SUPPLEMENTARY MATERIAL

A. THE ANALYSIS OF LARGE MARGIN LOSS FROM A SAMPLE RE-WEIGHTING PERSPECTIVE

Some of the proposed methods are based on two well known loss functions, noted as focal loss and large margin loss. We argue that these two loss functions can be both seen as sample-level re-weighting methods, which change the magnitude of gradient of the network output by multiplying a scalar. Specifically, focal loss would decrease the weights of well-classified samples, while large margin loss would increase the weights of all samples, especially for well-classified samples. Here, we provide an analysis of these loss functions from sample re-weighting perspective.

Following the formulation in the main text, we first consider a CNN trained using cross-entropy loss with one sample $\boldsymbol{x_i}$ and its one-hot label $\boldsymbol{y_i}$:

$$L_{CE}(\boldsymbol{x_i}, \boldsymbol{y_i}) = -\sum_{j=1}^{c} y_{ij} \log(p_{ij}) = -\boldsymbol{y_i} \cdot \log(\boldsymbol{p_i}), \tag{1}$$

where $\boldsymbol{p_i}$ is the calculated probability, which is a normalized term from the network output $\boldsymbol{z_i}$:

$$\boldsymbol{p_i} = \frac{\mathrm{e}^{\boldsymbol{z_i}}}{\mathrm{e}^{\boldsymbol{z_i}} \cdot \mathbf{1}}. \tag{2}$$

Let's look at the gradient of a general loss function $L(\boldsymbol{x_i}, \boldsymbol{y_i})$ with respect to the network parameters θ:

$$\frac{\partial L(\boldsymbol{x_i}, \boldsymbol{y_i})}{\partial \theta} = \frac{\partial L(\boldsymbol{x_i}, \boldsymbol{y_i})}{\partial \boldsymbol{z_i}} \frac{\partial \boldsymbol{z_i}}{\partial \theta}, \tag{3}$$

where the former term is associated with the design of loss functions and the latter is related to the network architecture. Considering a cross entropy loss $L_{CE}(\boldsymbol{x_i}, \boldsymbol{y_i})$, we can have:

$$\begin{aligned} \frac{\partial L_{CE}(\boldsymbol{x_i}, \boldsymbol{y_i})}{\partial \boldsymbol{z_i}} &= \frac{\partial L_{CE}(\boldsymbol{x_i}, \boldsymbol{y_i})}{\partial \boldsymbol{p_i}} \frac{\partial \boldsymbol{p_i}}{\partial \boldsymbol{z_i}} \\ &= -\frac{1}{\boldsymbol{p_i} \cdot \boldsymbol{y_i}} \boldsymbol{p_i} \cdot \boldsymbol{y_i} (\boldsymbol{y_i} - \boldsymbol{p_i}) = \boldsymbol{p_i} - \boldsymbol{y_i}. \end{aligned} \tag{4}$$

In instance-level, we can assign different weights for different samples $\boldsymbol{x_i}$ with different scalars w_i. The weights could be derive based on the frequency of samples or other heuristic rules. Typically, we multiply $L_{CE}(\boldsymbol{x_i}, \boldsymbol{y_i})$ by w_i, and the gradient after re-weighting would become:

$$\frac{\partial \Big(w_i L_{CE}(\boldsymbol{x_i}, \boldsymbol{y_i}) \Big)}{\partial \boldsymbol{z_i}} = w_i (\boldsymbol{p_i} - \boldsymbol{y_i}). \tag{5}$$

It can be seen that re-weighting would change the gradient of the network output for different samples and make the model fit better the chosen samples, which are assigned a larger weight w_i.

Similarly, we can also derive the gradient of the focal loss as:

$$\begin{aligned} \frac{\partial \Big(L_{CE_{focal}}(\boldsymbol{x_i}, \boldsymbol{y_i}) \Big)}{\partial \boldsymbol{z_i}} &= \frac{\partial \Big(L_{CE_{focal}}(\boldsymbol{x_i}, \boldsymbol{y_i}) \Big)}{\partial \boldsymbol{p_i}} \frac{\partial \boldsymbol{p_i}}{\partial \boldsymbol{z_i}} \\ &= \Big((1 - \boldsymbol{p_i} \cdot \boldsymbol{y_i})^{\gamma} - \gamma \boldsymbol{p_i} \cdot \boldsymbol{y_i} \log(\boldsymbol{p_i} \cdot \boldsymbol{y_i}) (1 - \boldsymbol{p_i} \cdot \boldsymbol{y_i})^{\gamma - 1} \Big) (\boldsymbol{p_i} - \boldsymbol{y_i}) \\ &= w_{i_{focal}} (\boldsymbol{p_i} - \boldsymbol{y_i}), \end{aligned} \tag{6}$$

The weight term of focal loss $w_{i_{focal}}$ is a scalar and related to the sample probability $\boldsymbol{p_i} \cdot \boldsymbol{y_i}$. Generally speaking, $w_{i_{focal}}$ would decrease when $\boldsymbol{p_i} \cdot \boldsymbol{y_i}$ is large, therefore focal loss would make the model fit the easy cases less.

