Abstention for Noise-Robust Learning in Medical Image Segmentation

presented by:
Wesam Moustafa [Mat. Nr. 3410585]

First Examiner:
Prof. Dr. Rafet Sifa

Second Examiner:
Prof. Dr. Christian Bauckhage

Supervisors:
Prof. Dr. Rafet Sifa & Dr. Helen Schneider

July 17, 2025

Table of contents

Introduction

Label Noise

What is Label Noise?

- Errors or inaccuracies in ground truth training labels.
- A pervasive problem in real-world datasets.

Why is it Bad?

- Deep Neural Networks tend to memorize these errors.
- This leads to poor generalization and unreliable models.

Amplified in Image Segmentation

- Segmentation demands pixel-perfect accuracy.
- This makes the annotation process uniquely tedious and error-prone, especially at object boundaries.

The Annotation Bottleneck

- Acquiring clean labels is extremely difficult and expensive.
- Requires time from scarce, highly-trained medical experts.
- Subject to significant inter-observer variability (experts disagree).

The High Stakes of Failure

- Medical segmentation is a critical, safety-sensitive task.
- Inaccurate models can directly impact patient diagnosis.
- Urgent need for models that are robust to noise.

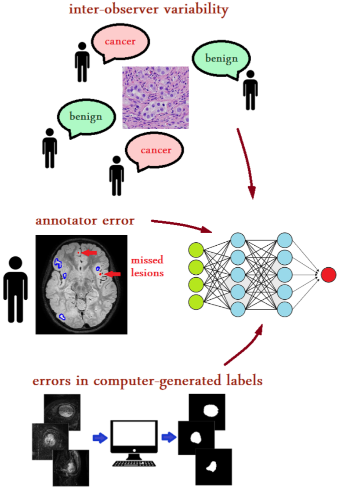


Figure: The primary sources of label noise in medical settings [3]

Robust Learning Methods in Classification

Extensive research exists for mitigating label noise in classification tasks:

- Label Cleaning and Pre-processing.
- Robust Network Architectures.
- Data Re-weighting.
- Curriculum Learning and Knowledge Distillation.
- Noise-robust Loss Functions.

Robust Learning Methods in Classification

Extensive research exists for mitigating label noise in classification tasks:

- Label Cleaning and Pre-processing.
- Robust Network Architectures.
- Data Re-weighting.
- Curriculum Learning and Knowledge Distillation.
- Noise-robust Loss Functions.
  - Relatively easy to implement.
  - Universal solutions.
  - Can be used alongside other methods.

Research Gap

- This critical area remains notably under-investigated for image segmentation.
  - Adapting existing methods to segmentation.
  - Developing new methods tailored for segmentation.
- Many existing methods are not directly suited for the spatial nature of segmentation noise and cannot be easily adapted.
- Developing new methods is complicated and requires significant research time and resources.

Our Contributions:

- We address this research gap by adapting Abstention to segmentation.
- We improve and expand abstention beyond its current definition.

Exploring Abstention

The Mechanism

- Model can choose to *not* make a classification decision on ambiguous data.
- Adds an extra output unit ($k + 1$) representing 'abstain' or 'ignore' class.
- The loss function is modified to reward abstention on uncertain samples.
- Higher abstention = lower loss = smaller contribution to the gradient.

Abstention

The Mechanism

- Model can choose to *not* make a classification decision on ambiguous data.
- Adds an extra output unit ($k + 1$) representing 'abstain' or 'ignore' class.
- The loss function is modified to reward abstention on uncertain samples.
- Higher abstention = lower loss = smaller contribution to the gradient.

The Benefits

- Avoids overfitting on noisy samples.
- Filters data during training with minimal computational overhead.
- No pre-processing required.
- Architecture (and potentially loss function) agnostic.

Deep Abstaining Classifier

$$\mathcal{L}_{DAC}(x_j) = (1 - p_{k+1}) \left(- \sum_{i=1}^{k} t_i \log \frac{p_i}{1 - p_{k+1}} \right) + \alpha \log \frac{1}{1 - p_{k+1}}$$

- Modified CE
- Abstention probability p_{k+1}.
- Regularization term $\left[\alpha \log \frac{1}{1 - p_{k+1}} \right]$.
- Incremental abstention penalty α.
- α is initialized to a small value after a warm-up period.

