Imperial College London

BioMedIA

Learning Strategies for Improving Neural Networks for Image Segmentation under Class Imbalance

Zeju Li
BioMedIA Group, Department of Computing, Imperial College London

August 30th 2022

Overview

Introduction

- Short Bio

PhD Research on Class Imbalance in Segmentation

- Overfitting under Class Imbalance
- Underfitting under Class Imbalance
- Automatic Data Augmentation
- Class Imbalance under Domain Shifts

Short Bio

Education

- 2018.10 – 2022.9
  - PhD in Computing, Imperial College London, London
  - Supervisor: Dr. Ben Glocker
- 2015.9 – 2018.7
  - Master in Biomedical Engineering, Fudan, Shanghai
- 2011.9 – 2015.7
  - Bachelor in Electronic Engineering, Fudan, Shanghai

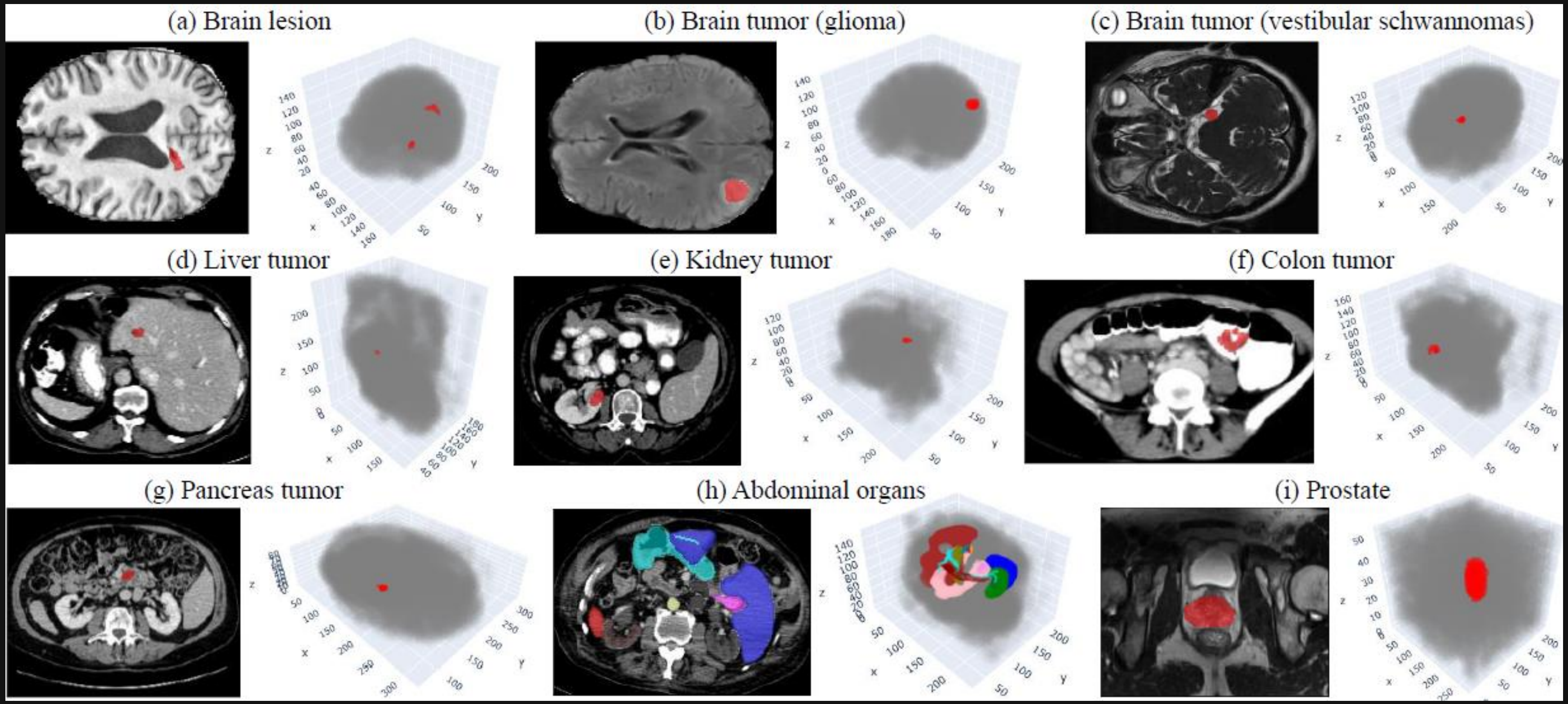
Intern

- 2019.7 – 2020.4
  - Huawei Noah's Ark Lab, London
- 2018.7 – 2018.9
  - MIRACLE, ICT, Beijing

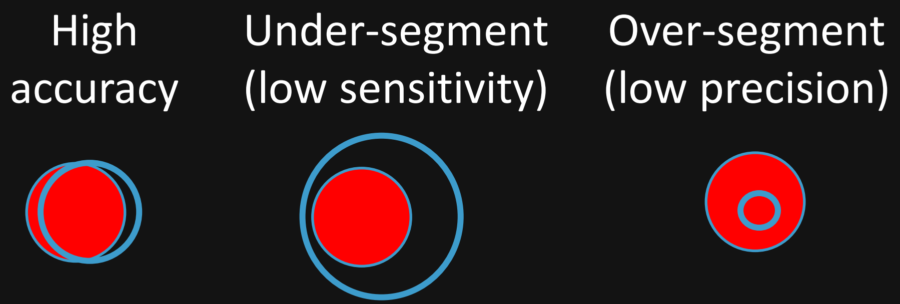
Motivations

Image Segmentation

- In medical image segmentation, **class imbalance** is not uncommon as tumor and organs are relatively small in medical imaging.

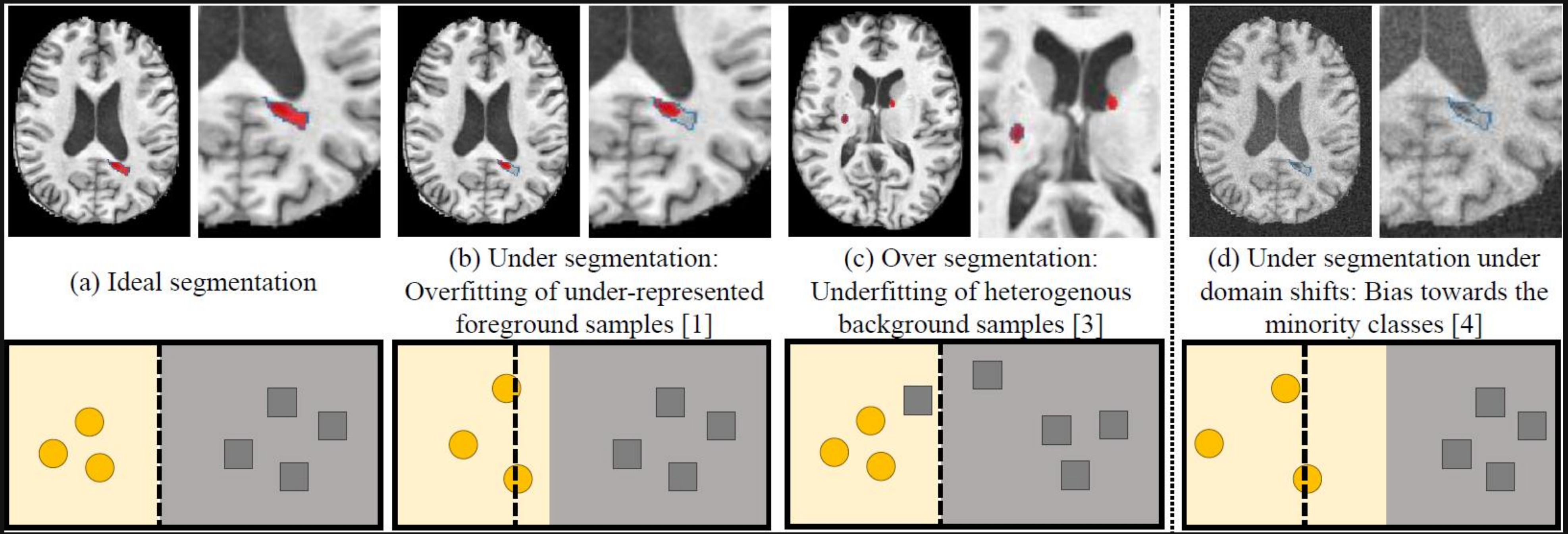


Motivations



Observations

- Class imbalance causes **under- and over-segmentation**.
- Understanding the effects of class imbalance in segmentation.

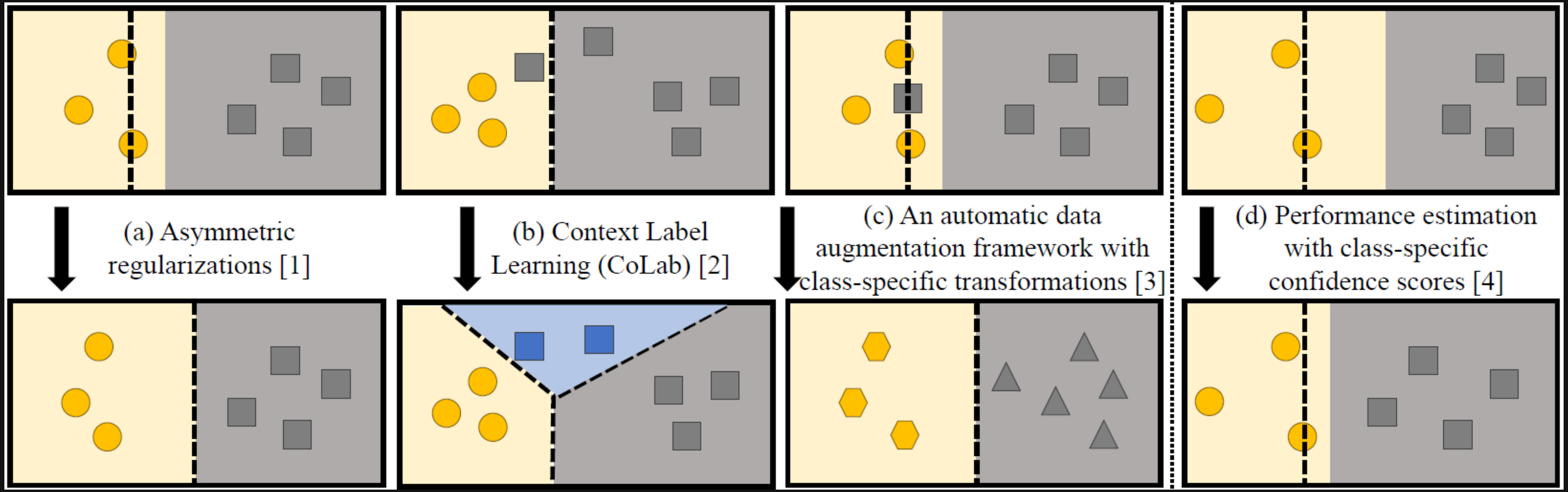


[1] **Z. Li**, *et al.* MICCAI 2019, TMI 2020 [2] **Z. Li**, *et al.* TMI Under Revision [3] **Z. Li**, *et al.* TMI Under Revision [4] **Z. Li**, *et al.* MICCAI 2022

Motivations

A glance of methodological contributions

- Improving neural networks for **segmentation under class imbalance**.