More interestingly, we next look into the effect of large margin loss on the gradient. Large margin loss would change the calculation of probability for the training process. Specifically, we substitute $\boldsymbol{p_i}$ with $\boldsymbol{q_i}$ to calculate the loss function, where we require:

$$\boldsymbol{q_i} = \frac{\mathrm{e}^{\boldsymbol{z_i} - \boldsymbol{y_i} m}}{\mathrm{e}^{\boldsymbol{z_i} - \boldsymbol{y_i} m} \cdot \mathbf{1}}, \tag{7}$$

where m is the hyper-parameter for the margin. In this case, the gradient of large margin loss can be derived as:

$$\begin{aligned} \frac{\partial L_{CE_M}(\boldsymbol{x_i}, \boldsymbol{y_i})}{\partial \boldsymbol{z_i}} &= \frac{\partial L_{CE}(\boldsymbol{x_i}, \boldsymbol{y_i})}{\partial \boldsymbol{q_i}} \frac{\partial \boldsymbol{q_i}}{\partial \boldsymbol{z_i}} \\ &= \frac{\mathrm{e}^{\boldsymbol{z_i}} \cdot \mathbf{1}}{\mathrm{e}^{\boldsymbol{z_i} - \boldsymbol{y_i} m} \cdot \mathbf{1}} (\boldsymbol{p_i} - \boldsymbol{y_i}) = w_{i_M} (\boldsymbol{p_i} - \boldsymbol{y_i}). \end{aligned} \tag{8}$$

The weight term of large margin loss w_{i_M} is also a scalar and related to the network output $\boldsymbol{z_i} \cdot \boldsymbol{y_i}$. It can be seen that the existence of a margin m would increase the gradient of sample $\boldsymbol{x_i}$. Moreover, w_{i_M} would be larger as $\boldsymbol{z_i} \cdot \boldsymbol{y_i}$ becomes larger, therefore large margin loss would make the model fit the easy cases more, and keep the distribution of $\boldsymbol{z_i}$ away from the decision boundary.

The analysis for L_{DSC} can be done in a similar way.

B. THE MAGNITUDE OF FOCAL DSC LOSS

The proposed focal DSC loss has similar behaviour with the original of focal loss, as shown in Figure 1. In addition, it does not change the magnitude of loss too much compared with existing solutions [1], [8], making it easier to be combined with other losses. We find it is particularly important for our experiments with 3D U-net [4] because this framework adopts a loss function which is a combination of cross entropy and DSC loss.

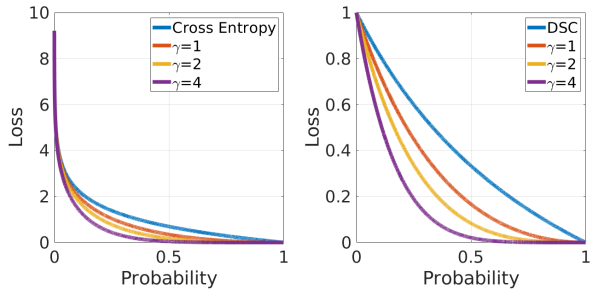

Fig. 1. The comparison of focal loss with cross entropy and DSC loss. The behavior of focal loss for cross entropy and DSC loss are similar using the formulation in equation 10 and 12.

C. Hyper-parameters of the regularization techniques

We summarize the hyper-parameters in Table I, Table II and Table III as a reference for practitioners. We find when the asymmetric regularization techniques are combined together, the network could be regularized too much. In this case, the model would not converge and even perform poorly on the training data. Therefore, we always choose hyper-parameters with smaller regularization magnitude for the experiments with the combined regularization. Empirically, we find decreasing the hyper-parameters of large margin loss and/or focal loss is a sensible choice.