Deep Abstaining Classifier

$$\mathcal{L}_{DAC}(x_j) = (1 - p_{k+1})\left(-\sum_{i=1}^{k} t_i \log \frac{p_i}{1 - p_{k+1}}\right) + \alpha \log \frac{1}{1 - p_{k+1}}$$

- Modified CE
- Abstention probability p_{k+1}.
- Regularization term $\left[\alpha \log \frac{1}{1-p_{k+1}}\right]$.
- Incremental abstention penalty α.
- α is initialized to a small value after a warm-up period.

Algorithm 1 α auto-tuning

Input: total iter. (T), current iter. (t), total epochs (E), abstention-free epochs (L), current epoch (e), α init factor (ρ), final α (α_{final}), mini-batch cross-entropy over true classes ($\mathcal{H}_c(P^M_{1...K})$)

$\alpha_{set} = False$
for $t := 0$ to T **do**
  if $e < L$ **then**
    $\beta = (1 - P^M_{k+1})\mathcal{H}_c(P^M_{1...K})$
    if $t = 0$ **then**
      $\tilde{\beta} = \beta$\{ // initialize moving average\}
    end if
    $\tilde{\beta} \leftarrow (1-\mu)\tilde{\beta} + \mu\beta$
  end if
  if $e = L$ **and not** α_{set} **then**
    $\alpha := \tilde{\beta}/\rho$\{ // initialize α at start of epoch L\}
    $\delta_\alpha := \frac{\alpha_{final}-\alpha}{E-L}$
    $update_{epoch} = L$
    $\alpha_{set} = True$
  end if
  if $e > update_{epoch}$ **then**
    $\alpha \leftarrow \alpha + \delta_\alpha$\{ //then update α once every epoch\}
    $update_{epoch} = e$
  end if
end for

Figure: DAC's α auto-tuning algorithm [5].

Informed Deep Abstaining Classifier

$$\mathcal{L}_{IDAC}(x_j) = (1 - p_{k+1})\left(-\sum_{i=1}^{k} t_i \log \frac{p_i}{1 - p_{k+1}}\right) + \alpha(\tilde{\eta} - \hat{\eta})^2$$

- Extension of DAC.
- α is fixed during training.
- Uses noise estimation $\tilde{\eta}$ to guide or 'inform' abstention $\hat{\eta}$.
- $\hat{\eta} = \sum_{l=1}^{N} \frac{p_{l,k+1}}{N}$.

Informed Deep Abstaining Classifier

$$\mathcal{L}_{IDAC}(x_j) = (1 - p_{k+1}) \left(- \sum_{i=1}^{k} t_i \log \frac{p_i}{1 - p_{k+1}} \right) + \alpha(\tilde{\eta} - \hat{\eta})^2$$

- Extension of DAC.
- α is fixed during training.
- Uses noise estimation $\tilde{\eta}$ to guide or 'inform' abstention $\hat{\eta}$.
- $\hat{\eta} = \sum_{l=1}^{N} \frac{p_{l,k+1}}{N}$.

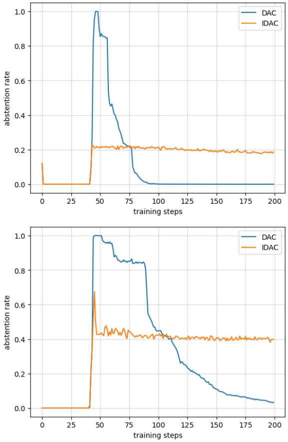


Figure: Abstention behaviour in DAC and IDAC at 10% (top) and 20% (bottom) label noise.

Generalized Cross Entropy (GCE)

$$\mathcal{L}_{GCE}(x_j) = \frac{1 - f(x_j)^q}{q}$$

Potential Baselines

Generalized Cross Entropy (GCE)

$$\mathcal{L}_{GCE}(x_j) = \frac{1 - f(x_j)^q}{q}$$

Symmetric Cross Entropy (SCE)

$$\mathcal{L}_{RCE}(x_j) = -\sum_{i=1}^{k} p_i \log(t_i)$$

$$\mathcal{L}_{SCE}(x_j) = \alpha \mathcal{L}_{CE}(x_j) + \beta \mathcal{L}_{RCE}(x_j)$$

