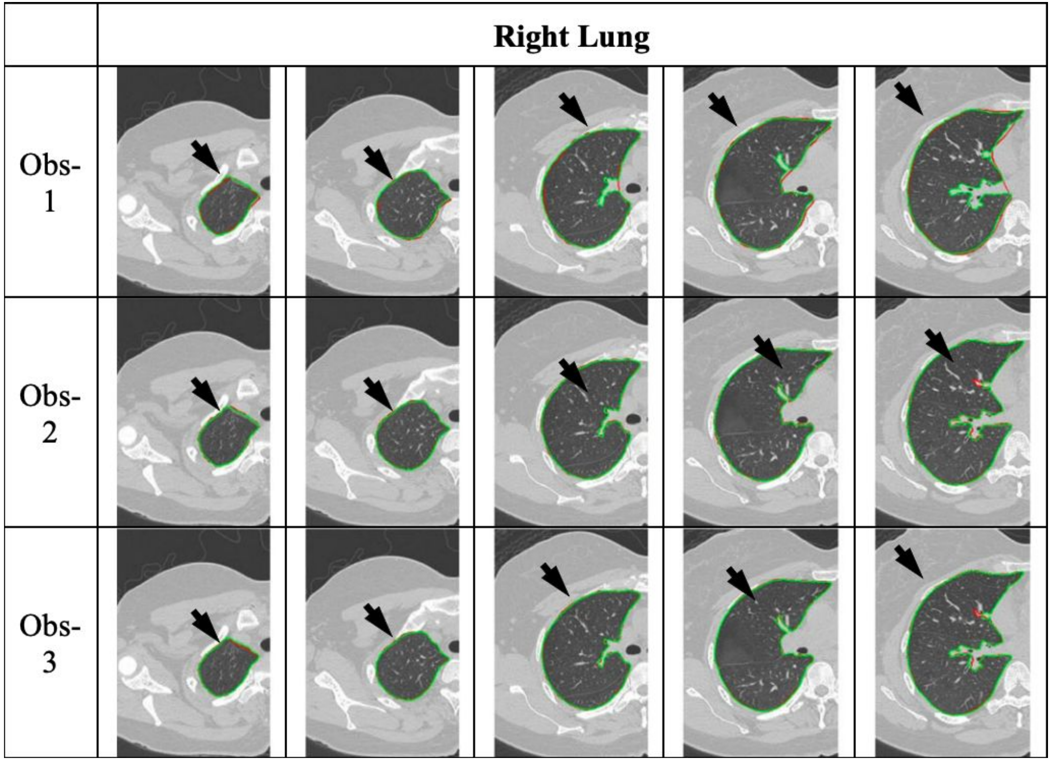
Disentangling Human Error from the Ground Truth in Segmentation of Medical Images

Le Zhang, Ryutaro Tanno, Mou-Cheng Xu, Chen Jin, Joseph Jacob, Olga Ciccarelli, Frederik Barkhof and Daniel C. Alexander

NeurIPS (2020)

Problem



- Segmentation of anatomical structures suffers from **high inter-observer variability** in medical images (biases, level of expertise, etc.)

- Noisy labels limits the performance of automatic segmentation algorithms

Inter-observer Variability Analysis of Automatic Lung Delineation in Normal and Disease Patients. Saba et al. Journal of Medical Systems (2016)

Problem

Existing approaches (to deal with multiple annotations):

- Majority vote
- Weighing annotation according to estimated reliability of expert (image-wise of pixel-wise)

Objective

- Jointly learn from noisy observations the **reliability of individual annotators** and the **true segmentation label distributions**

  → Produce a better segmentation at test-time

Hypothesis:

There is a **single, true segmentation map**, and each annotator produces a noisy approximation of it **according to its individual characteristics**