TABLE I
HYPER-PARAMETERS OF EXPERIMENTS USING DEEPMEDIC WITH BRATS.

BRATS		5% data	10% data	20% data	50% data
Individual	large margin m	1	0.2	1	1
	focal γ	2	4	4	4
	adversarial ϵ	1e-5	1e-5	1e-5	1e-5
	adversarial l	10	10	10	20
	mixup λ (symmetric)	0.2	0.2	0.2	0.2
	mixup λ (asymmetric)	1	1	1	1
	mixup m	0.2	0.2	0.2	0.1
Combination	large margin m	1	0	0	0
	focal γ	1.5	2	2	4
	adversarial ϵ	1e-5	1e-5	1e-5	1e-5
	adversarial l	10	10	10	20
	mixup λ (symmetric)	0.2	0.2	0.2	0.2
	mixup λ (asymmetric)	1	1	1	1
	mixup m	0.2	0.2	0.2	0.1

TABLE II
HYPER-PARAMETERS OF EXPERIMENTS USING DEEPMEDIC WITH ATLAS.

ATLAS		30% data	50% data	100% data
Individual	large margin m	0.1	3	2
	focal γ	4	4	4
	adversarial ϵ	1e-5	1e-5	1e-5
	adversarial l	10	10	10
	mixup λ (symmetric)	0.2	0.2	0.2
	mixup λ (asymmetric)	1	1	1
	mixup m	0.8	0.8	0.2
	Probability of background samples being augmented	50%	50%	25%
Combination	large margin m	0.1	0	1
	focal γ	4	3	2
	adversarial ϵ	1e-5	1e-5	1e-5
	adversarial l	10	10	10
	mixup λ (symmetric)	0.2	0.2	0.2
	mixup λ (asymmetric)	——	——	1
	mixup m	——	——	0.2
	Probability of background samples being augmented	50%	50%	——

D. Sensitivity analysis

We conduct a series of controlled experiments with different hyper-parameters to provide more practical details of the proposed regularization techniques. Specifically, we use a baseline DeepMedic model for brain tumor core segmentation with 5% training data of BRATS. The experimental details are consistent with descriptions in Section V. We summarize the results with and without any post-processing in Table IV. We can see from the results that the proposed methods can improve the baseline segmentation results with varied hyper-parameters in most cases. Specifically, asymmetric large margin loss yields improvements for most cases, however, a specific hyper-parameter may yield unexpected results (i.e.

TABLE III
HYPER-PARAMETERS OF EXPERIMENTS USING 3D U-NET WITH KITS.

KiTS		10% data	50% data	100% data
Individual	large margin m	0.8	0.6	0.8
	focal γ	6	6	6
	adversarial ϵ	1e-5	1e-5	1e-5
	adversarial l	100	50	50
	mixup λ (symmetric)	0.2	0.2	0.2
	mixup λ (asymmetric)	1	1	1
	mixup m	0.05	0.05	0.2
	Probability of background samples being augmented	0%	0%	50%
Combination	large margin m	0.8	0.2	0.8
	focal γ	4	6	2
	adversarial ϵ	——	——	——
	adversarial l	——	——	——
	mixup λ (symmetric)	——	——	——
	mixup λ (asymmetric)	——	——	——
	mixup m	——	——	——
	Probability of background samples being augmented	0%	——	——

$m = 0.5$). A potential reason is that the model which focuses on a small portion of easy under-represented samples (c.f. equation 8 in Section A) would overfit more. Asymmetric large margin loss with larger m makes the model emphasize on more under-represented samples and therefore generalize better. Asymmetric adversarial training and asymmetric mixup yields considerable improvements when the perturbation in data augmentation is larger (i.e. $l > 2.5$ for asymmetric adversarial training and $m < 0.8$ for asymmetric mixup). Asymmetric focal loss is robust and can improve the segmentation results with all chosen hyper-parameters. Therefore, we recommend to choose asymmetric focal loss at first for new applications.

E. The intensity histogram of different datasets

Empirically, we find the asymmetric mixup is the most effective method for tumor segmentation with BRATS. However, asymmetric mixup show limited improvements for ATLAS and KiTS. We think it is because the multi-channel information in BRATS could create more useful information, as shown in Figure 2.