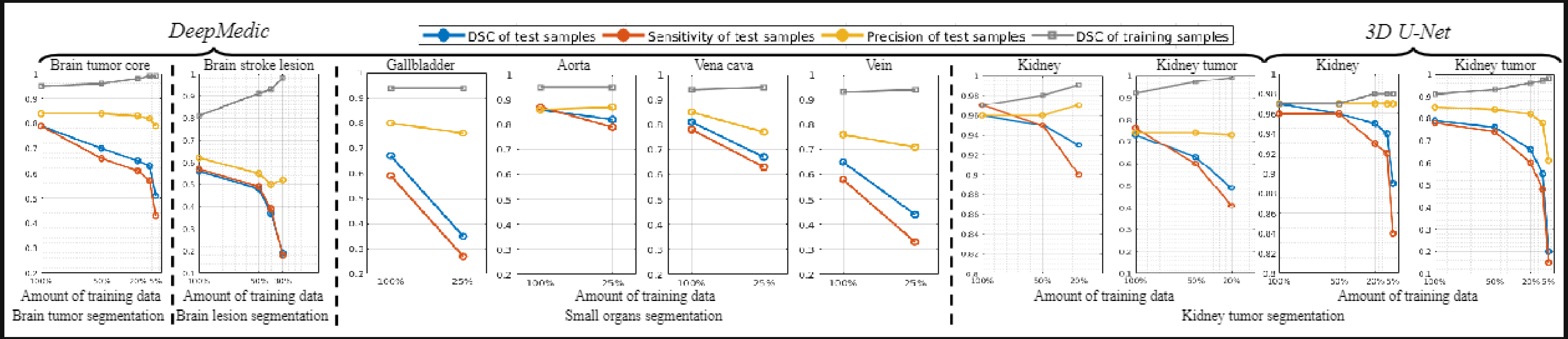


[1] **Z. Li**, *et al.* MICCAI 2019, TMI 2020 [2] **Z. Li**, *et al.* TMI Under Revision [3] **Z. Li**, *et al.* TMI Under Revision [4] **Z. Li**, *et al.* MICCAI 2022

Overfitting under Class Imbalance

Analysis

➢ With less training data, performances decline due to the drastic **reduction of sensitivity**, while precision is retained.



[1] **Z. Li**, *et al.* MICCAI 2019, TMI 2020

Overfitting under Class Imbalance

Analysis

- CNN maps training and testing samples of the background class to similar logit values.
- However, **mean activation for testing data shifts** significantly for the foreground class towards and sometimes across the decision boundary.

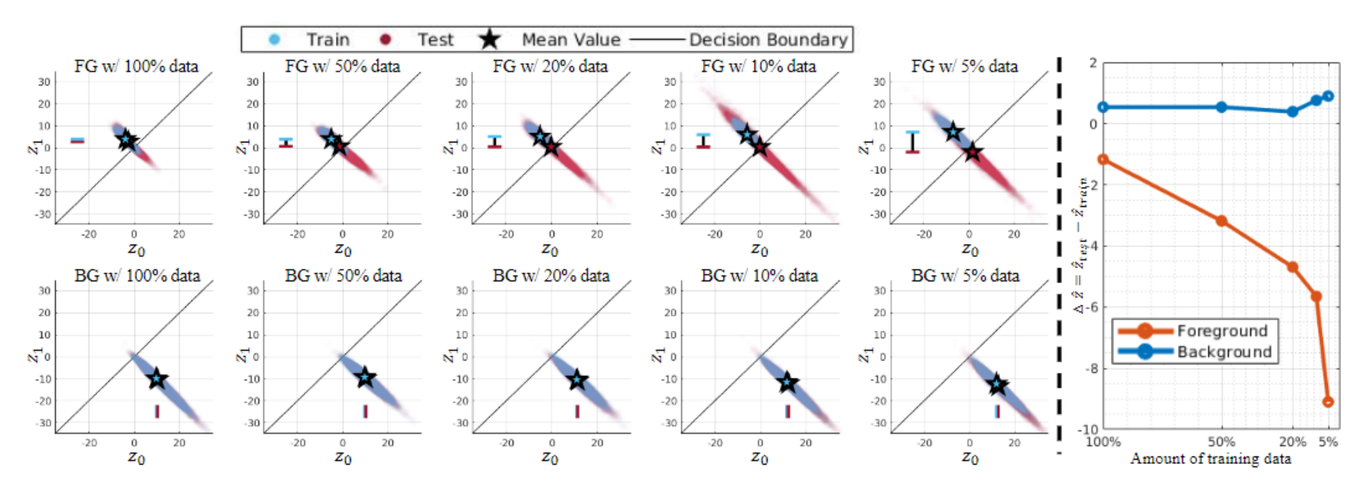
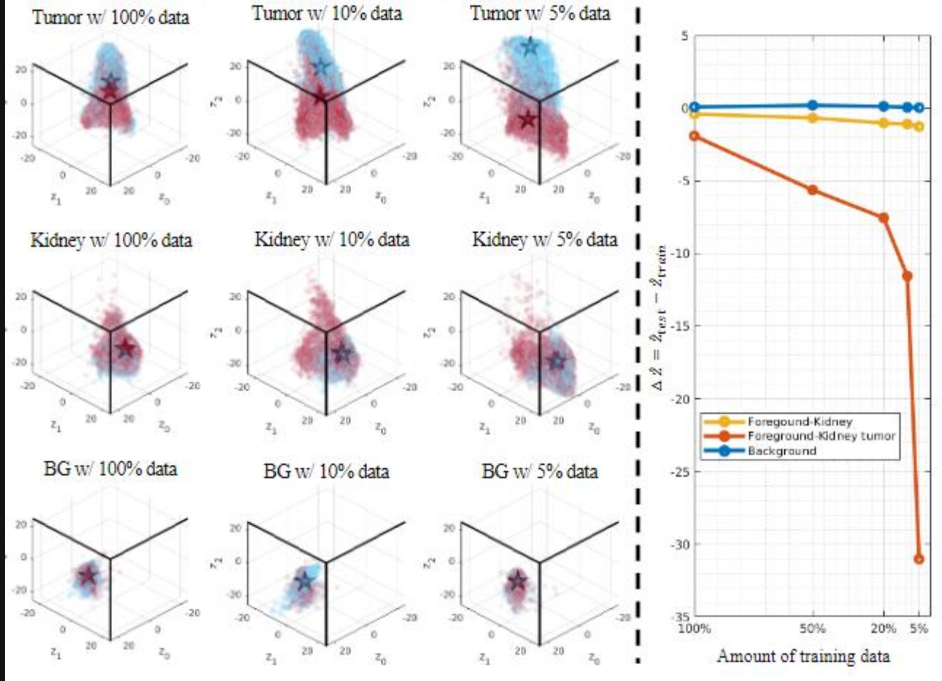


Fig. 3. (Left part) Activations of the classification layer (logit z_0 for background, logit z_1 for brain tumor core) when processing (top) tumor and (bottom) background samples of BRATS with DeepMedic, using different amounts of training data. The CNN maps training and testing samples of the background class to similar logit values. However, mean activation for testing data shifts significantly for the tumor class towards and sometimes across the decision boundary. (Right part) The shift of mean value of logits observed when processing training and testing data ($\Delta\hat{z} = \hat{z}_{test} - \hat{z}_{train}$).



[1] **Z. Li**, *et al.* MICCAI 2019, TMI 2020

Overfitting under Class Imbalance

Method

- We make the logit activations of foreground class far away from the decision boundary by **setting bias for the foreground class** in different ways.

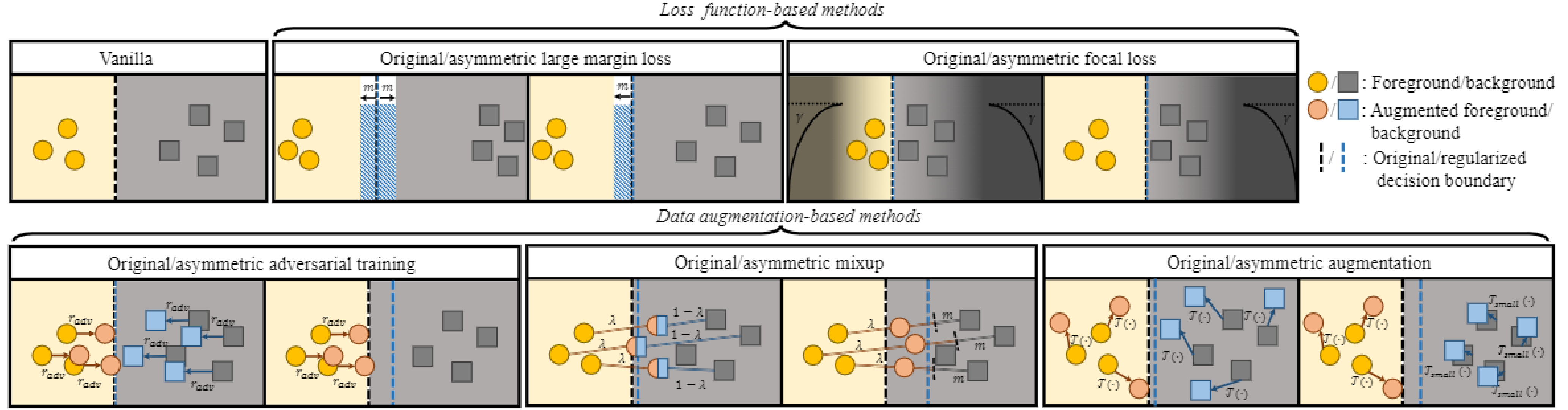


Fig. 6. The illustration of the proposed asymmetric modifications for the existing loss functions and regularization techniques. We make the logit activations of foreground class far away from the decision boundary by setting a bias for the foreground class in different ways.

[1] **Z. Li**, *et al.* MICCAI 2019, TMI 2020

Overfitting under Class Imbalance

Results

➢ The proposed variants of regularization and techniques can **reduce overfitting** and improve performance.

TABLE II

EVALUATION OF BRAIN TUMOR CORE SEGMENTATION USING DEEPMEDIC WITH DIFFERENT AMOUNTS OF TRAINING DATA AND DIFFERENT TECHNIQUES TO COUNTER OVERFITTING. THE RESULTS ARE CALCULATED WITH POST-PROCESSING. RESULTS WHICH HAVE WORSE DSC THAN THE VANILLA BASELINE ARE HIGHLIGHTED WITH SHADING. THE BEST AND SECOND BEST RESULTS ARE IN **BOLD** WITH THE BEST ALSO **UNDERLINED**.