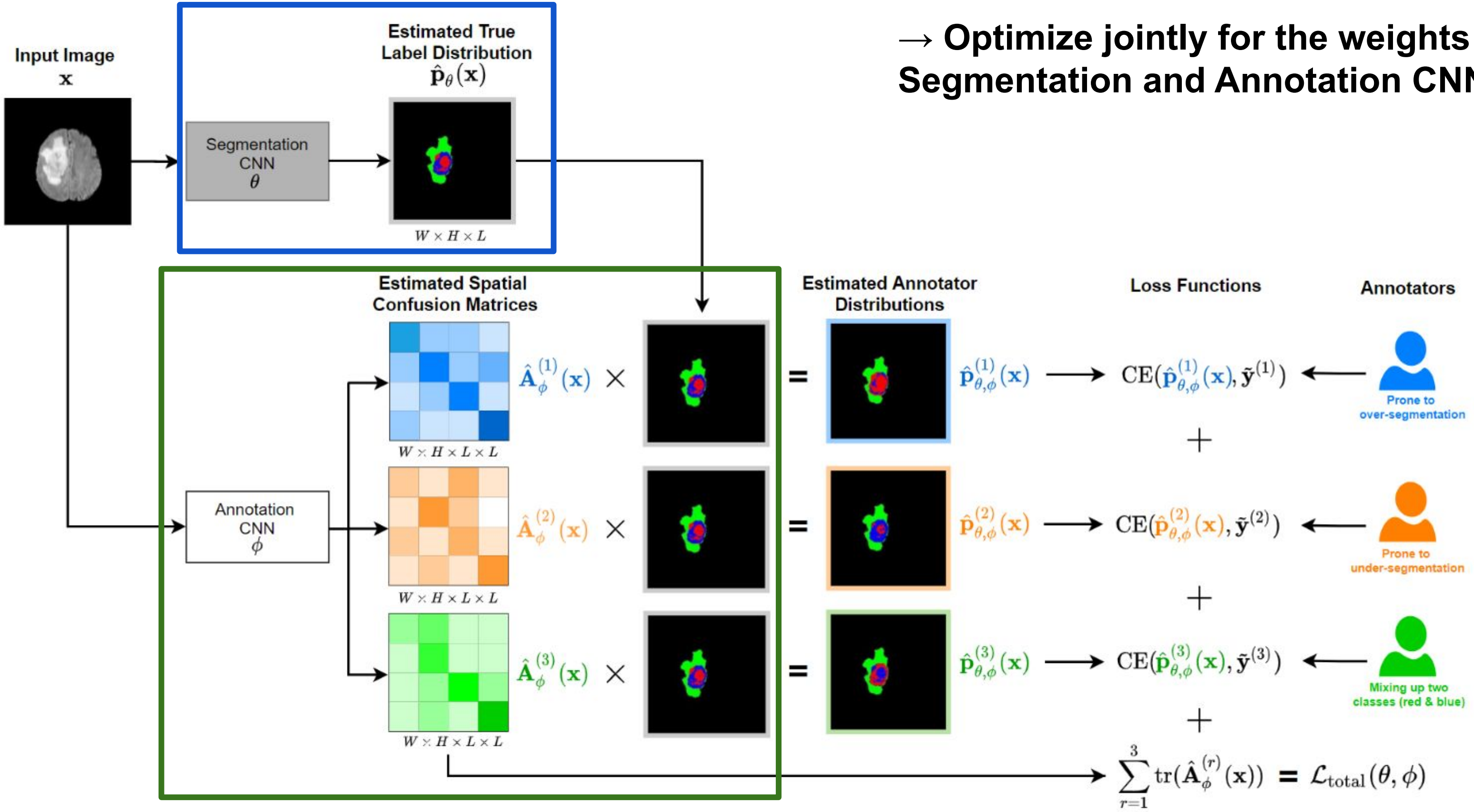
Method

- Jointly train a Segmentation CNN and Annotation CNN
- Set of images, with multiple noisy segmentation masks

Assumptions:

1) **Annotators** are statistically **independent**
2) **Annotations noise** is **independent** of input image

