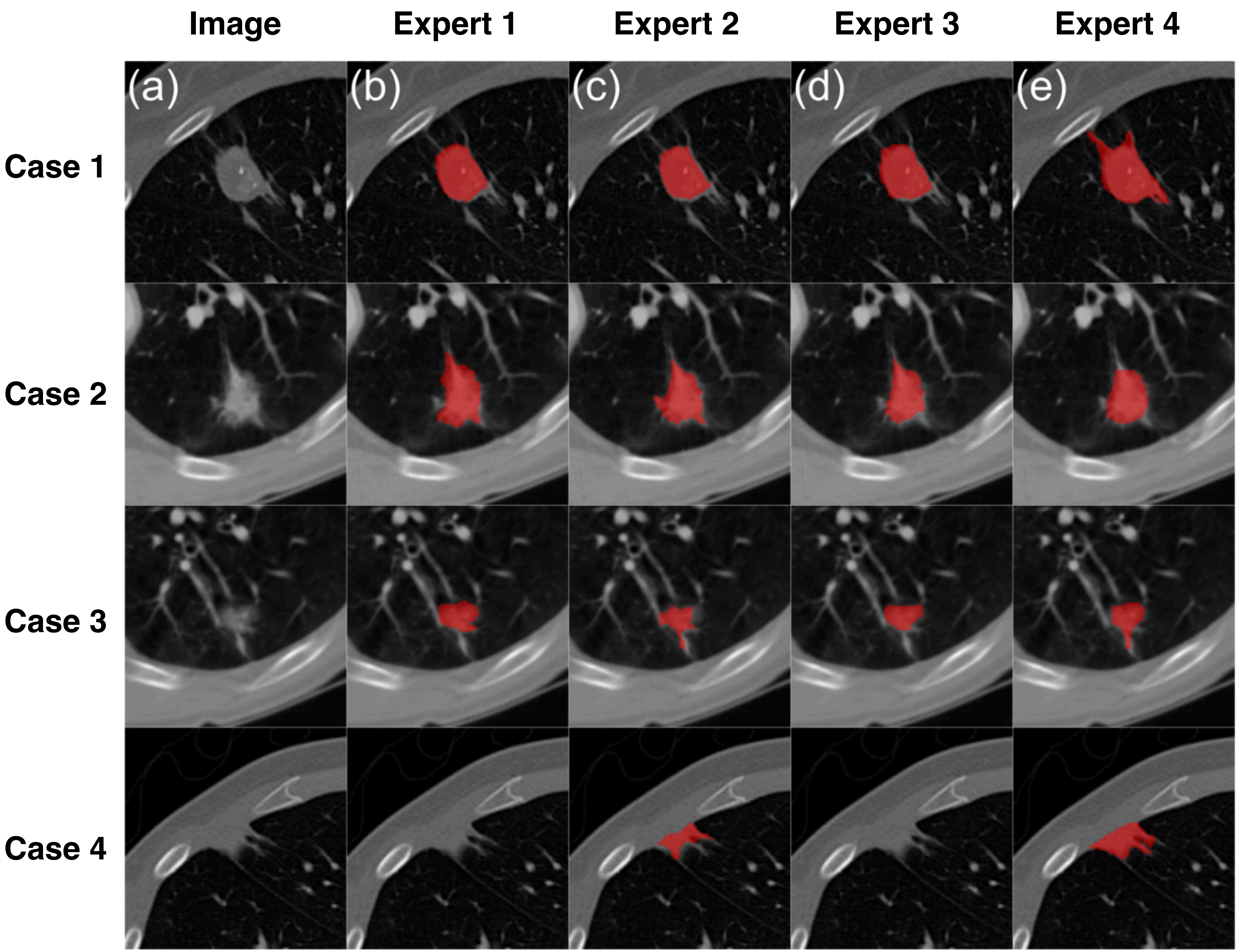
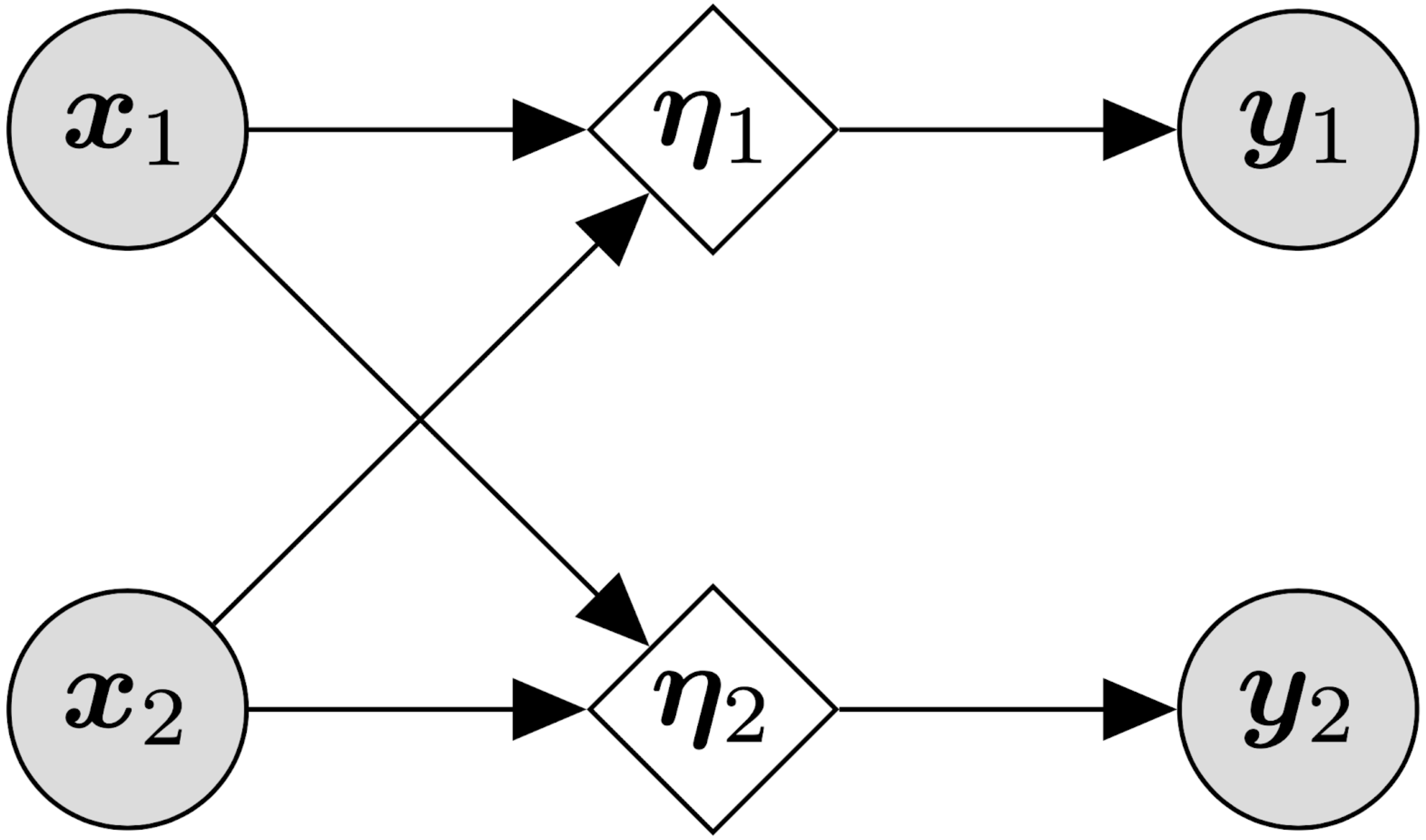