Method	5% training				10% training				20% training				50% training			
	DSC	SEN	PRC	HD	DSC	SEN	PRC	HD	DSC	SEN	PRC	HD	DSC	SEN	PRC	HD
Vanilla - CE [20]	50.4	41.0	83.5	18.0	62.5	56.0	83.1	14.3	64.9	59.8	85.7	13.8	69.4	65.4	85.3	15.7
Vanilla - CE - 80% tumor	45.5	36.0	86.7	17.8	61.5	54.2	81.7	18.5	65.3	59.6	85.0	15.1	68.6	64.1	86.1	14.8
Vanilla - F1 (DSC)	47.2	37.4	86.6	15.9	58.9	51.1	83.6	20.1	64.3	58.1	83.5	16.3	67.1	62.5	86.5	15.3
Vanilla - F2 [14]	45.8	36.9	81.9	17.9	59.3	52.2	84.9	18.0	66.4	61.1	83.4	14.1	68.8	66.0	83.4	13.7
Vanilla - F4 [14]	51.6	42.5	83.8	18.1	59.6	53.0	82.9	18.4	65.9	61.9	85.4	14.2	67.5	64.5	84.9	13.7
Vanilla - F8 [14]	47.4	38.7	83.1	19.6	59.8	52.4	87.0	15.4	64.5	60.3	85.2	14.7	67.9	65.4	81.6	14.9
Large margin loss [31]	44.5	35.9	82.8	20.2	60.9	53.5	84.0	17.6	67.0	61.6	86.1	14.4	66.5	62.2	88.1	13.7
Asymmetric large margin loss	56.8	48.9	83.4	**15.0**	64.0	56.8	87.0	13.9	67.4	62.9	84.1	15.9	68.9	64.9	86.5	14.1
Focal loss [29]	54.0	44.8	82.6	16.0	62.6	55.1	84.3	17.7	64.9	60.0	84.4	19.5	67.0	62.3	87.0	16.5
Asymmetric focal loss	58.8	51.4	81.6	**15.0**	66.8	62.0	83.2	**13.2**	68.9	66.2	83.3	**12.5**	**71.5**	70.6	83.7	12.1
Adversarial training [12]	53.2	44.6	85.0	19.2	62.0	55.0	84.8	20.6	64.6	59.4	84.6	17.3	65.6	61.2	86.0	19.4
Asymmetric adversarial training	58.5	50.8	80.1	16.2	63.9	58.2	83.1	17.2	67.7	63.7	84.2	17.0	70.5	68.4	83.0	14.8
Mixup [47]	49.7	40.9	83.0	19.6	60.3	53.9	83.1	21.2	63.9	58.5	84.1	18.2	66.4	61.5	86.8	19.0
Asymmetric mixup	**59.8**	56.8	74.7	17.7	**68.5**	65.1	80.7	15.3	**70.8**	67.9	85.1	**11.6**	70.7	67.9	85.4	**11.8**
Symmetric combination	50.0	42.0	84.6	21.1	60.3	53.1	84.7	25.1	64.1	58.3	86.6	19.1	67.2	63.1	86.6	15.1
Asymmetric combination	**63.4**	63.1	75.9	**15.1**	**72.4**	72.9	78.3	**10.8**	**71.6**	72.0	80.1	13.7	**74.1**	76.0	82.4	**10.7**

[1] **Z. Li**, *et al.* MICCAI 2019, TMI 2020

Overfitting under Class Imbalance

Results

➢ Asymmetric modifications lead to **better separation of the logits** of unseen foreground samples.

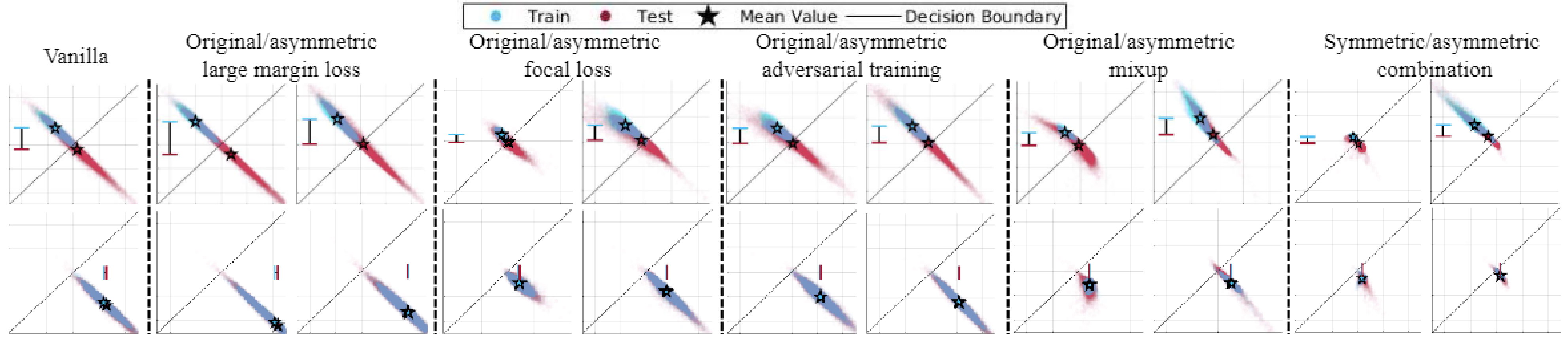


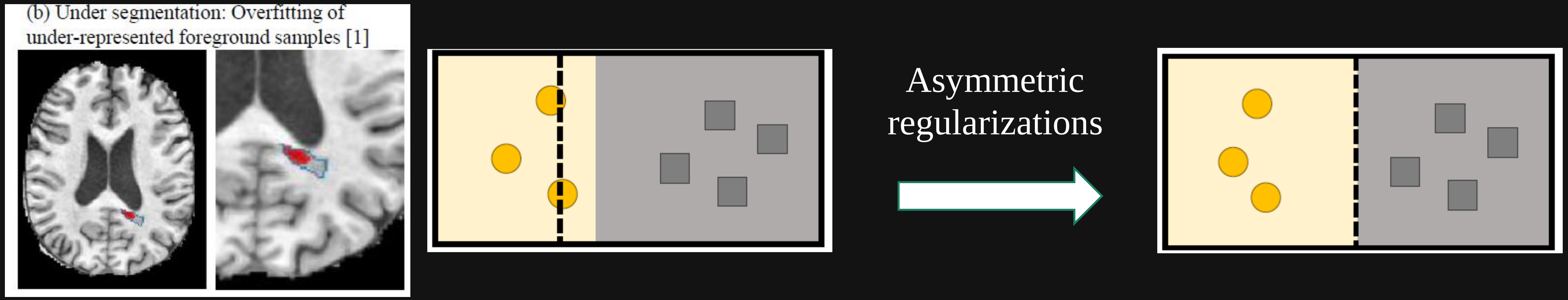
Fig. 7. Activations of the classification layer when processing tumor (top) and background (bottom) samples of BRATS with DeepMedic, using 5% training data. Asymmetric modifications lead to better separation of the logits of unseen tumor samples.

[1] **Z. Li**, *et al.* MICCAI 2019, TMI 2020

Overfitting under Class Imbalance

Conclusion

- Overfitting under class imbalance leads to **loss of sensitivity**.
- The distribution of logit activations when processing unseen test samples of an **under-represented class tends to shift** towards and even across the decision boundary.
- We propose several asymmetric techniques based on our observations of logit distribution.

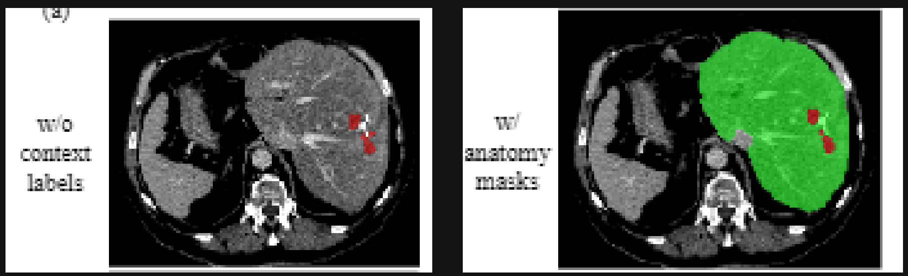


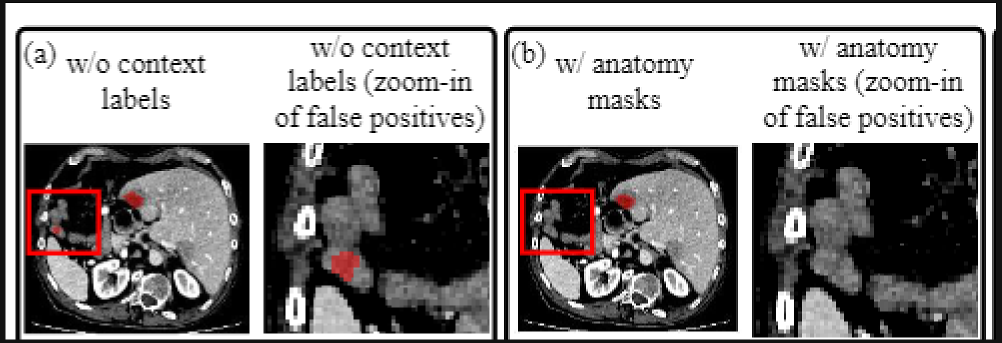
[1] **Z. Li**, *et al.* MICCAI 2019, TMI 2020

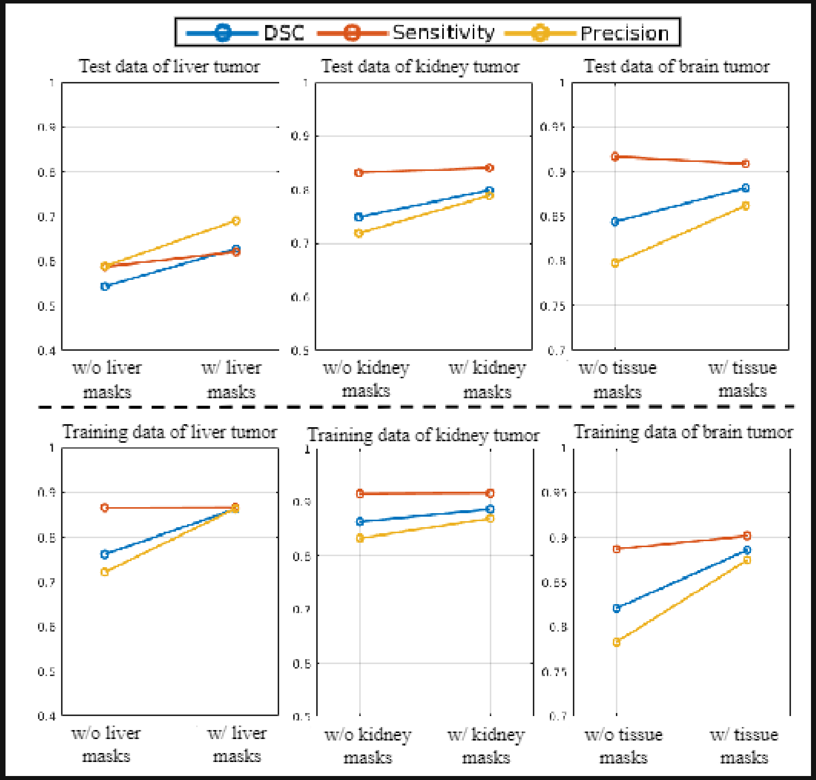
Underfitting under Class Imbalance

Analysis

- With heterogeneous background, performances decline due to the drastic **reduction of precision**, while sensitivity is retained.