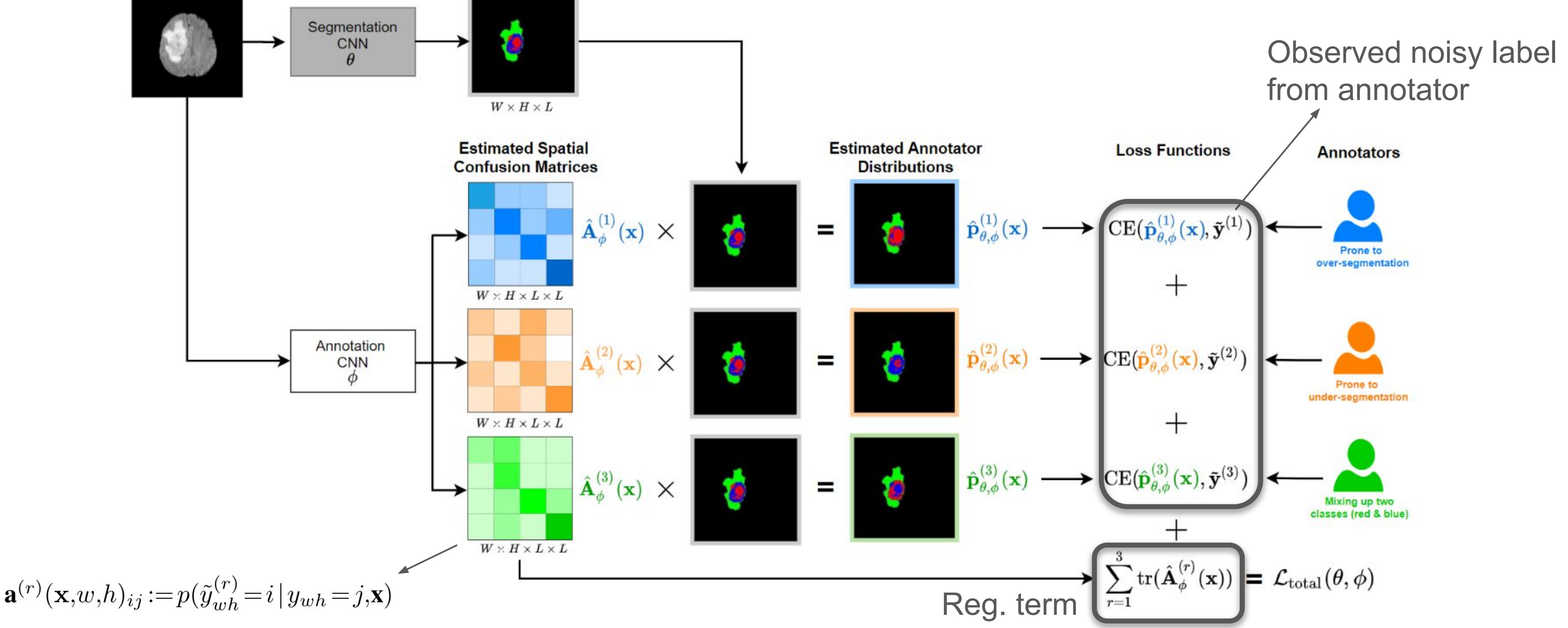
Method



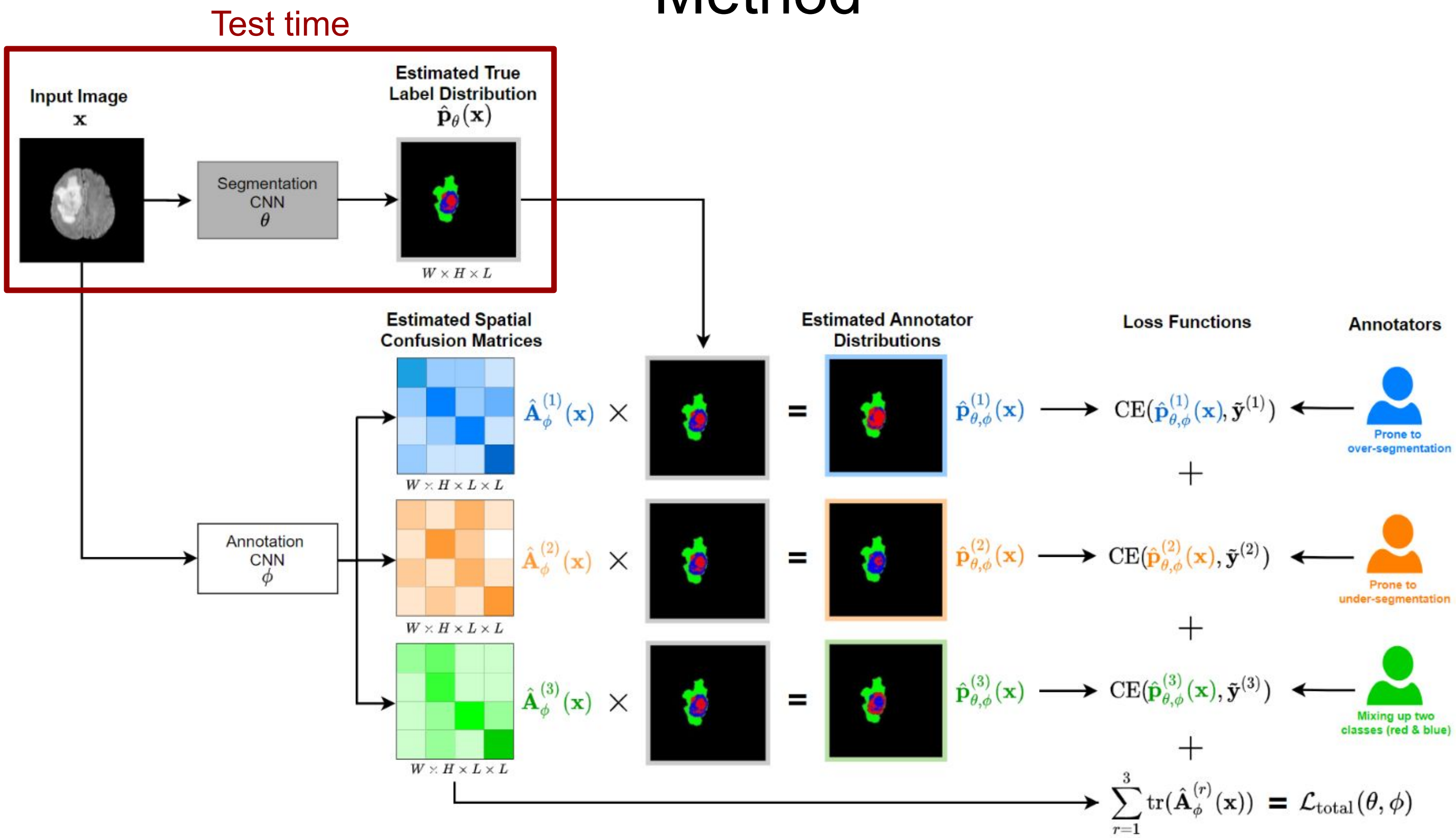
→ **Optimize jointly for the weights of the Segmentation and Annotation CNN**

Method

Minimising CE → ensures fidelity with observed noisy annotators
Minimising trace → encourages estimated annotators to be maximally unreliable



Method


Test time
Input Image
x
Segmentation CNN θ
Estimated True Label Distribution
$\hat{\mathbf{p}}_\theta(\mathbf{x})$
$W \times H \times L$
Annotation CNN ϕ
Estimated Spatial Confusion Matrices
$W \times H \times L \times L$
Estimated Annotator Distributions
Loss Functions
Annotators
$\hat{\mathbf{A}}^{(1)}_\phi(\mathbf{x}) \times$ =