Potential Baselines

Generalized Cross Entropy (GCE)

$$\mathcal{L}_{GCE}(x_j) = \frac{1 - f(x_j)^q}{q}$$

Symmetric Cross Entropy (SCE)

$$\mathcal{L}_{RCE}(x_j) = -\sum_{i=1}^{k} p_i \log(t_i)$$

$$\mathcal{L}_{SCE}(x_j) = \alpha\mathcal{L}_{CE}(x_j) + \beta\mathcal{L}_{RCE}(x_j)$$

Dice Loss (Dice Similarity Coefficient)

$$\mathcal{DSC}(x_j) = \frac{2\sum_{i=1}^{N} p_i t_i}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} t_i}$$

$$\mathcal{L}_{Dice}(x_j) = 1 - \mathcal{DSC}(x_j)$$

Universal Abstention Framework

Universal Abstention Framework

$$\mathcal{L}_{abstention}(x_j) = (1 - p_{k+1})\mathcal{L}_{\mathcal{X}}(x_j) + \alpha \left| \log \frac{1 - \tilde{\eta}}{1 - p_{k+1}} \right| \tag{1}$$

$$\mathcal{L}_{abstention}(x_j) = (1 - p_{k+1})\mathcal{L}_{\mathcal{X}}(x_j) + \alpha \left| \log \frac{1 - \tilde{\eta}}{1 - p_{k+1}} \right| \quad (1)$$

Informed Regularization

- Combines DAC and IDAC.
- Allows for more freedom to abstain when noise level is high.
- Reduces overfitting in the final stages of training.
- Defaults back to DAC if $\tilde{\eta}$ is unknown.

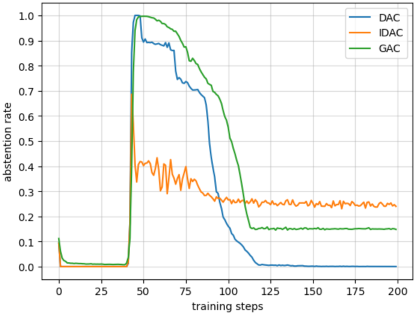


Figure: Abstention behaviour in DAC, IDAC, and GAC at 15% noise.

Power-law auto-tuning

$$\alpha = \alpha_{final} * \left(\frac{e - L}{E - L}\right)^{\gamma} \qquad (2)$$

current epoch e, total epochs E, warm-up epochs L.

- Replaces DAC's complicated auto-tuning algorithm with a simpler and more flexible calculation.
- γ controls the rate of growth for α.
- Higher $\gamma \rightarrow$ smaller $\alpha \rightarrow$ more abstention.
- Still allows for DAC's linear growth ($\gamma = 1$).

Power-law auto-tuning

$$\alpha = \alpha_{final} * \left(\frac{e - L}{E - L}\right)^{\gamma} \qquad (2)$$

current epoch e, total epochs E, warm-up epochs L.

- Replaces DAC's complicated auto-tuning algorithm with a simpler and more flexible calculation.
- γ controls the rate of growth for α.
- Higher $\gamma \rightarrow$ smaller $\alpha \rightarrow$ more abstention.
- Still allows for DAC's linear growth ($\gamma = 1$).

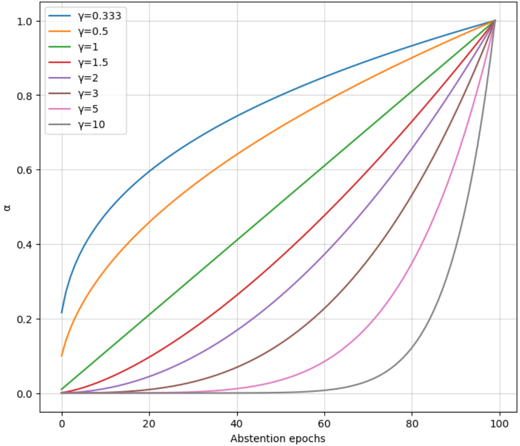


Figure: The effect of different values of γ on the growth of α with $\alpha_{final} = 1$.