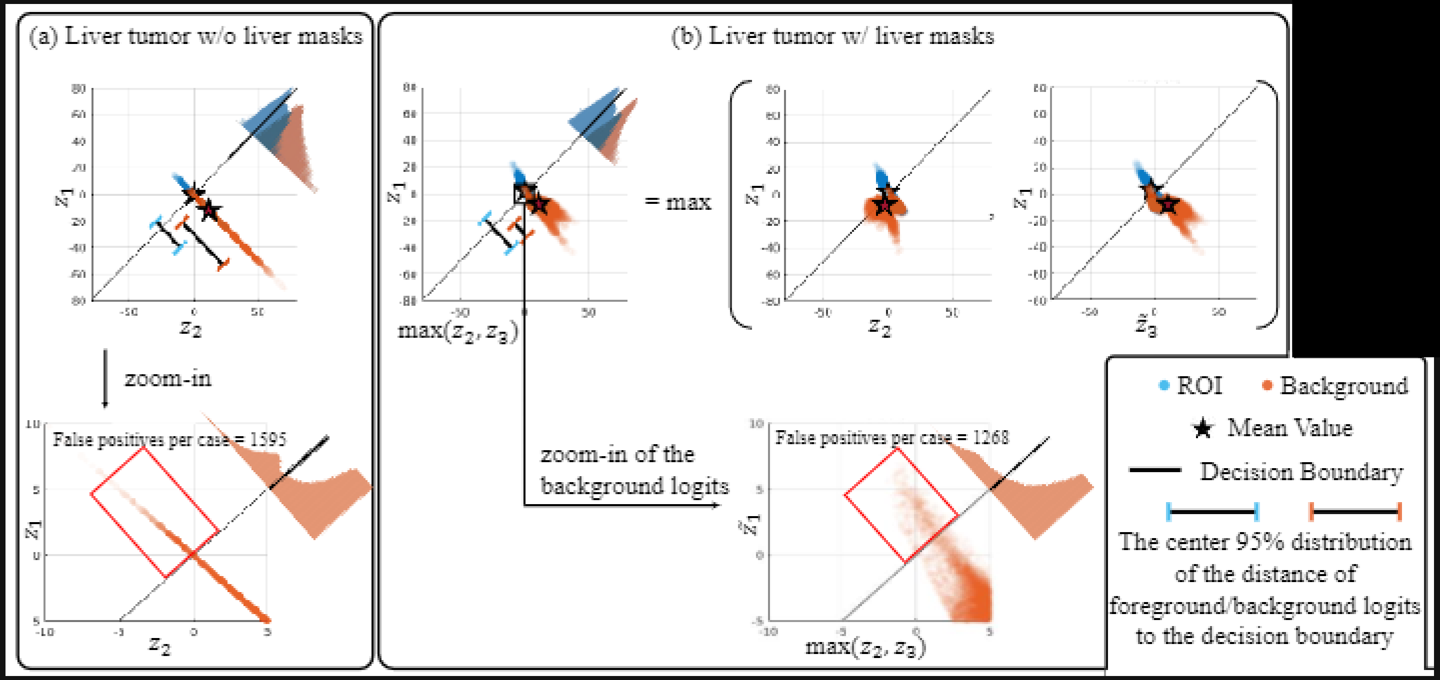


[2] **Z. Li**, *et al.* TMI Under Revision

Underfitting under Class Imbalance

Analysis

- Neural networks could not map the heterogeneous background samples to **compact clusters** in feature space.
- As a result, the logit activations of background would approach and even move across the decision boundary.

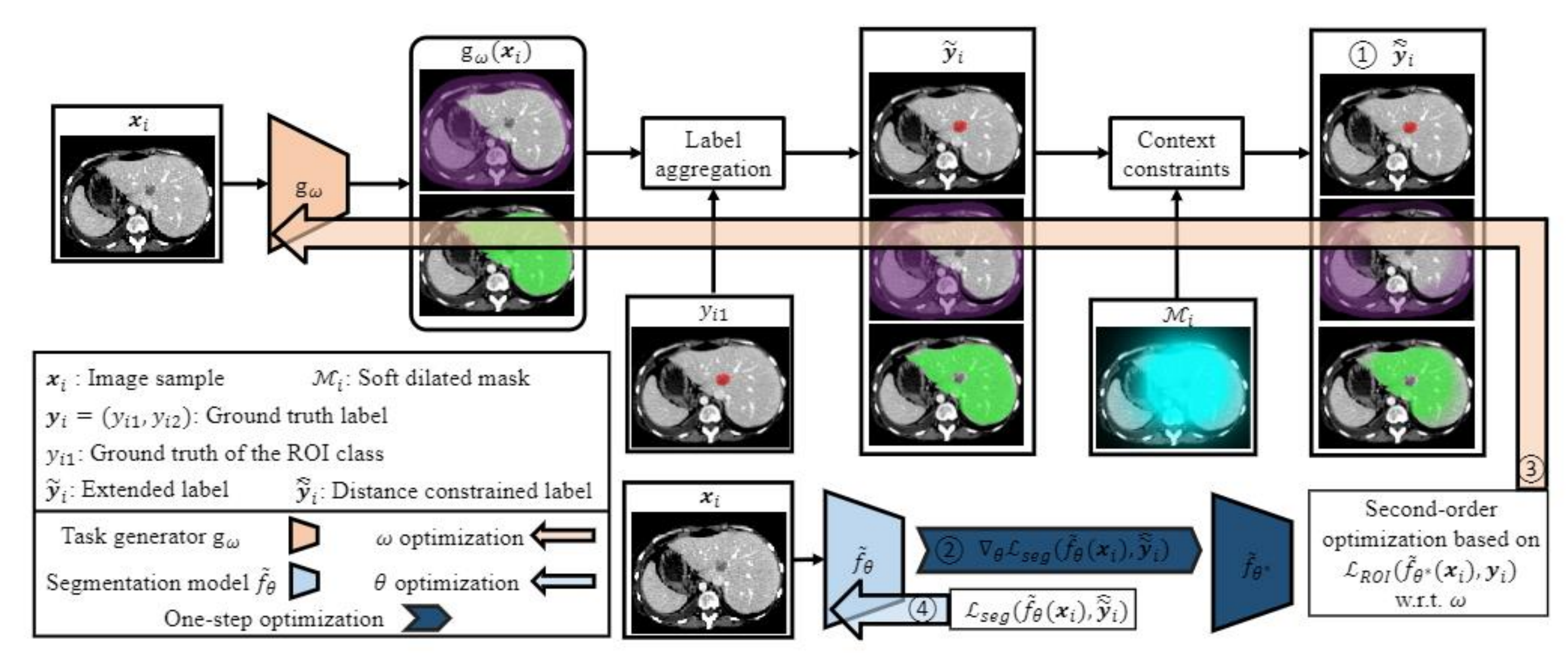


[2] **Z. Li**, *et al.* TMI Under Revision

Underfitting under Class Imbalance

Method

- **Context label learning (CoLab)**
- We train an auxiliary network as a task generator, along with the primary segmentation model, to automatically generate context labels that positively affect the ROI segmentation accuracy.

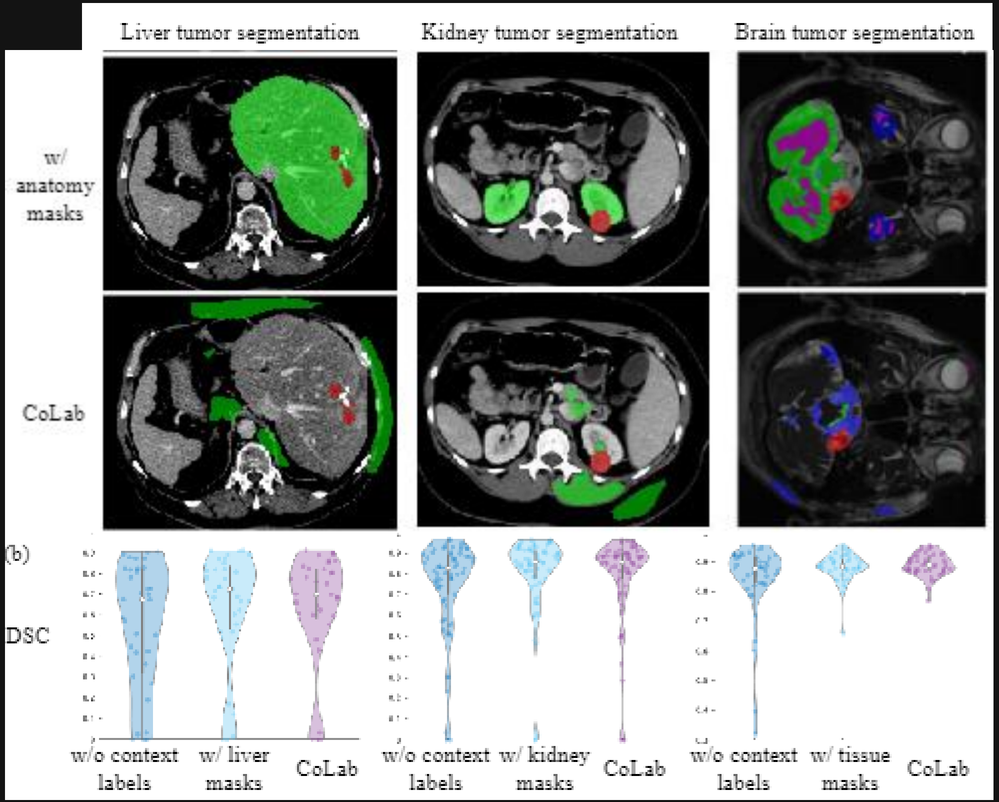


[2] **Z. Li**, *et al.* TMI Under Revision

Underfitting under Class Imbalance

Results

➢ Similar and sometimes better effect in improving segmentation accuracy when compared with human-defined context labels.



Task	Method	t	DSC	SEN	PRC	HD
Liver tumor [5]	w/o liver masks	1	54.4	58.8	58.9	111.1
	K-means [1]	2	**61.4**	**61.4**	67.0	71.9
	Dilated masks [26]	2	60.7	59.8	**68.0**	67.6
	CoLab	2	**62.5**	**62.8**	**67.3**	69.4
	CoLab	4	57.3	60.5	62.3	**56.3**
	CoLab	6	59.7	60.3	65.2	**43.6**
	w/ model-predicted liver masks [16]	2	62.4	61.6	70.6	44.1
	w/ liver masks [5]	2	62.8	62.1	69.1	53.5
Kidney tumor [12]	w/o kidney masks	1	74.9	83.2	71.9	120.4
	K-means [1]	2	**76.8**	**83.5**	74.3	87.1
	Dilated masks [26]	2	76.4	**83.9**	73.1	95.3
	CoLab	2	**78.5**	82.2	**77.7**	**75.7**
	CoLab	4	76.4	80.6	**76.5**	**63.7**
	CoLab	6	74.9	81.0	73.3	79.4
	w/ model-predicted kidney masks [16]	2	79.2	81.3	82.7	38.1
	w/ kidney masks [12]	2	79.9	84.1	78.9	54.7

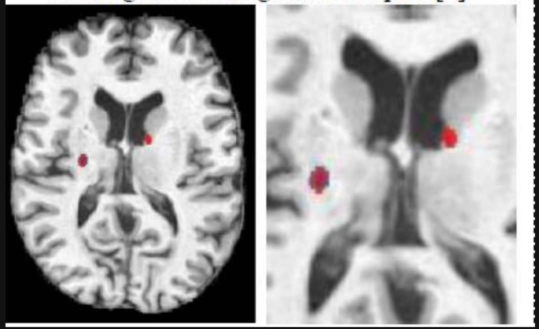
[2] **Z. Li**, *et al.* TMI Under Revision

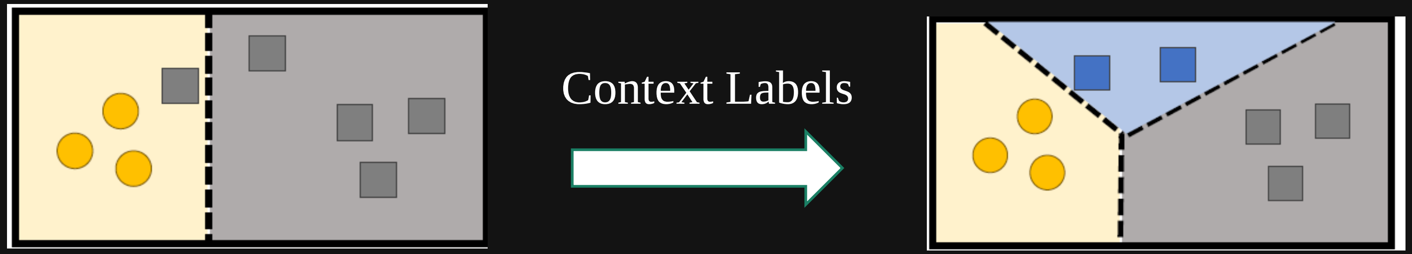
Underfitting under Class Imbalance

Conclusion

- Overfitting under class imbalance leads to **loss of precision**.
- The distribution over background logit activations may shift across the decision boundary, leading to systematic over-segmentation.
- Context labels improve the context representations by **decomposing the background class** into several subclasses.