$\hat{\mathbf{p}}^{(1)}_{\theta,\phi}(\mathbf{x}) \rightarrow \mathrm{CE}(\hat{\mathbf{p}}^{(1)}_{\theta,\phi}(\mathbf{x}), \tilde{\mathbf{y}}^{(1)})$
Prone to over-segmentation
$\hat{\mathbf{A}}^{(2)}_\phi(\mathbf{x}) \times$ =
$\hat{\mathbf{p}}^{(2)}_{\theta,\phi}(\mathbf{x}) \rightarrow \mathrm{CE}(\hat{\mathbf{p}}^{(2)}_{\theta,\phi}(\mathbf{x}), \tilde{\mathbf{y}}^{(2)})$
Prone to under-segmentation
$\hat{\mathbf{A}}^{(3)}_\phi(\mathbf{x}) \times$ =
$\hat{\mathbf{p}}^{(3)}_{\theta,\phi}(\mathbf{x}) \rightarrow \mathrm{CE}(\hat{\mathbf{p}}^{(3)}_{\theta,\phi}(\mathbf{x}), \tilde{\mathbf{y}}^{(3)})$
Mixing up two classes (red & blue)
$+ \sum_{r=1}^{3} \mathrm{tr}(\hat{\mathbf{A}}^{(r)}_\phi(\mathbf{x})) = \mathcal{L}_{\text{total}}(\theta, \phi)$