Image segmentation is inherently ambiguous.



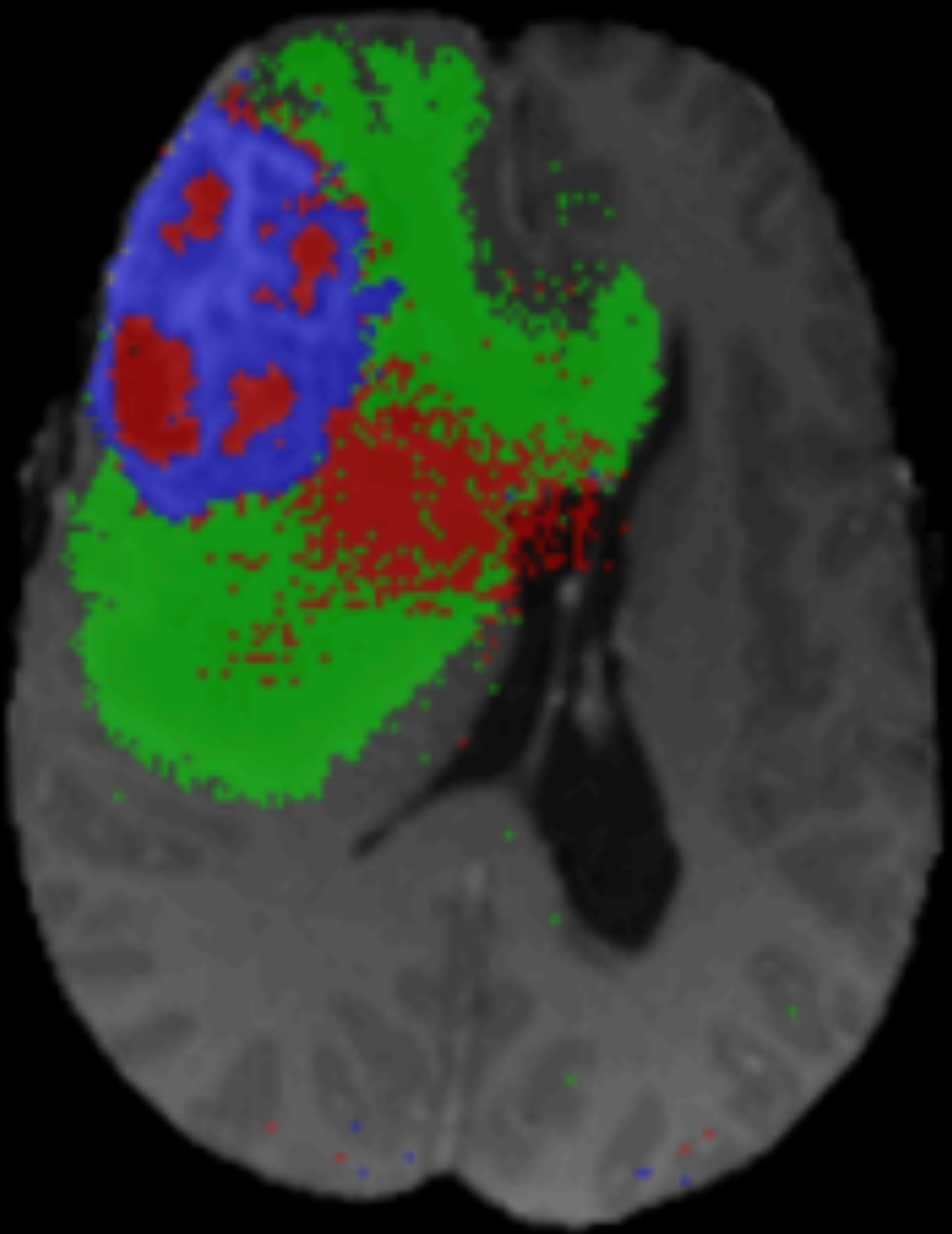


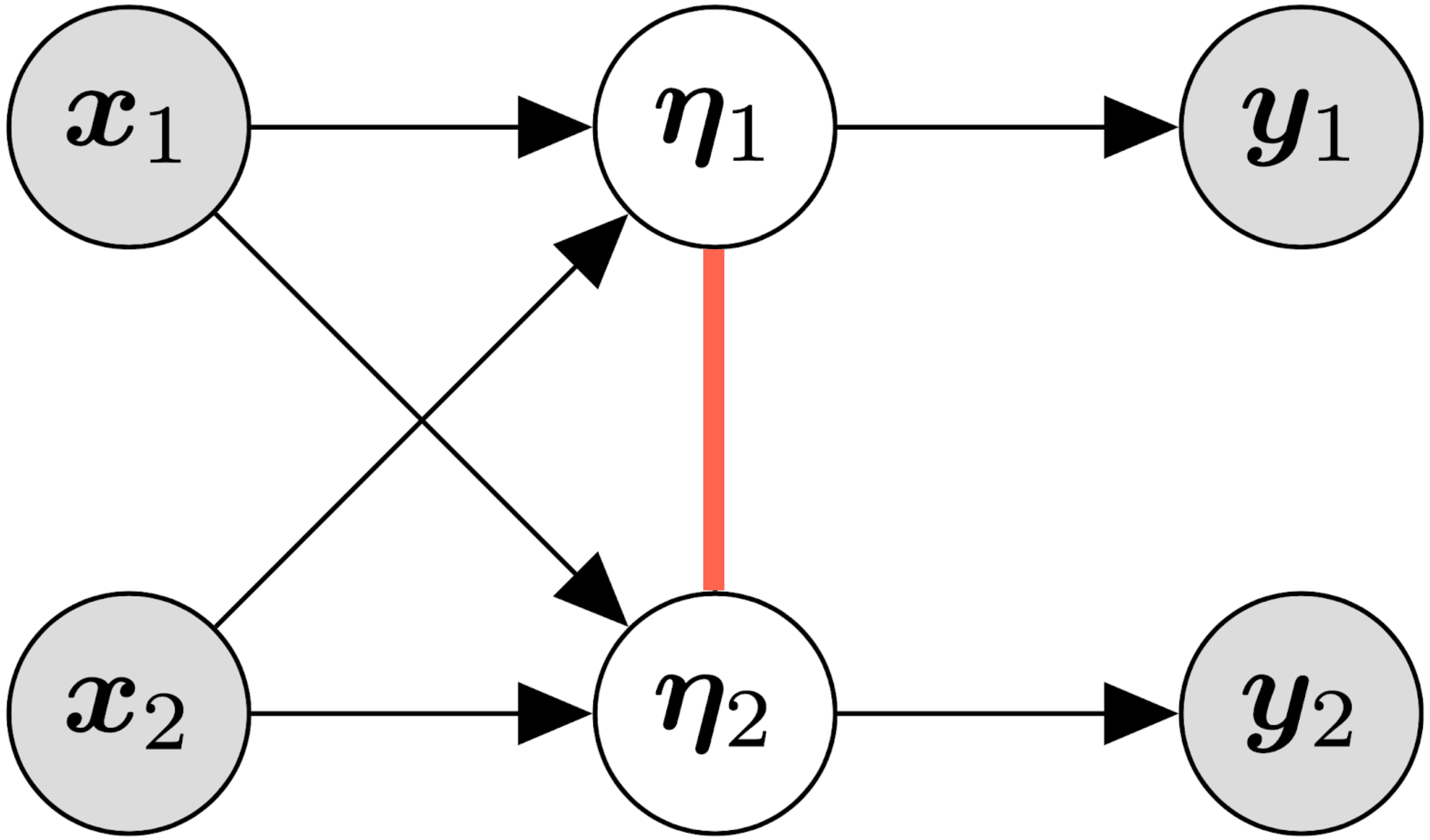
Assumptions:

1) Logits are independent given the image;
2) Labels are independent given the logits.

Pixel-wise independent loss function:

$$-\log p(\boldsymbol{y} \quad \boldsymbol{x}) = -\log \int p(\boldsymbol{y} \quad \boldsymbol{\eta}) p_{\phi}(\boldsymbol{\eta} \quad \boldsymbol{x}) d\boldsymbol{\eta} = -\log \prod_{i=1}^{\#pixels} p(\boldsymbol{y}_i \quad \boldsymbol{\eta}_i)$$





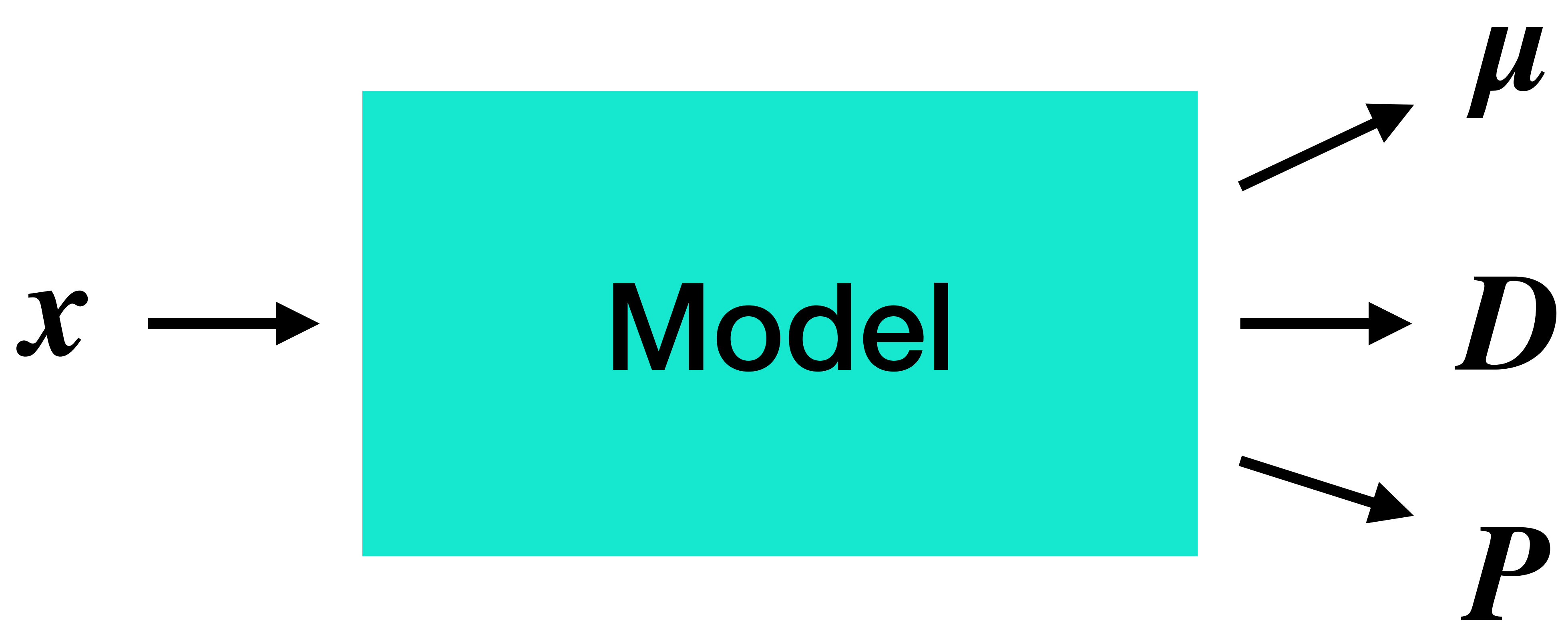
Make the logits dependent of each other by using a low rank multivariate normal