(c) Over segmentation: Underfitting of heterogenous background samples [3]



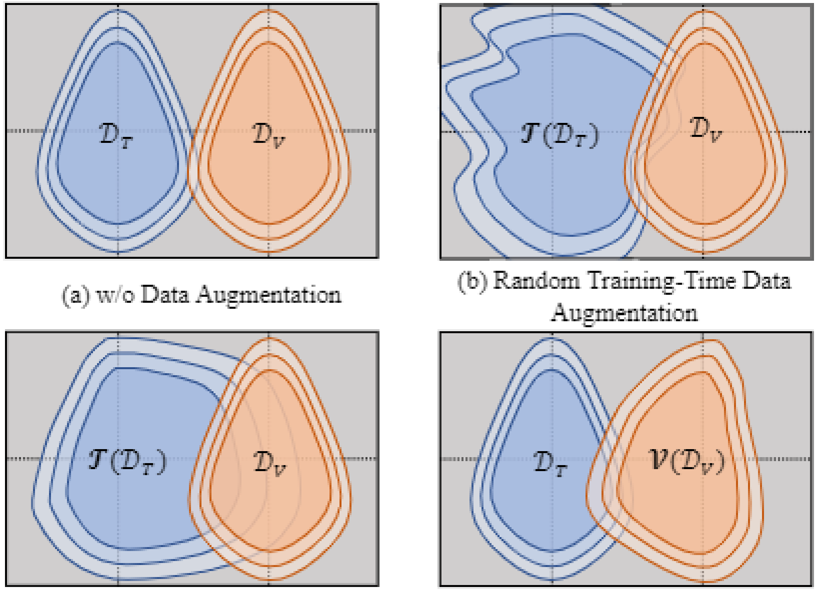


[2] **Z. Li**, *et al.* TMI Under Revision

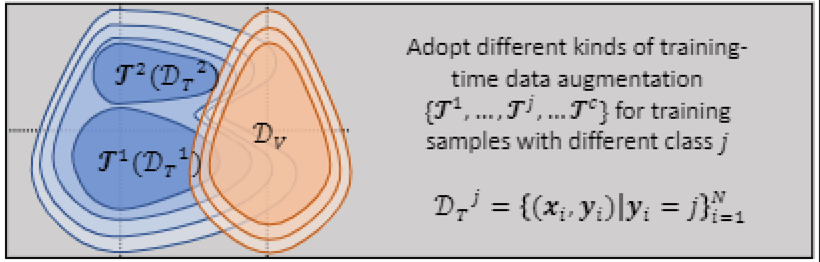
Automatic Data Augmentation

Method

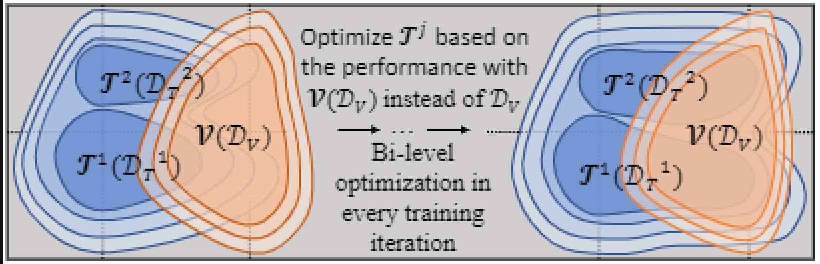
- Data augmentation improves model performance by aligning the training and validation/test data distributions.
- Training-time data augmentation (TRA) and test-time data augmentation (TEA) are closely connected as both aim to **align the training and test data distribution**.



(a) w/o Data Augmentation
(b) Random Training-Time Data Augmentation
(c) Heuristic/Learned Training-Time Data Augmentation
(d) Heuristic/Learned Test-Time Data Augmentation



(e) Learned Class-Specific Training-Time Data Augmentation



(f) Joint Learned Class-Specific Training- and Test- time Data Augmentation

$\mathcal{D}_T$: Distribution of training dataset	$\mathcal{T}$: Training-Time Data Augmentation
$\mathcal{D}_V$: Distribution of validation dataset	$\mathcal{V}$: Test-Time Data Augmentation

[3] **Z. Li**, *et al.* TMI Under Revision

Automatic Data Augmentation

Method

➢ A meta-learning based data augmentation framework, **building a balance** between foreground and background.

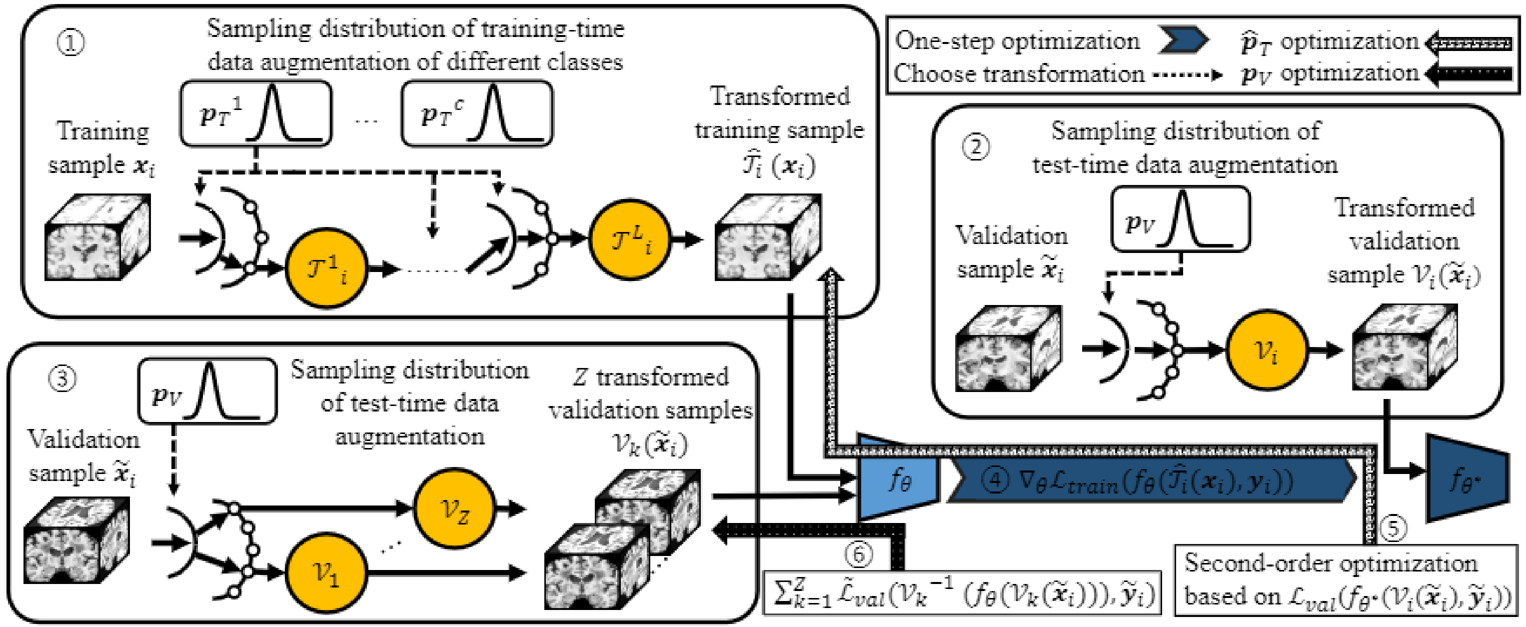


Fig. 2. The optimization process of the proposed method. In this study, data augmentation is formulated as the probability distribution of multiple predefined transformations, as demonstrated in ①, ② and ③. During the same iteration, class-specific TRA is optimized based on meta-gradients with ⑤ while TEA is optimized based on the validation losses of Z transformed samples with ⑥.

[3] **Z. Li**, *et al.* TMI Under Revision

Automatic Data Augmentation

Results

- Consistently improve segmentation performance in various applications.
- Potential to replace the heuristically chosen augmentation policies currently used in most previous works.