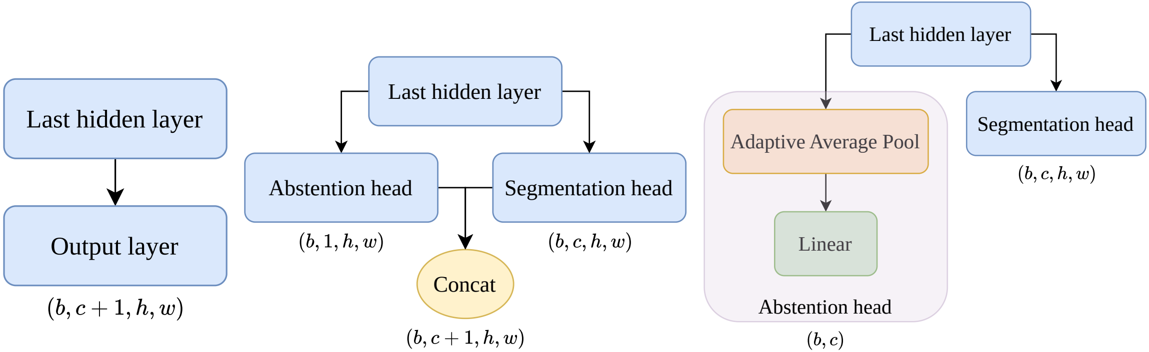
Novel Abstaining Loss Functions

- **Generalized Abstaining Classifier (GAC):** GCE + Abstention
- **Symmetric Abstaining Classifier (SAC):** SCE + Abstention
- **Abstaining Dice Segmenter (ADS):** Dice + Abstention

Novel Abstaining Loss Functions

- **Generalized Abstaining Classifier (GAC):** GCE + Abstention
- **Symmetric Abstaining Classifier (SAC):** SCE + Abstention
- **Abstaining Dice Segmenter (ADS):** Dice + Abstention
  - needs to adapt Abstention to Dice's class-wise nature.

ADS Class-wise Abstention



(a) Output layer with abstention channel

(b) Output layer with split abstention channel

(c) ADS output layer with class-wise abstention

Figure: Transforming the output layer of an abstaining model from pixel-wise to class-wise abstention.

Experiments

Datasets: CaDIS

- 4,670 high-quality annotated images from cataract surgery.
- Dense annotations.
- Has 3 variants for number of classes.
- We used the first variant (8 classes).
- Normalized and resized to 480x256.

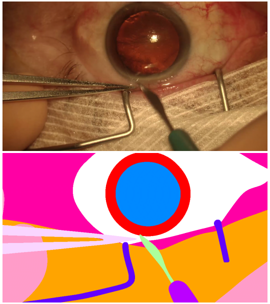


Figure: Example image frame (top) and semantic segmentation labels (bottom) from the CaDIS Dataset [2].

- 13,195 laparoscopic annotated images.
- Binary segmentations for 11 anatomical structures.
- 1,430 stomach images used for multi-organ segmentation (7 organs).
- Sparse annotations ($\approx$ 82% background).
- Normalized and resized to 480x384.

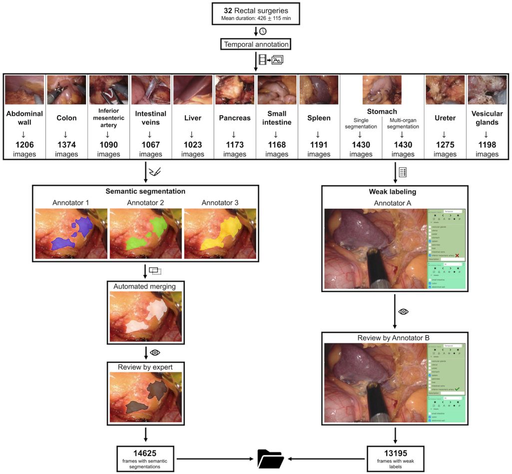


Figure: Overview of the data acquisition and validation process of DSAD [1].

Noise Synthesis

- Morphological operations: Erosion and Dilation.
- Random label flipping.
- 5 noise level for each dataset.
- CaDIS: 5-25%.
- DSAD: 3-15%.

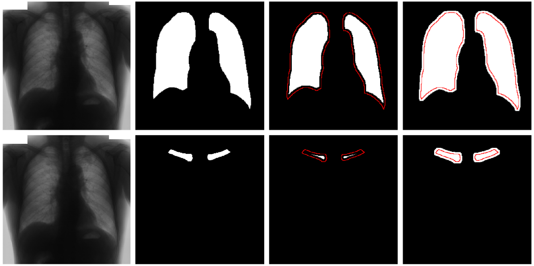


Figure: Two examples of Erosion and Dilation. Correct segmentation boundaries in red [6].