$$p(\boldsymbol{\eta} \quad \boldsymbol{x}) = \mathcal{N}(\boldsymbol{\eta} \quad \boldsymbol{\mu}(\boldsymbol{x}), \boldsymbol{\Sigma}(\boldsymbol{x}))$$

$$\boldsymbol{\Sigma} = \boldsymbol{P}\boldsymbol{P}^T + \boldsymbol{D}$$

Pixel dependencies are modelled in loss function:

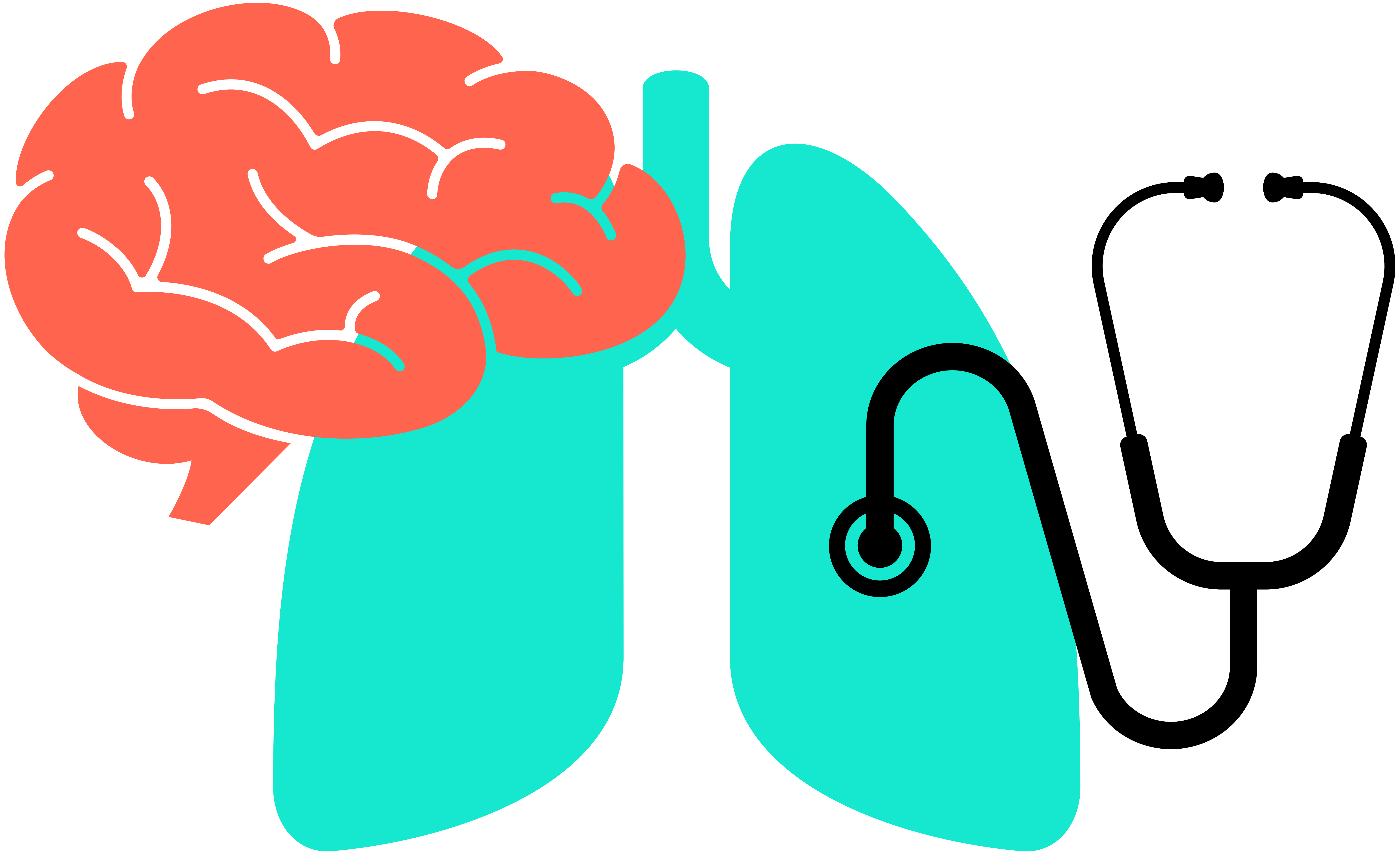
$$-\log p(\boldsymbol{y} \quad \boldsymbol{x}) = \approx -\log \frac{1}{M} \sum_{m=1}^{M} p(\boldsymbol{y} \quad \boldsymbol{\eta}^{(m)}) \ , \quad \boldsymbol{\eta}^{(m)} \quad \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{x}), \boldsymbol{\Sigma}(\boldsymbol{x}))$$