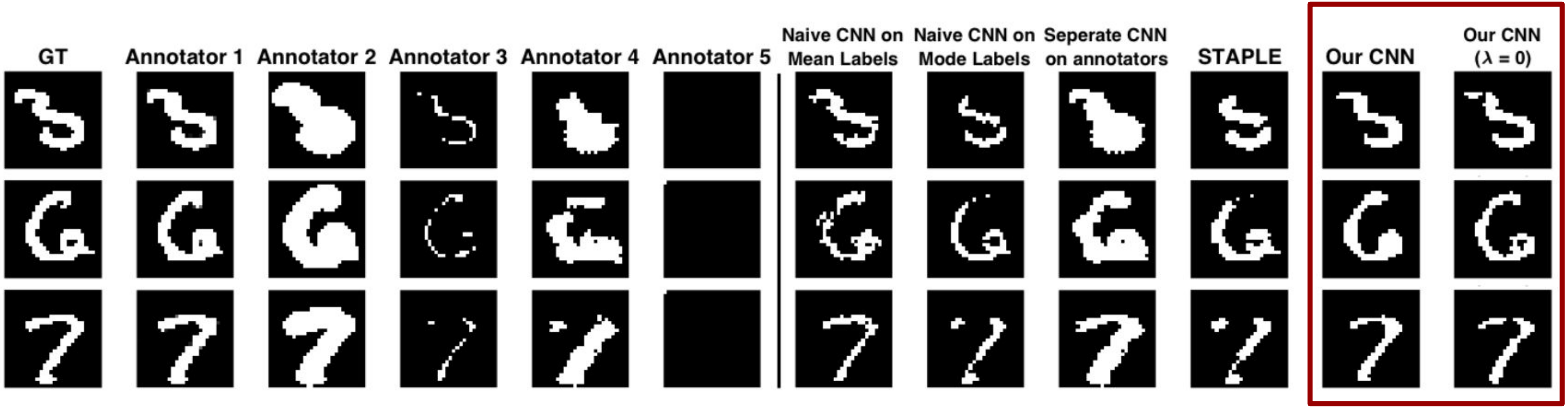
Experiment

- Dataset
  - MNIST
  - MSLSC (multiple-sclerosis lesions)
  - BraTS (brain tumours)
  - LIDC-IDRI (lung abnormalities) → multiple annotations