- Most commonly used segmentation architecture.
- Designed for medical image segmentation.
- encoder captures context and decoder enables precise localization.
- Skip connections bridge the two paths.
- Used with pretrained ResNet-50 backbone.

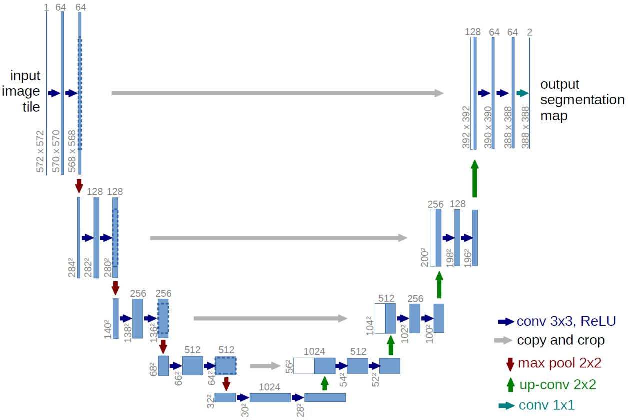


Figure: The U-Net architecture [4].

Experimental Setup

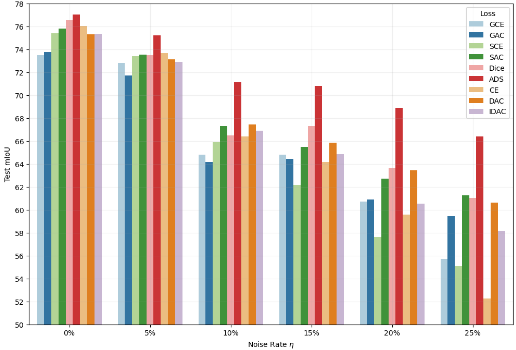
- Optimized hyperparameters with U-Net for highest noise level for each dataset.
- Trained for 50 epochs.
- AdamW with lr=0.003.
- lr divided by 5 every 10 epochs.
- A single NVIDIA A100 80GB GPU.

Loss	CaDIS	DSAD
DAC	$\alpha_{final} = 1$ $L = 10$	$\alpha_{final} = 2$ $L = 18$
IDAC	$\alpha = 1$ $L = 10$	$\alpha = 1$ $L = 10$
GCE	q=0.5	q=0.1
GAC	$\alpha_{final} = 3$ $L = 10$ $\gamma = 3$	$\alpha_{final} = 2$ $L = 15$ $\gamma = 2$
SCE	$\alpha = 1$ $\beta = 1$	$\alpha = 0.5$ $\beta = 1$
SAC	$\alpha_{final} = 1$ $L = 10$ $\gamma = 1.5$	$\alpha_{final} = 1$ $L = 20$ $\gamma = 3$
ADS	$\alpha_{final} = 1$ $L = 10$ $\gamma = 3$ $w = 16$	$\alpha_{final} = 4$ $L = 10$ $\gamma = 1.5$ $w = 16$

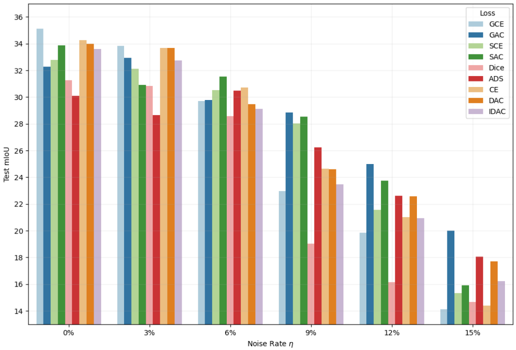
Table: The hyperparameters used in our experiments.

Evaluations

Results



(a) CaDIS



(b) DSAD

Figure: Test mIoU (%) scores of a U-Net model trained on CaDIS (a) and DSAD (b) at 5 different noise levels.