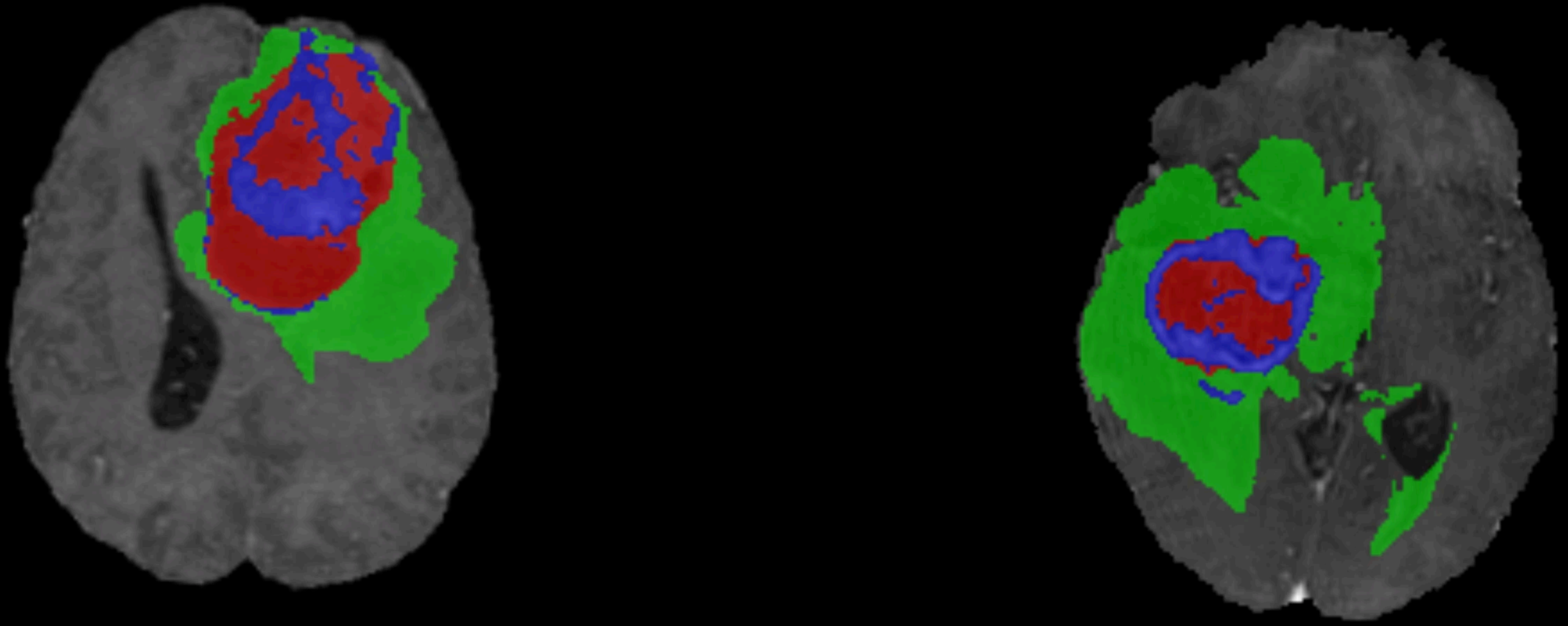
A full covariance matrix scales with the square of the number of pixels times the number of classes.

- The mean, $\boldsymbol{\mu}$, has shape: $n\_classes \times n\_pixels$;
- The covariance diagonal, $\boldsymbol{D}$, is a diagonal matrix with $n\_classes \times n\_pixels$ diagonal elements;
- The covariance factor, $\boldsymbol{P}$, is a rectangular matrix with shape: $(n\_classes \times n\_pixels) \times rank$.

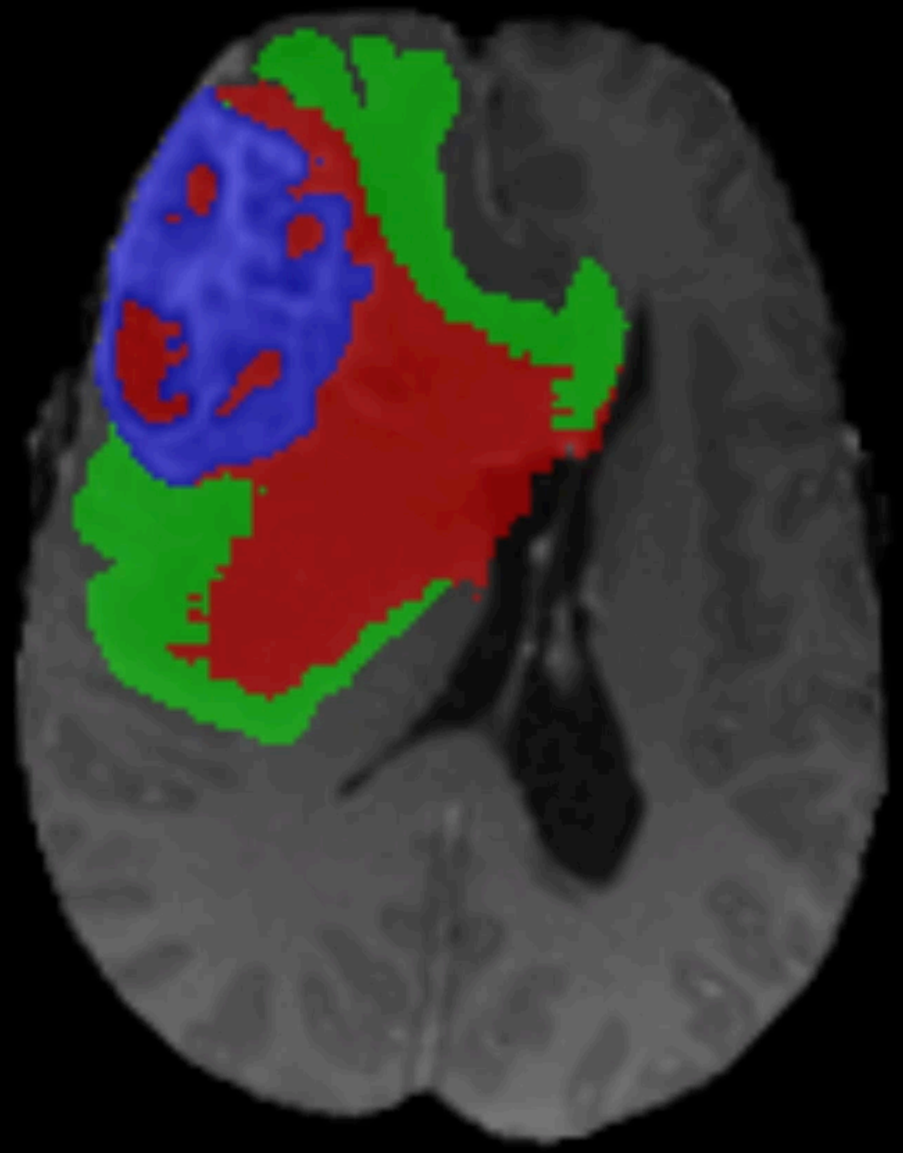
Results



Brain tumour segmentation



Real time sample manipulation



Are the samples better than the mean?