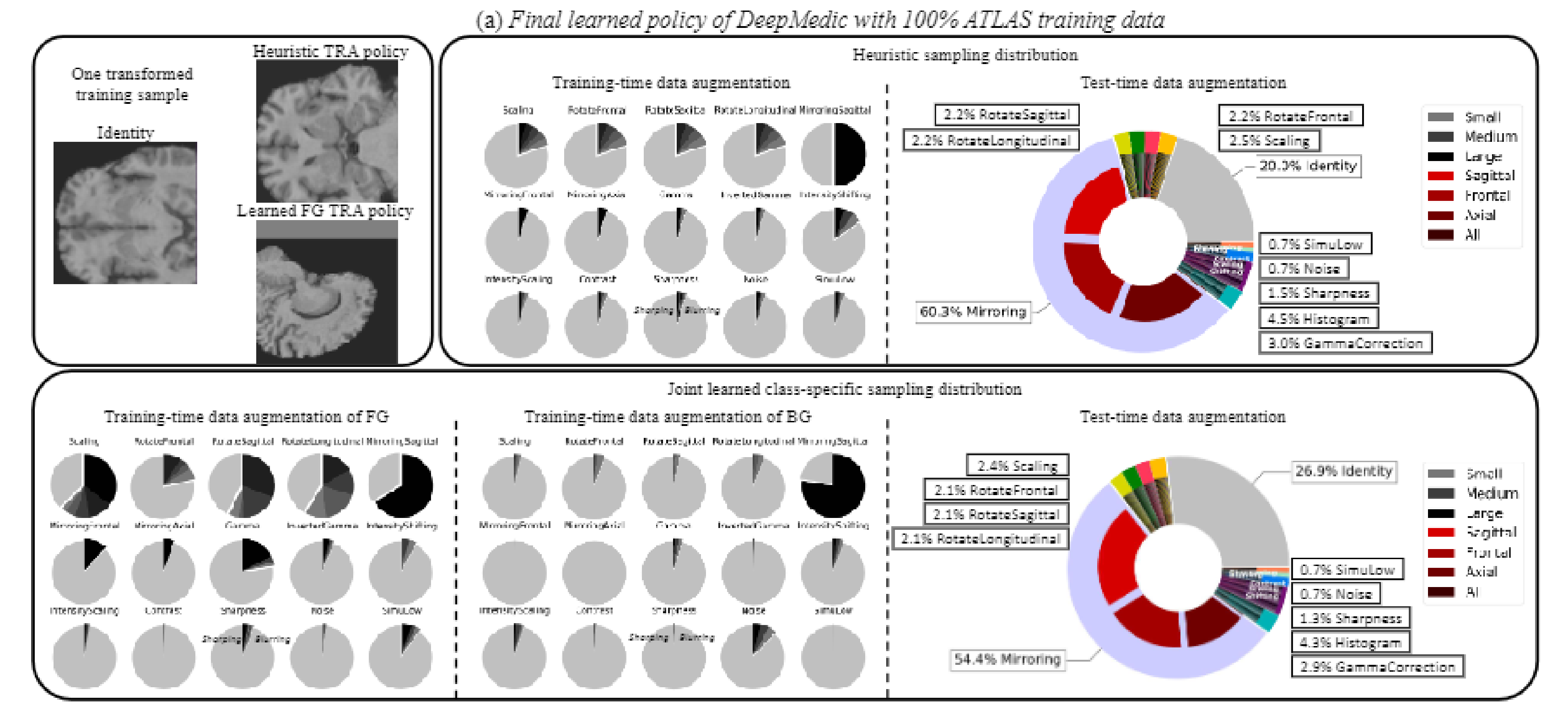
Model	Training data	Training-time data augmentation	Test-time data augmentation	Kidney DSC	Kidney SEN	Kidney PRC	Kidney HD	Tumor DSC	Tumor SEN	Tumor PRC	Tumor HD
DeepMedc [18]	50%	None	None	92.6	89.7	97.0	**8.1**	40.1	35.4	56.2	93.0
		Heuristic [18]	None	95.5	95.1	96.4	16.2	66.6	69.3	72.4	76.3
		Learned [7], [24], [27]	None	**95.8**	95.4	96.4	11.7	**69.1**	71.3	75.1	**61.8**
		Learned Class-Specific	None	**95.7**	94.7	96.9	**9.8**	**71.6**	72.8	76.8	**66.5**
		Heuristic [18]	Heuristic [16]	**95.8**	95.0	97.1	11.0	70.5	70.5	78.5	58.7
		Learned Class-Specific	Heuristic [16]	**95.7**	94.9	97.0	**6.0**	72.5	73.4	78.3	**48.1**
		Learned Class-Specific	Learned [20], [32]	**95.8**	94.9	97.1	**5.9**	**72.8**	73.6	78.5	**47.9**
		Joint Learned Class-Specific		**95.8**	94.9	97.0	11.0	**73.3**	73.5	79.7	48.4
	100%	None	None	94.7	93.5	96.7	**10.6**	51.1	50.0	62.0	**72.8**
		Heuristic [18]	None	**96.0**	96.8	95.3	**18.0**	**69.5**	77.2	69.8	**76.3**
		Learned [7], [24], [27]	None	95.0	97.2	93.2	23.5	**69.5**	79.3	67.6	89.4
		Learned Class-Specific	None	**95.8**	97.1	94.8	19.4	**71.2**	78.1	70.4	88.3
		Heuristic [18]	Heuristic [16]	**96.3**	96.8	96.0	**13.3**	72.9	77.9	74.1	62.5
		Learned Class-Specific	Heuristic [16]	96.0	97.1	95.1	22.3	73.1	79.5	73.2	**57.2**
		Learned Class-Specific	Learned [20], [32]	**96.1**	97.1	95.2	21.9	**73.3**	79.3	73.6	**60.5**
		Joint Learned Class-Specific		**96.1**	96.8	95.5	**12.7**	**74.1**	79.6	74.0	71.7
3D U-Net [5]	50%	None	None	95.3	94.0	97.3	5.7	43.5	39.5	60.9	104.6
		Heuristic [16]	None	**96.6**	96.4	96.9	**2.6**	76.6	80.2	77.4	**40.6**
		RandAugment-S [8]	None	96.4	96.0	97.0	2.7	74.4	76.6	78.3	45.5
		RandAugment-M [8]	None	96.4	95.7	97.2	2.8	77.5	79.0	79.7	**34.2**
		RandAugment-L [8]	None	96.4	96.0	97.0	2.7	**77.6**	82.1	77.0	61.4
		Learned [7], [24], [27]	None	96.5	96.3	96.8	2.8	76.7	82.0	76.1	55.7
		Learned Class-Specific	None	**96.8**	96.6	96.9	**2.5**	**78.4**	82.2	78.0	47.2
		Heuristic [16]	Heuristic [16]	**96.9**	96.6	97.2	**2.3**	**78.8**	82.1	79.4	**37.2**
		Learned Class-Specific	Heuristic [16]	96.8	96.5	97.1	2.5	78.7	81.7	79.6	**42.0**
		Learned Class-Specific	Learned [20], [32]	96.8	96.5	97.1	2.5	**78.8**	81.7	79.6	**42.0**
		Joint Learned Class-Specific		**97.0**	96.9	97.2	**2.2**	**79.3**	82.2	79.7	45.4

[3] **Z. Li**, *et al.* TMI Under Revision

Automatic Data Augmentation

Results

- The learned policies would adopt larger transformations to the foreground than the background samples, **implicitly alleviating the class imbalance issue**.

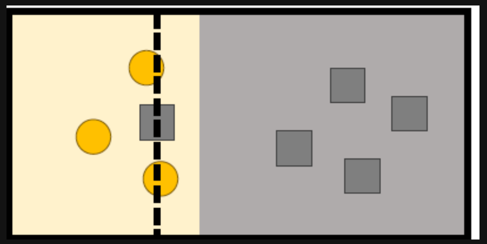


[3] **Z. Li**, *et al.* TMI Under Revision

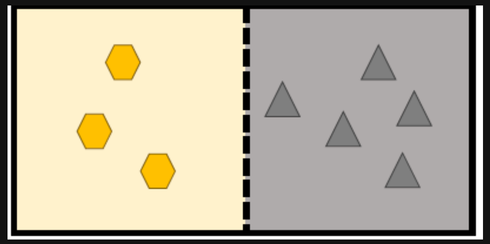
Automatic Data Augmentation

Conclusion

- A data augmentation framework which **bridges the gap** between training and test data distributions.
- We present class-specific TRA, implicitly **addressing the class imbalance problem**.
- We propose to the **joint optimization** of TRA and TEA, which improves alignment of training and test sample distributions and yields better generalization



An automatic data augmentation framework with class-specific transformations

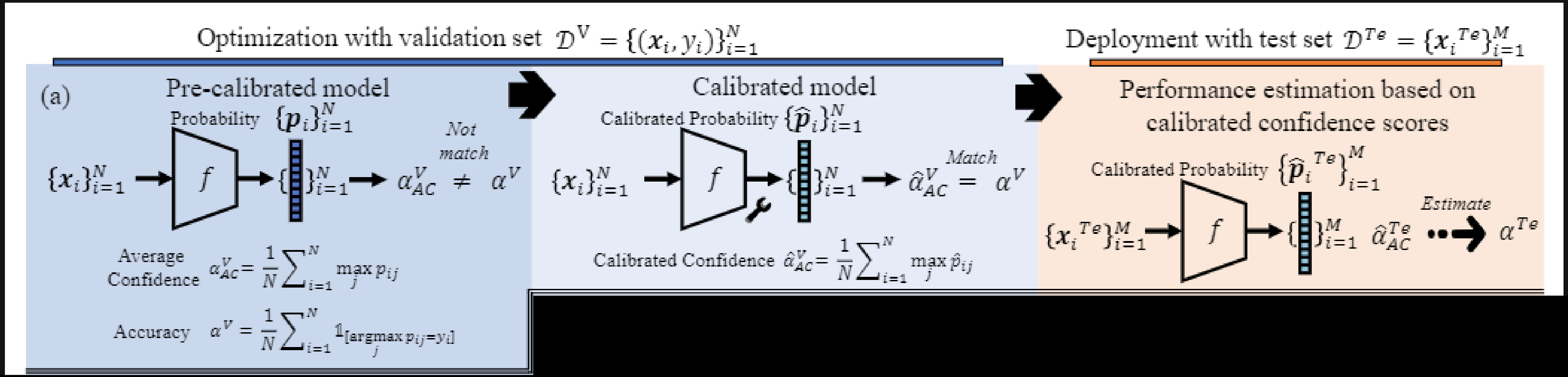


[3] **Z. Li**, *et al.* TMI Under Revision

Class Imbalance under Domain Shifts

Background

- Effect of class imbalance on confidence-based model evaluation methods.

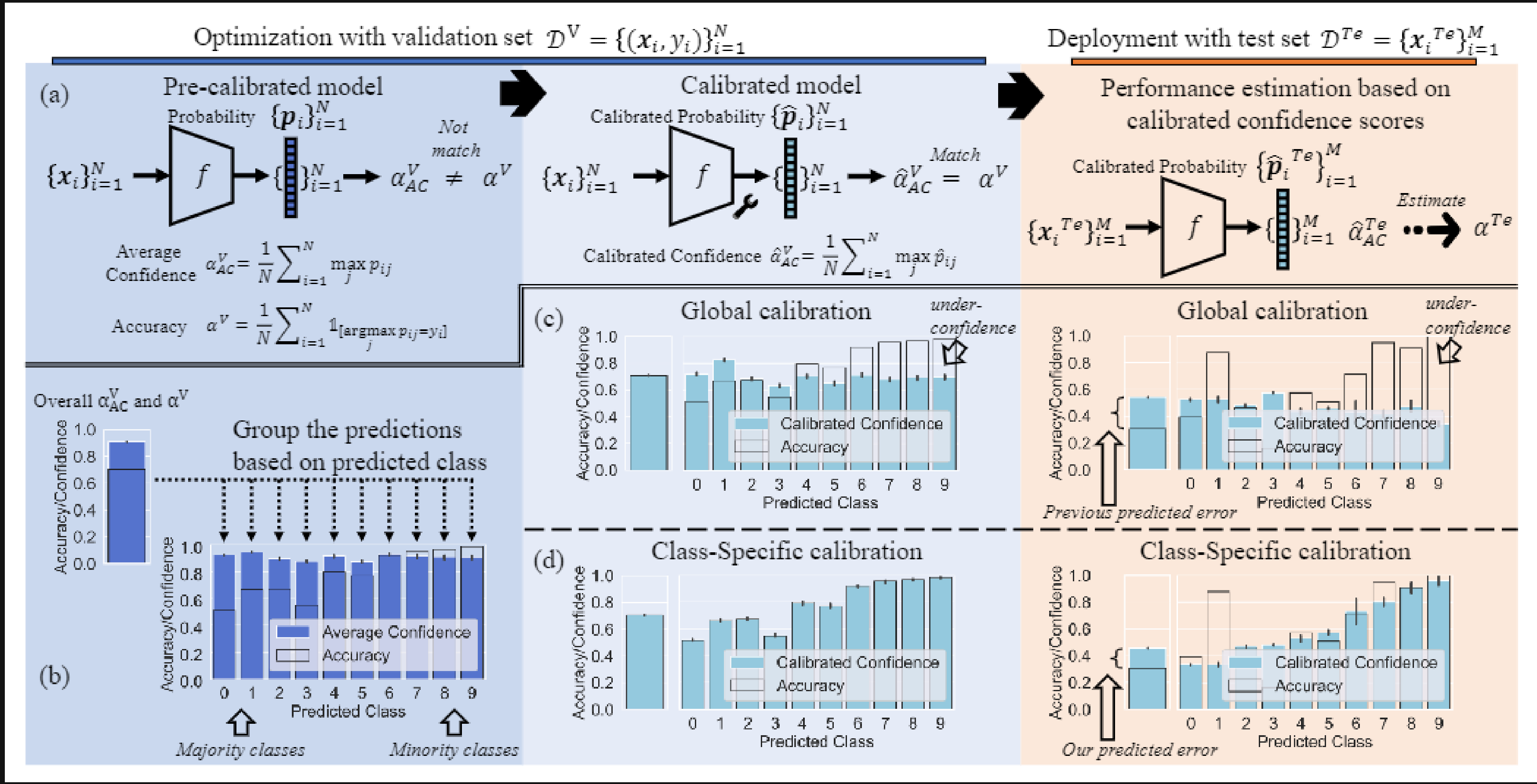


[4] **Z. Li**, *et al.* MICCAI 2022

Class Imbalance under Domain Shifts

Background

➢ Effect of class imbalance on confidence-based model evaluation methods.