U-Net Performance

Dataset	Noise rate η (%)	Loss function								
		CE	DAC	IDAC	GCE	GAC	SCE	SAC	Dice	ADS
CaDIS	0	**76.02**±**0.70**	75.29±0.79	75.36±0.73	73.49±3.27	**73.76**±**2.80**	75.38±0.75	**75.83**±**0.62**	76.52±0.47	**77.04**±**0.37**
	5	**73.67**±**1.03**	73.14±0.46	72.89±0.41	**72.83**±**1.11**	71.73±2.79	73.41±0.71	**73.51**±**1.59**	73.48±0.28	**75.22**±**0.85**
	10	66.39±0.17	**67.43**±**0.49**	66.92±0.49	**64.82**±**0.86**	64.16±2.57	65.92±0.91	**67.29**±**1.65**	66.51±0.61	**71.12**±**0.55**
	15	64.15±2.47	**65.85**±**1.05**	64.87±0.91	**64.81**±**0.46**	64.44±2.70	62.16±1.99	**65.48**±**2.11**	67.31±0.73	**70.80**±**1.08**
	20	59.56±1.21	**63.42**±**0.87**	60.54±2.27	60.73±1.41	**60.91**±**1.64**	57.62±4.22	**62.70**±**0.31**	63.64±0.82	**68.88**±**0.49**
	25	52.27±1.70	**60.63**±**2.73**	58.19±4.77	55.71±1.30	**59.46**±**0.76**	55.08±0.93	**61.27**±**1.22**	61.04±1.41	**66.39**±**0.67**
DSAD	0	**34.25**±**2.50**	34.01±0.96	33.60±0.72	**35.14**±**1.65**	32.26±0.53	32.78±1.19	**33.86**±**1.83**	**31.28**±**0.87**	30.09±1.10
	3	**33.69**±**1.85**	33.67±2.01	32.76±2.03	**33.84**±**2.56**	32.94±2.23	**32.11**±**1.09**	30.90±2.76	**30.83**±**4.78**	28.64±2.76
	6	**30.70**±**2.47**	29.47±1.97	29.11±2.10	29.69±1.96	**29.78**±**4.27**	30.51±2.16	**31.55**±**2.43**	28.56±1.00	**30.48**±**3.61**
	9	**24.65**±**2.90**	24.58±2.61	23.47±2.48	22.95±2.93	**28.84**±**4.17**	28.02±2.37	**28.55**±**1.29**	19.04±1.92	**26.23**±**2.05**
	12	21.00±3.15	**22.59**±**4.35**	20.94±1.86	19.84±2.89	**25.00**±**4.13**	21.57±0.67	**23.73**±**0.68**	16.15±1.49	**22.63**±**0.51**
	15	14.41±2.59	**17.69**±**3.97**	16.24±1.45	14.12±2.91	**20.01**±**2.56**	15.31±0.75	**15.91**±**3.53**	14.65±1.50	**18.05**±**1.63**

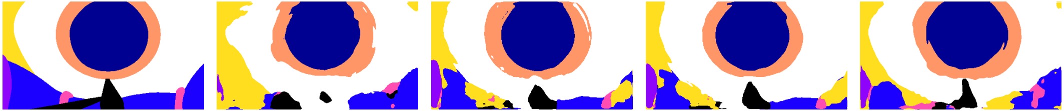
Table: Average test mIoU (%) and standard deviation (5 runs) of a U-Net model trained on CaDIS and DSAD datasets with various rate of label noise, comparing five abstaining loss functions [DAC, IDAC, GAC, SAC, ADS] against their non-abstaining baselines [CE, GCE, SCE, Dice]. Best results in each bracket are in **bold**.

DeepLabV3+

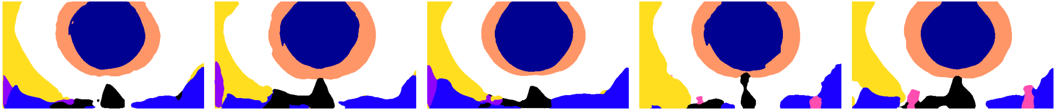
Dataset	Loss function								
	CE	DAC	IDAC	GCE	GAC	SCE	SAC	Dice	ADS
CaDIS	56.02±1.30	**57.02±0.81**	56.29±1.05	55.56±2.08	**58.08±1.43**	58.37±0.53	**59.77±1.17**	59.55±1.66	**61.84±2.23**
DSAD	**16.73±2.34**	15.90±3.19	16.20±1.37	16.26±1.37	**19.01±1.69**	12.74±2.03	**14.03±3.53**	12.46±0.86	**17.16±2.02**

Table: Average test mIoU (%) and standard deviation (5 runs) of a DeepLabV3+ model trained on CaDIS and DSAD datasets at 25% and 15% label noise, respectively. Best results in each bracket are in **bold**.