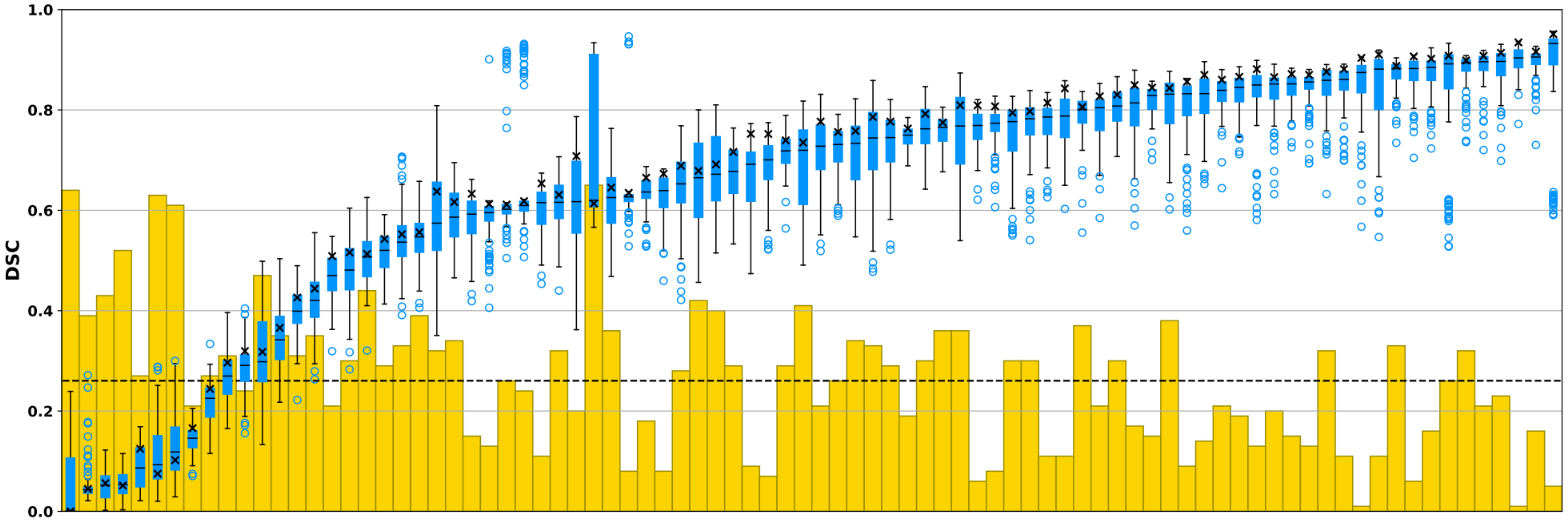
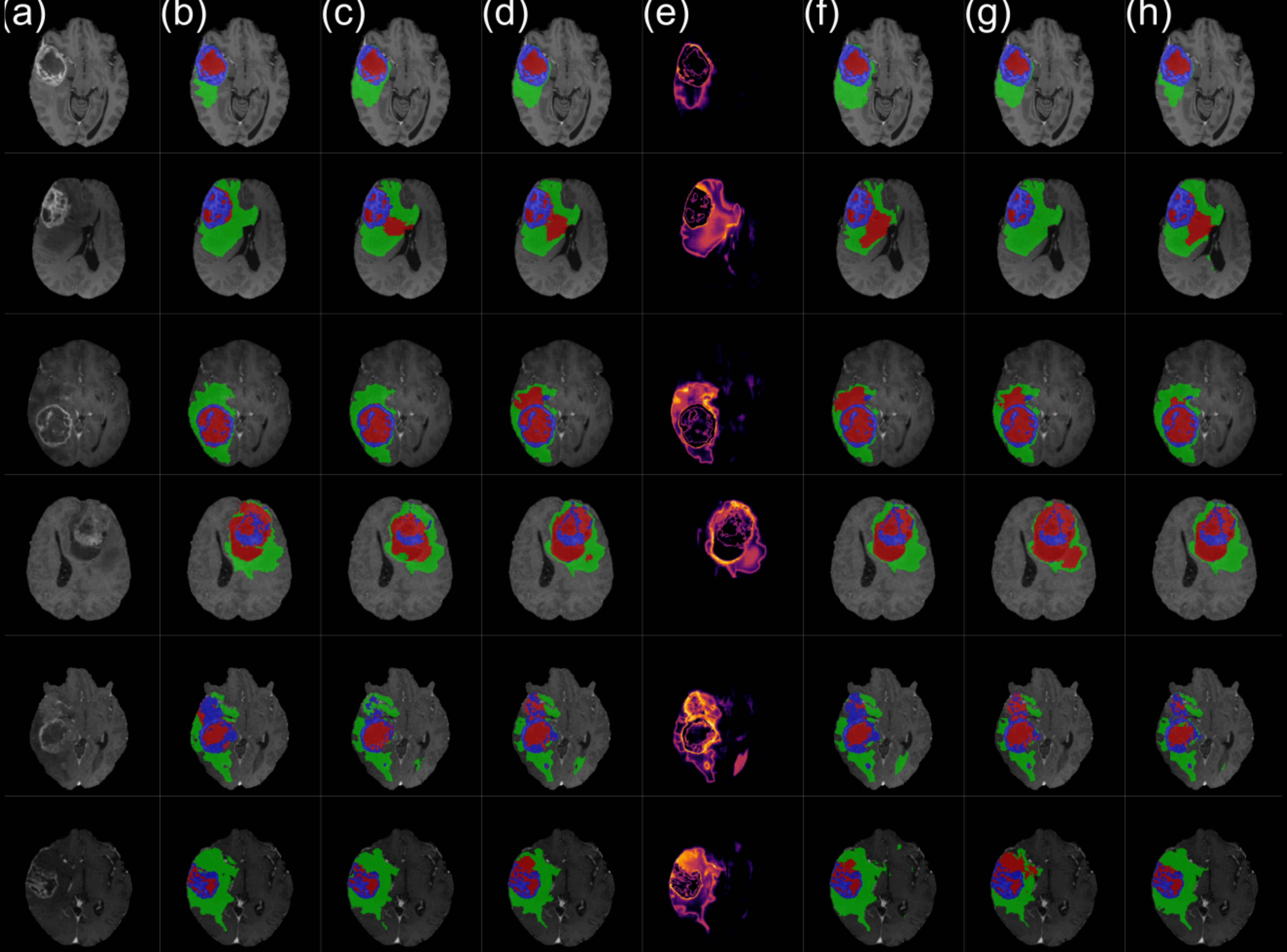