[4] **Z. Li**, *et al.* MICCAI 2022

Class Imbalance under Domain Shifts

Method

- Introduce **class-wise calibration** within the framework of performance estimation for imbalanced datasets.

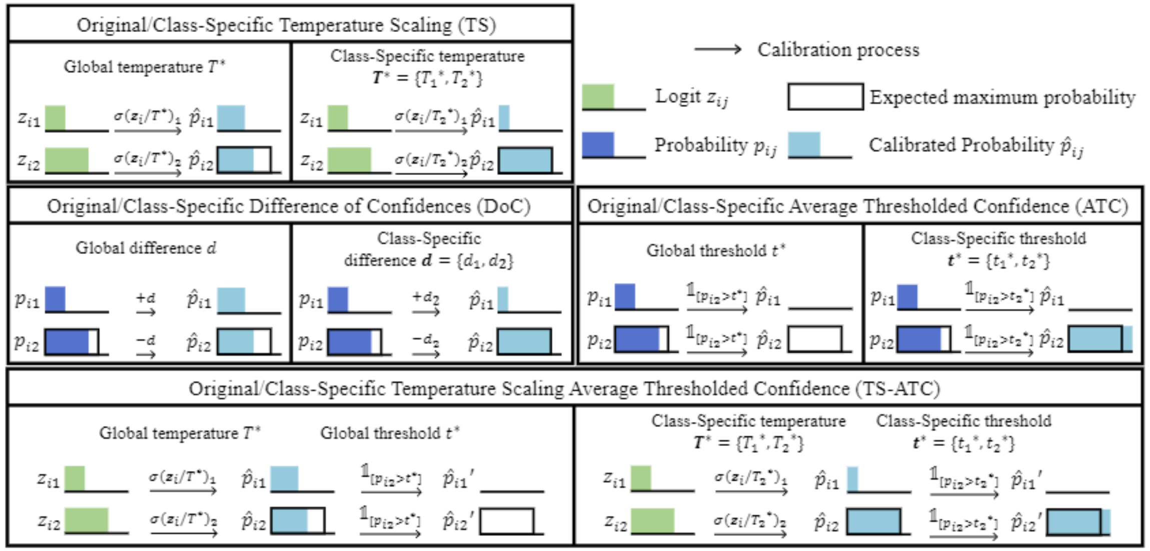


Fig. 2. Illustration of proposed Class-Specific modifications for four existing model evaluation methods. We show the calibration process of an under-confident prediction made for sample from minority class $c = 2$. Prior calibration methods use a global parameter for all classes, which leads to sub-optimal calibration and therefore bias for the minority class. The proposed variants adapt separate parameters per class, enabling improved, class-wise calibration.

[4] **Z. Li**, *et al.* MICCAI 2022

Class Imbalance under Domain Shifts

Results

➢ Consistently improve model estimation accuracy, especially for **segmentation tasks**.

Table 1. Evaluation on different tasks under varied types of domain shifts based on Mean Absolute Error (MAE). Lower MAE is better. Best results with lowest MAE in **bold**. Class-specific calibration as proposed (CS methods) improves all baselines. This is most profound in segmentation tasks, which present extreme class imbalance.

Task	Classification			Segmentation		
Training dataset	CIFAR-10	HAM10000		ATLAS	Prostate	
Test domain shifts	Synthetic	Synthetic	Natural	Synthetic	Synthetic	Natural
AC	31.3 ± 8.2	12.3 ± 5.1	20.1 ± 13.4	35.6 ± 2.1	8.7 ± 4.9	18.7 ± 5.9
QC [30]	—	—	—	3.0 ± 1.7	5.2 ± 6.6	19.3 ± 7.1
TS [12]	5.7 ± 5.6	3.9 ± 4.3	12.1 ± 8.3	9.7 ± 2.5	3.7 ± 5.4	9.2 ± 4.9
VS [12]	3.8 ± 2.1	4.2 ± 4.2	13.6 ± 9.6	11.4 ± 2.5	4.8 ± 5.1	11.2 ± 4.9
NORCAL [29]	7.6 ± 3.8	4.2 ± 4.6	13.7 ± 9.6	6.7 ± 2.4	5.8 ± 5.7	7.3 ± 4.7
CS TS	$5.5^{\sim} \pm 5.6$	$3.7^{\sim} \pm 4.0$	$11.9^{\sim} \pm 8.0$	$1.6^{**} \pm 1.8$	$3.0^{**} \pm 5.7$	$7.8^{**} \pm 4.8$
DoC [11]	10.8 ± 8.2	4.6 ± 5.0	15.3 ± 9.7	4.2 ± 3.2	3.7 ± 5.8	13.9 ± 6.5
CS DoC	$9.4^{**} \pm 7.2$	$4.5^{\sim} \pm 4.9$	$14.7^{*} \pm 9.2$	$\mathbf{1.3^{**} \pm 1.9}$	$3.5^{*} \pm 6.1$	$12.1^{*} \pm 5.9$
ATC [10]	4.6 ± 4.4	3.4 ± 3.9	7.1 ± 6.3	30.4 ± 1.8	8.6 ± 3.3	16.7 ± 5.3
CS ATC	$2.8^{**} \pm 2.9$	$\mathbf{3.3^{\sim} \pm 4.8}$	$\mathbf{5.8^{\sim} \pm 7.6}$	$1.6^{**} \pm 1.5$	$\mathbf{1.1^{**} \pm 1.7}$	$4.3^{**} \pm 2.2$
TS-ATC [10, 12]	5.3 ± 3.9	4.2 ± 4.2	7.3 ± 7.1	30.4 ± 1.8	8.5 ± 3.3	16.7 ± 5.3
CS TS-ATC	$\mathbf{2.7^{**} \pm 2.3}$	$4.2^{\sim} \pm 5.4$	$5.9^{\sim} \pm 8.4$	$\mathbf{1.3^{**} \pm 1.4}$	$1.2^{**} \pm 1.7$	$\mathbf{4.2^{**} \pm 2.2}$

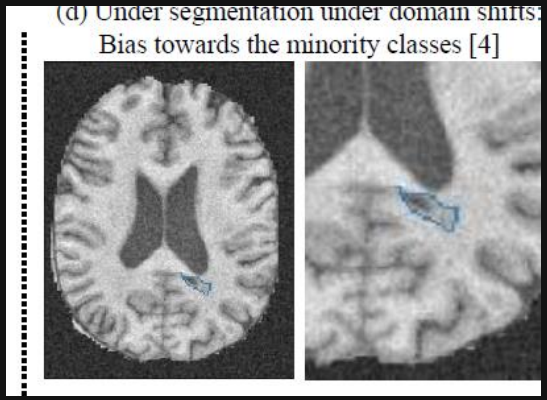
$^{*}p$-value < 0.05; $^{**}p$-value < 0.01; $^{\sim}p$-value ≥ 0.05 (compared with their class-agnostic counterparts)

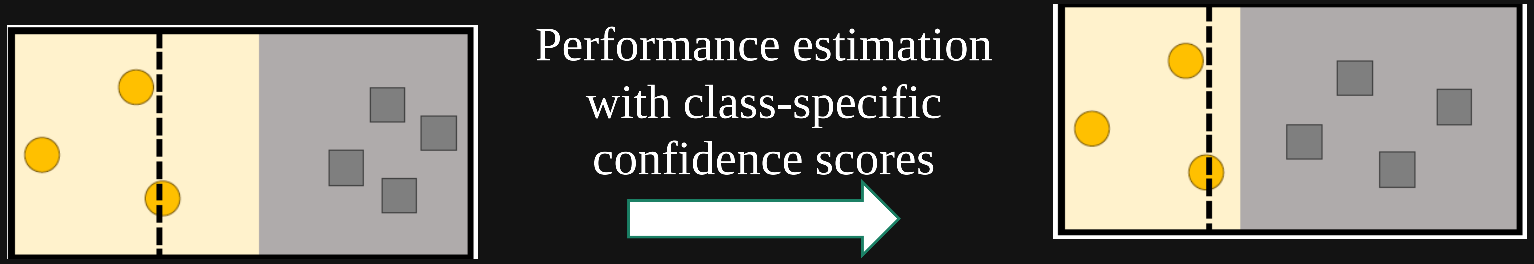
[4] **Z. Li**, *et al.* MICCAI 2022

Class Imbalance under Domain Shifts

Conclusion

- Existing model estimation methods do not account for **bias induced by class imbalance**, thus cannot perform well.
- We derive **class-specific modifications** of state-of-the-art confidence-based model evaluation methods.
- We expect the proposed methods to be useful for **safe deployment** of machine learning in real-world settings.





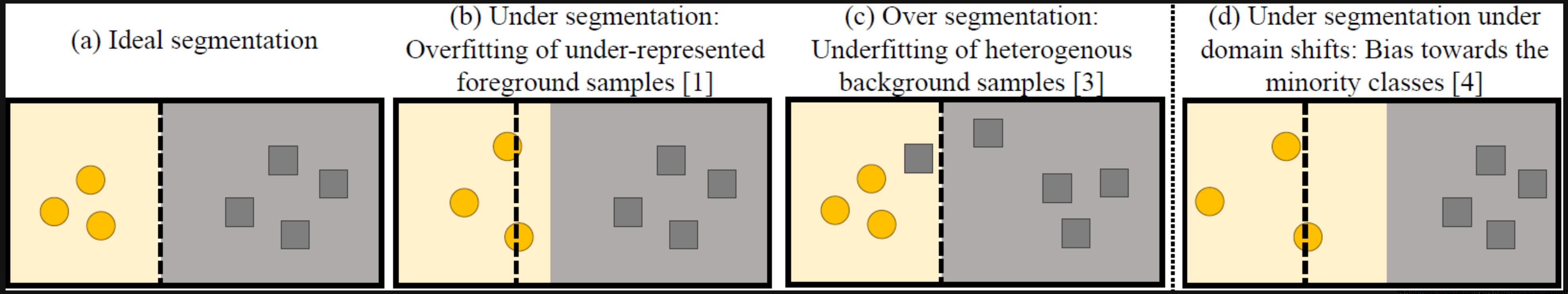
[4] **Z. Li**, *et al.* MICCAI 2022

Conclusion

Take home message

- Class imbalance cause under-segmentation because of **overfitting foreground samples**, while over-segmentation because of **underfitting background samples**.
- Plotting **logit distributions** is useful network inspection tool to gain a better understanding network behaviour under different training scenario, helping us identify the limitations that render problems.

- **Asymmetric** loss functions and regularization techniques help counter overfitting under class imbalance.
- **Context labels** help alleviate underfitting under class imbalance.
- **Class-specific parameters** are beneficial for improving data augmentation and tackling domain shifts.



Thank you!