Visualizations

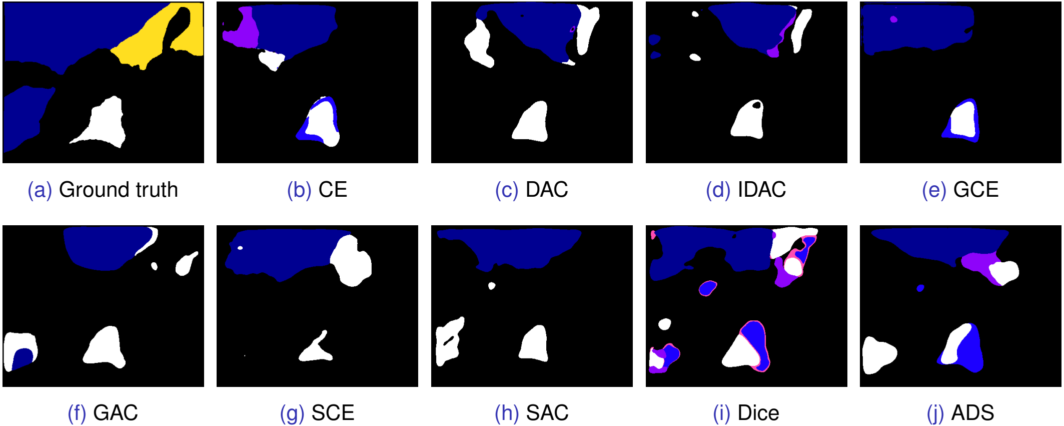


(a) Ground truth (b) CE (c) DAC (d) IDAC (e) GCE



(f) GAC (g) SCE (h) SAC (i) Dice (j) ADS

Figure: Visualisation of a sample clean ground truth from CaDIS and the segmentation predictions of a U-Net model trained with each loss function at 25% noise.



(a) Ground truth (b) CE (c) DAC (d) IDAC (e) GCE

(f) GAC (g) SCE (h) SAC (i) Dice (j) ADS

Figure: Visualisation of a sample clean ground truth from DSAD and the segmentation predictions of a U-Net model trained with each loss function at 15% noise.

Conclusions

Contributions & Impact

- Adapted abstention for Medical Image Segmentation.
- Enhanced abstention with Informed regularization and flexible α-tuning.
- Integration with different and distinct loss functions.
- Empirical proof: Abstention boosts robustness across losses, datasets, and architectures.
- Abstention is a modular and easy-to-use extension for robust learning in medical imaging.

Future Work

- **Dynamic Noise Estimation:** Develop methods to learn the noise rate directly from the data.
- **Real-World Noise Validation:** Test on clinical datasets with naturally occurring, unsimulated noise.
- **Abstention as an Uncertainty Metric:** Use the model's abstention signal to flag difficult cases for expert review, creating a human-in-the-loop system.

Thank You

References I

[1] Matthias Carstens et al. "The Dresden Surgical Anatomy Dataset for Abdominal Organ Segmentation in Surgical Data Science". In: *Sci Data* 10.1 (Jan. 2023), p. 3. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01719-2.

[2] Maria Grammatikopoulou et al. *CaDIS: Cataract Dataset for Image Segmentation*. Feb. 2022. DOI: 10.48550/arXiv.1906.11586. arXiv: 1906.11586 [cs].

[3] Davood Karimi et al. "Deep Learning with Noisy Labels: Exploring Techniques and Remedies in Medical Image Analysis". In: *Medical Image Analysis* 65 (Oct. 2020), p. 101759. ISSN: 13618415. DOI: 10.1016/j.media.2020.101759.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. May 2015. DOI: 10.48550/arXiv.1505.04597. arXiv: 1505.04597 [cs].

References II

[5] Sunil Thulasidasan et al. *Combating Label Noise in Deep Learning Using Abstention*. Aug. 2019. DOI: 10.48550/arXiv.1905.10964. arXiv: 1905.10964 [stat].

[6] Haidong Zhu, Jialin Shi, and Ji Wu. *Pick-and-Learn: Automatic Quality Evaluation for Noisy-Labeled Image Segmentation*. July 2019. DOI: 10.48550/arXiv.1907.11835. arXiv: 1907.11835 [cs].