Figure 5: Distribution of sample average class DSC per case. The yellow bars denote the fraction of samples whose DSC is higher than the mean prediction, which is represented by a cross. The dashed line is the average fraction of samples better than the mean prediction (average height of the bars).


(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)

Lung nodule segmentation

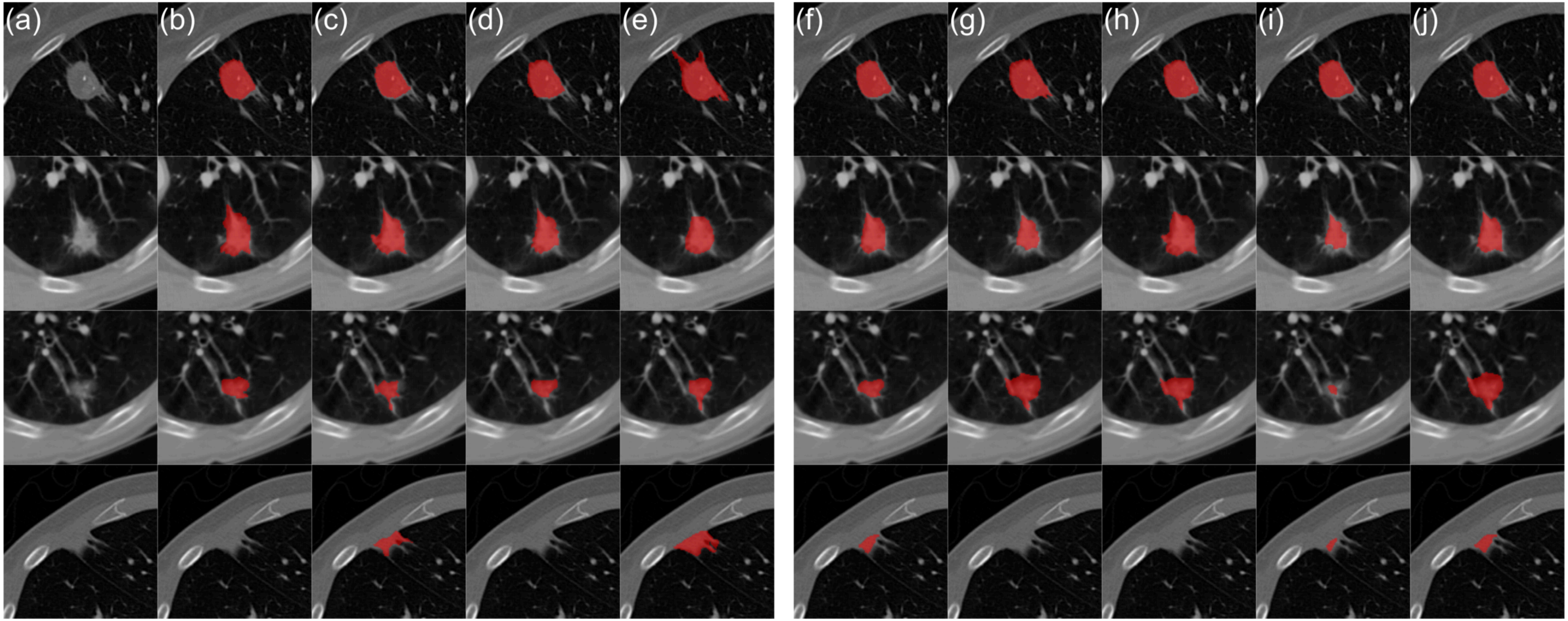


Figure 3: Qualitative results on the LIDC-IDRI dataset for the proposed model trained on four expert annotations: (a) CT image; (b-e) radiologist segmentations; (f) mean prediction; (g-j) samples.

Comparison to state of the art

model	trained on	DSC (%) ↑	DSC_{nod} (%) ↑	D^2_{GED} ↓	sample diversity
deterministic U-Net	set 0	37.5 ± 0.4	50.3 ± 0.4	0.698 ± 0.009	0.000 ± 0.000
probabilistic U-Net		38.4 ± 0.4	57.2 ± 0.4	0.516 ± 0.007	0.290 ± 0.004
PHiSeg		39.1 ± 0.4	51.3 ± 0.5	0.456 ± 0.008	0.215 ± 0.003
proposed (diagonal)		37.1 ± 0.4	51.2 ± 0.4	0.734 ± 0.009	0.001 ± 0.000
proposed (low-rank)		40.7 ± 0.4	58.6 ± 0.4	0.365 ± 0.005	0.399 ± 0.004
deterministic U-Net	all sets	35.9 ± 0.4	43.5 ± 0.5	0.607 ± 0.009	0.000 ± 0.000
probabilistic U-Net		39.0 ± 0.4	50.6 ± 0.5	0.252 ± 0.004	0.469 ± 0.003
PHiSeg		33.8 ± 0.4	40.3 ± 0.5	0.224 ± 0.004	0.496 ± 0.003
proposed (diagonal)		37.0 ± 0.4	46.2 ± 0.5	0.622 ± 0.009	0.007 ± 0.001
proposed (low-rank)		43.6 ± 0.4	68.5 ± 0.3	0.225 ± 0.002	0.609 ± 0.002

Comparison to state of the art

Better predictive performance for the mean sample.