Contact: zeju.li18@imperial.ac.uk

Appendix

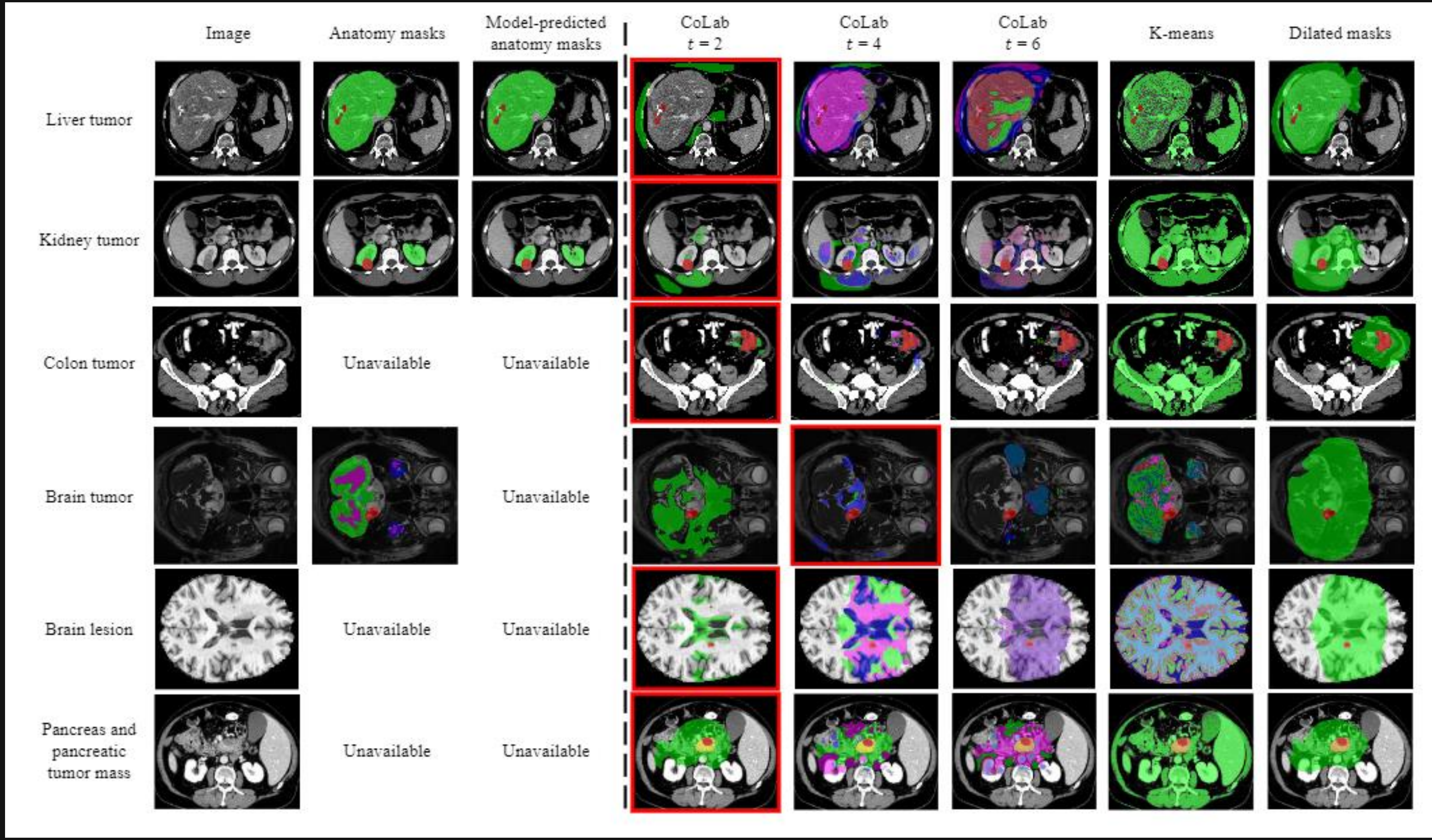
Backup results

Method	Kidney											
	10% training				50% training				100% training			
	DSC	SEN	PRC	HD	DSC	SEN	PRC	HD	DSC	SEN	PRC	HD
Vanilla - w/ augmentation [18]	93.3	91.2	96.9	5.4	96.4	95.8	97.1	2.7	96.6	96.1	97.3	2.4
Vanilla - w/o augmentation	92.3	89.3	96.8	12.1	96.1	95.6	96.7	2.8	96.3	95.8	96.9	2.7
Vanilla - asymmetric augmentation	94.3	92.2	97.0	5.2	94.9	94.5	95.5	5.9	96.1	95.8	96.4	3.8
Large margin loss [31]	**94.6**	92.7	97.1	4.8	96.4	95.9	97.0	2.8	96.1	95.9	96.3	3.2
Asymmetric large margin loss	93.8	91.4	97.2	5.3	96.1	95.5	96.9	2.9	**96.8**	96.6	97.1	**2.2**
Focal loss [29]	91.4	85.9	99.2	10.6	94.1	89.6	99.2	4.2	94.3	90.0	99.1	4.2
Asymmetric focal loss	92.0	86.7	99.0	6.0	94.7	90.9	98.9	3.5	94.8	90.9	99.1	3.1
Adversarial training [12]	94.1	91.9	97.3	9.1	96.3	95.7	97.1	2.6	96.6	96.2	97.2	**2.3**
Asymmetric adversarial training	94.4	92.5	97.2	5.7	**96.6**	96.0	97.3	**2.5**	**96.8**	96.4	97.3	**2.3**
Mixup [47]	**95.0**	93.2	97.3	**4.2**	**96.8**	96.2	97.5	**2.3**	**96.9**	96.4	97.5	**2.2**
Asymmetric mixup	**94.6**	92.6	97.3	**4.5**	96.0	95.2	97.0	3.1	96.4	95.7	97.3	2.7
Symmetric combination	94.1	91.4	97.5	5.1	94.6	91.0	98.7	4.3	96.7	96.2	97.2	**2.2**
Asymmetric combination	93.5	89.7	98.5	5.2	93.9	90.0	98.3	5.3	96.7	95.6	97.9	**2.2**

Method	Kidney tumor											
	10% training				50% training				100% training			
	DSC	SEN	PRC	HD	DSC	SEN	PRC	HD	DSC	SEN	PRC	HD
Vanilla - w/ augmentation [18]	54.6	46.0	80.0	**53.2**	76.0	72.8	86.1	25.1	79.2	77.0	86.2	**17.8**
Vanilla - w/o augmentation	37.4	31.5	65.6	96.0	62.8	58.7	75.9	47.8	73.0	69.1	83.4	18.9
Vanilla - asymmetric augmentation	55.9	48.2	76.4	71.5	74.3	70.3	85.2	33.3	78.4	76.9	85.7	19.8
Large margin loss [31]	52.2	44.3	77.2	68.5	78.2	74.3	87.8	26.6	80.2	79.1	84.5	25.5
Asymmetric large margin loss	55.5	48.3	77.4	71.6	**78.4**	74.9	87.5	24.1	**82.3**	81.4	86.0	**16.9**
Focal loss [29]	47.1	37.5	78.2	74.5	73.0	66.0	87.6	40.2	79.0	73.2	90.0	20.3
Asymmetric focal loss	**57.9**	48.9	78.4	61.4	77.4	74.4	85.0	**20.2**	81.5	80.6	86.7	19.4
Adversarial training [12]	50.9	42.5	81.3	62.0	73.2	69.6	83.9	44.1	81.9	81.1	85.8	27.6
Asymmetric adversarial training	55.2	47.8	79.6	66.7	78.3	74.9	87.9	23.7	82.1	81.1	87.4	19.7
Mixup [47]	53.3	45.2	81.6	57.8	77.0	72.9	87.3	32.1	80.3	78.5	85.9	34.1
Asymmetric mixup	56.8	48.1	84.6	66.5	77.9	74.0	89.2	22.0	79.7	78.1	87.3	19.3
Symmetric combination	53.9	45.1	81.3	70.2	73.9	67.1	87.7	39.6	80.9	79.3	86.5	19.6
Asymmetric combination	**59.2**	52.2	80.3	**49.5**	**79.4**	77.0	86.7	**15.5**	**82.7**	82.1	87.0	18.8

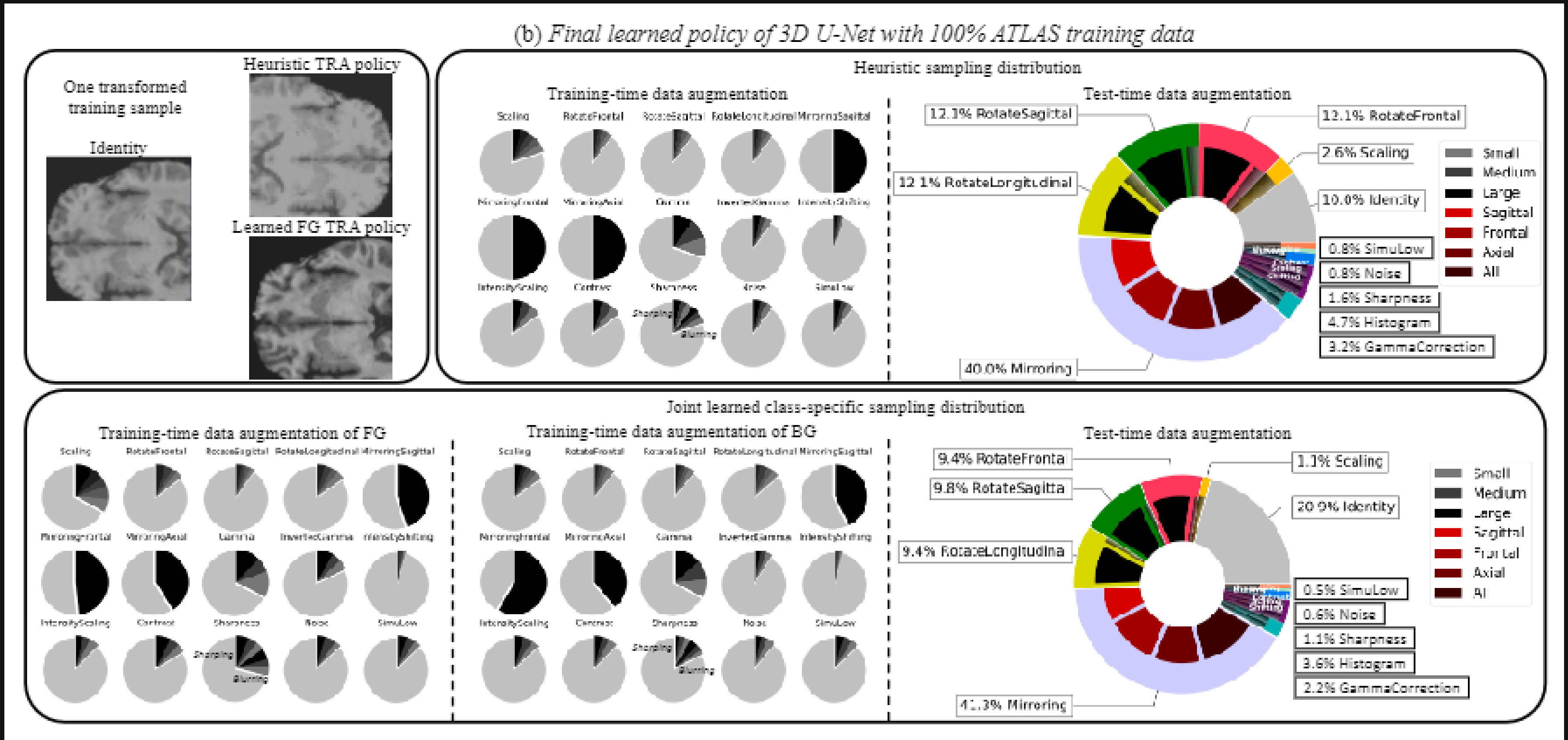
Appendix

Backup results



Appendix

Backup results



Appendix

Backup results