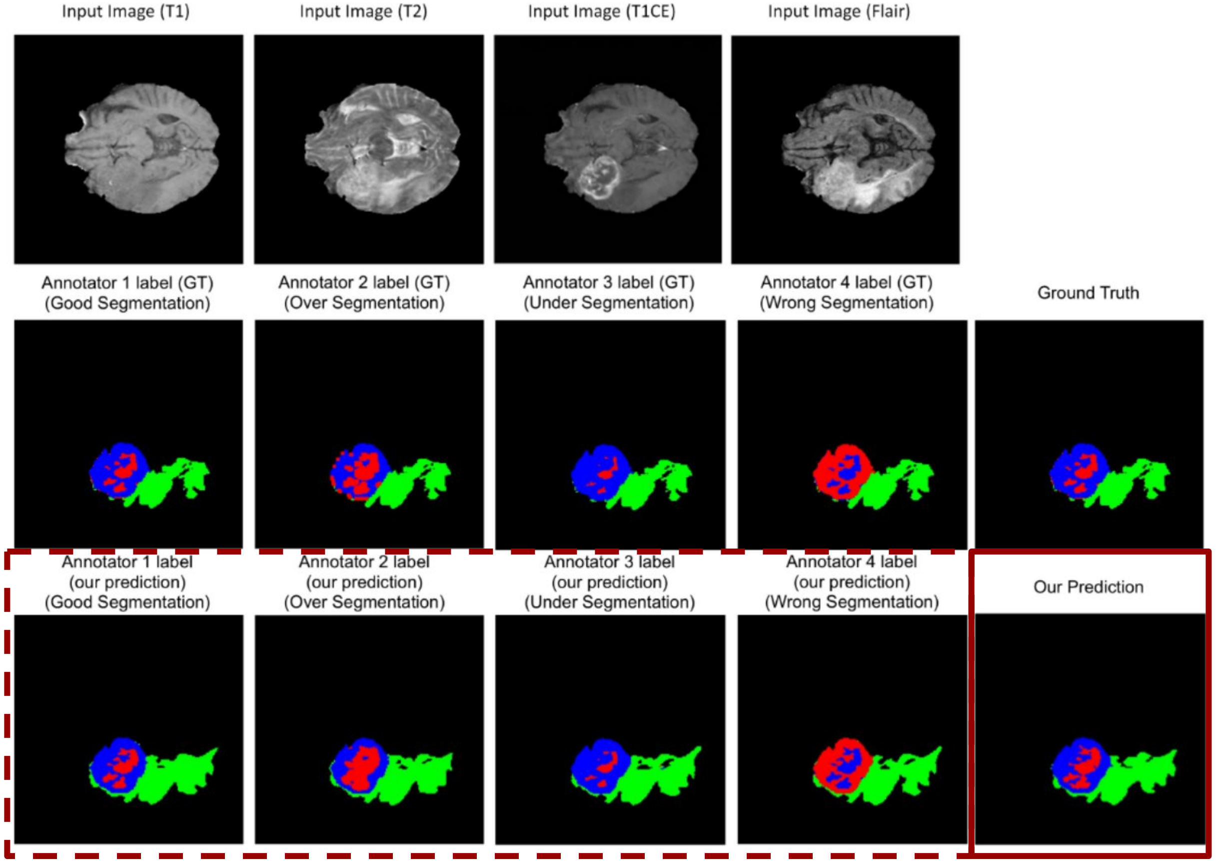
Synthetically noisy labels

- Baselines
  - Mean of noisy labels
  - Majority vote
  - STAPLE
  - Spatial STAPLE
  - Probabilistic UNet

Results


GT
Annotator 1
Annotator 2
Annotator 3
Annotator 4
Annotator 5
Naive CNN on Mean Labels
Naive CNN on Mode Labels
Seperate CNN on annotators
STAPLE
Our CNN
Our CNN ($\lambda = 0$)

Results


Input Image (T1)
Input Image (T2)
Input Image (T1CE)
Input Image (Flair)
Annotator 1 label (GT)
(Good Segmentation)
Annotator 2 label (GT)
(Over Segmentation)
Annotator 3 label (GT)
(Under Segmentation)
Annotator 4 label (GT)
(Wrong Segmentation)
Ground Truth
Annotator 1 label
(our prediction)
(Good Segmentation)
Annotator 2 label
(our prediction)
(Over Segmentation)
Annotator 3 label
(our prediction)
(Under Segmentation)
Annotator 4 label
(our prediction)
(Wrong Segmentation)
Our Prediction

Results

Models	BraTS DICE (%)	BraTS CM estimation	LIDC-IDRI DICE (%) ↑	LIDC-IDRI CM estimation ↓
Naive CNN on mean labels	29.42 ± 0.58	n/a	56.72 ± 0.61	n/a
Naive CNN on mode labels	34.12 ± 0.45	n/a	58.64 ± 0.47	n/a
Probabilistic U-net [24]	40.53 ± 0.75	n/a	61.26 ± 0.69	n/a
STAPLE [9]	46.73 ± 0.17	0.2147 ± 0.0103	69.34 ± 0.58	0.0832 ± 0.0043
Spatial STAPLE [14]	47.31 ± 0.21	0.1871 ± 0.0094	70.92 ± 0.18	0.0746 ± 0.0057
Ours with Global CMs	47.33 ± 0.28	0.1673 ± 0.1021	70.94 ± 0.19	0.1386 ± 0.0052
Ours without Trace	49.03 ± 0.34	0.1569 ± 0.0072	71.25 ± 0.12	0.0482 ± 0.0038
Ours	$\mathbf{53.47 \pm 0.24}$	$\mathbf{0.1185 \pm 0.0056}$	$\mathbf{74.12 \pm 0.19}$	$\mathbf{0.0451 \pm 0.0025}$
Oracle (Ours but with known CMs)	67.13 ± 0.14	0.0843 ± 0.0029	79.41 ± 0.17	0.0381 ± 0.0021

Conclusion

A supervised segmentation method for jointly estimating the spatial **characteristics of labelling errors** (annotation noise) from multiple human annotators and the **ground-truth label distribution.**

→ Finds the **maximal amount of confusion** which explains the noisy observations well.

→ Improves **robustness** against label noise