model	trained on	DSC (%) $\uparrow$	DSC_{nod} (%) $\uparrow$	D^2_{GED} $\downarrow$	sample diversity
deterministic U-Net	set 0	37.5 ± 0.4	50.3 ± 0.4	0.698 ± 0.009	0.000 ± 0.000
probabilistic U-Net		38.4 ± 0.4	57.2 ± 0.4	0.516 ± 0.007	0.290 ± 0.004
PHiSeg		39.1 ± 0.4	51.3 ± 0.5	0.456 ± 0.008	0.215 ± 0.003
proposed (diagonal)		37.1 ± 0.4	51.2 ± 0.4	0.734 ± 0.009	0.001 ± 0.000
proposed (low-rank)		40.7 ± 0.4	58.6 ± 0.4	0.365 ± 0.005	0.399 ± 0.004
deterministic U-Net	all sets	35.9 ± 0.4	43.5 ± 0.5	0.607 ± 0.009	0.000 ± 0.000
probabilistic U-Net		39.0 ± 0.4	50.6 ± 0.5	0.252 ± 0.004	0.469 ± 0.003
PHiSeg		33.8 ± 0.4	40.3 ± 0.5	0.224 ± 0.004	0.496 ± 0.003
proposed (diagonal)		37.0 ± 0.4	46.2 ± 0.5	0.622 ± 0.009	0.007 ± 0.001
proposed (low-rank)		43.6 ± 0.4	68.5 ± 0.3	0.225 ± 0.002	0.609 ± 0.002

Comparison to state of the art

Better distance to expert distribution when trained on only one expert.

model	trained on	DSC (%) $\uparrow$	DSC_{nod} (%) $\uparrow$	$D^2_{GED} \downarrow$	sample diversity
deterministic U-Net	set 0	37.5 ± 0.4	50.3 ± 0.4	0.698 ± 0.009	0.000 ± 0.000
probabilistic U-Net		38.4 ± 0.4	57.2 ± 0.4	0.516 ± 0.007	0.290 ± 0.004
PHiSeg		39.1 ± 0.4	51.3 ± 0.5	0.456 ± 0.008	0.215 ± 0.003
proposed (diagonal)		37.1 ± 0.4	51.2 ± 0.4	0.734 ± 0.009	0.001 ± 0.000
proposed (low-rank)		40.7 ± 0.4	58.6 ± 0.4	0.365 ± 0.005	0.399 ± 0.004
deterministic U-Net	all sets	35.9 ± 0.4	43.5 ± 0.5	0.607 ± 0.009	0.000 ± 0.000
probabilistic U-Net		39.0 ± 0.4	50.6 ± 0.5	0.252 ± 0.004	0.469 ± 0.003
PHiSeg		33.8 ± 0.4	40.3 ± 0.5	0.224 ± 0.004	0.496 ± 0.003
proposed (diagonal)		37.0 ± 0.4	46.2 ± 0.5	0.622 ± 0.009	0.007 ± 0.001
proposed (low-rank)		43.6 ± 0.4	68.5 ± 0.3	0.225 ± 0.002	0.609 ± 0.002

Comparison to state of the art

model	trained on	DSC (%) ↑	DSC_{nod} (%) ↑	D^2_{GED} ↓	sample diversity
deterministic U-Net	set 0	37.5 ± 0.4	50.3 ± 0.4	0.698 ± 0.009	0.000 ± 0.000
probabilistic U-Net		38.4 ± 0.4	57.2 ± 0.4	0.516 ± 0.007	0.290 ± 0.004
PHiSeg		39.1 ± 0.4	51.3 ± 0.5	0.456 ± 0.008	0.215 ± 0.003
proposed (diagonal)		37.1 ± 0.4	51.2 ± 0.4	0.734 ± 0.009	0.001 ± 0.000
proposed (low-rank)		40.7 ± 0.4	58.6 ± 0.4	0.365 ± 0.005	0.399 ± 0.004
deterministic U-Net	all sets	35.9 ± 0.4	43.5 ± 0.5	0.607 ± 0.009	0.000 ± 0.000
probabilistic U-Net		39.0 ± 0.4	50.6 ± 0.5	0.252 ± 0.004	0.469 ± 0.003
PHiSeg		33.8 ± 0.4	40.3 ± 0.5	0.224 ± 0.004	0.496 ± 0.003
proposed (diagonal)		37.0 ± 0.4	46.2 ± 0.5	0.622 ± 0.009	0.007 ± 0.001
proposed (low-rank)		43.6 ± 0.4	68.5 ± 0.3	0.225 ± 0.002	0.609 ± 0.002

Equal or better distance to expert distribution when trained on all experts.

Comparison to state of the art

More diverse samples

model	trained on	DSC (%) ↑	DSC_{nod} (%) ↑	D^2_{GED} ↓	sample diversity
deterministic U-Net	set 0	37.5 ± 0.4	50.3 ± 0.4	0.698 ± 0.009	0.000 ± 0.000
probabilistic U-Net		38.4 ± 0.4	57.2 ± 0.4	0.516 ± 0.007	0.290 ± 0.004
PHiSeg		39.1 ± 0.4	51.3 ± 0.5	0.456 ± 0.008	0.215 ± 0.003
proposed (diagonal)		37.1 ± 0.4	51.2 ± 0.4	0.734 ± 0.009	0.001 ± 0.000
proposed (low-rank)		40.7 ± 0.4	58.6 ± 0.4	0.365 ± 0.005	0.399 ± 0.004
deterministic U-Net	all sets	35.9 ± 0.4	43.5 ± 0.5	0.607 ± 0.009	0.000 ± 0.000
probabilistic U-Net		39.0 ± 0.4	50.6 ± 0.5	0.252 ± 0.004	0.469 ± 0.003
PHiSeg		33.8 ± 0.4	40.3 ± 0.5	0.224 ± 0.004	0.496 ± 0.003
proposed (diagonal)		37.0 ± 0.4	46.2 ± 0.5	0.622 ± 0.009	0.007 ± 0.001
proposed (low-rank)		43.6 ± 0.4	68.5 ± 0.3	0.225 ± 0.002	0.609 ± 0.002

Algorithmic benefits over SOTA

- Infinitely many samples from one forward pass;
- No variational inference which requires tuning many regularisation hyper-parameters;
- Lightweight and flexible, enabling it to be used over any existing architecture, including 3D CNNs.

Demo