F. The quantitative results of abdominal organ segmentation

We evaluate one of our proposed techniques, asymmetric focal loss, with the application of abdominal organ segmentation to demonstrate our method can be feasibly applied to multi-class segmentation. Specifically, we train a model of basic DeepMedic using 25% of the training data, with the same setting in empirical experiments in Section III. Considering the class distribution of the dataset, as shown in Figure 3, we take class 4, class 5, class 8, class 9, class 10, class 11, class 12 and class 13 as rare classes. Specifically, we initiate the one-hot vector $\boldsymbol{r}$ as $[0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1]^\intercal$. We use $\gamma = 4$ in this experiments. We adopt post-processing described in Section V separately to the results of every classes. The results are shown in Table V. The asymmetric focal loss can get better overall segmentation results than cross entropy or its symmetric variant. More importantly. it can get better segmentation results with higher sensitivity for most rare classes. Specifically, asymmetric focal loss can improve

TABLE IV
THE SENSITIVITY ANALYSIS OF DIFFERENT HYPER-PARAMETERS. WE CONDUCT EXPERIMENTS WITH DIFFERENT PARAMETERS WITH BRAIN TUMOR CORE SEGMENTATION (5% TRAINING DATA) WITH BRATS USING DEEPMEDIC. RESULTS WHICH HAVE WORSE DSC THAN THE VANILLA BASELINE ARE HIGHLIGHTED WITH GRAY SHADING.

Method	Parameter		DSC	SEN	PRC	HD
w/ post-processing						
Vanilla - CE	——		50.4	41.0	83.5	18.0
Asymmetric large margin loss	m = 0.2		53.6	44.8	84.8	15.6
	m = 0.5		48.4	39.4	81.5	16.9
	m = 1		56.8	48.9	83.4	15.0
	m = 1.5		54.1	45.6	81.7	15.3
	m = 2		51.6	42.8	84.0	16.7
	m = 3		54.4	45.7	82.3	14.3
Asymmetric focal loss	γ = 0.5		53.4	44.8	79.6	16.6
	γ = 1		53.9	45.2	81.9	17.9
	γ = 1.5		56.5	48.3	87.8	13.8
	γ = 2		58.8	51.4	81.6	15.0
	γ = 3		57.5	49.0	85.5	14.2
	γ = 4		55.2	48.5	78.3	15.5
Asymmetric adversarial training	ϵ = 1e-5	l = 2.5	50.3	41.8	82.0	17.2
	ϵ = 1e-5	l = 5	58.1	50.0	84.7	14.1
	ϵ = 1e-5	l = 10	58.5	50.8	80.1	16.2
	ϵ = 1e-5	l = 15	53.8	46.2	80.1	16.9
	ϵ = 1e-5	l = 20	56.6	50.7	76.9	18.8
	ϵ = 1e-4	l = 10	57.6	51.1	78.9	16.1
	ϵ = 1e-6	l = 10	56.2	48.5	81.1	17.8
Asymmetric mixup	m = 0.1		52.1	47.3	73.8	20.7
	m = 0.15		58.1	53.7	75.0	19.9
	m = 0.2		59.8	56.8	74.7	17.7
	m = 0.25		60.4	55.0	82.0	15.6
	m = 0.3		59.1	54.3	82.0	15.3
	m = 0.4		52.1	44.2	84.2	21.4
	m = 0.8		50.3	41.6	85.5	17.7
w/o post-processing						
Vanilla - CE	——		51.0	42.6	78.6	17.5
Asymmetric large margin loss	m = 0.2		53.2	46.0	79.8	18.3
	m = 0.5		48.8	40.8	78.1	17.5
	m = 1		55.5	50.6	76.2	23.9
	m = 1.5		52.6	47.2	73.1	25.8
	m = 2		51.4	44.2	78.2	18.8
	m = 3		53.4	47.3	75.1	21.3
Asymmetric focal loss	γ = 0.5		54.2	48.0	76.2	22.0
	γ = 1		53.7	46.8	76.0	22.8
	γ = 1.5		54.3	49.6	76.3	25.9
	γ = 2		57.3	52.7	76.4	24.4
	γ = 3		55.7	50.3	75.4	24.6
	γ = 4		54.4	50.3	71.1	25.4
Asymmetric adversarial training	ϵ = 1e-5	l = 2.5	50.5	43.6	76.3	21.3
	ϵ = 1e-5	l = 5	56.6	51.3	76.1	21.9
	ϵ = 1e-5	l = 10	56.8	51.8	74.8	23.6
	ϵ = 1e-5	l = 15	53.3	47.6	74.8	21.5
	ϵ = 1e-5	l = 20	55.4	53.2	72.0	26.2
	ϵ = 1e-4	l = 10	56.9	53.3	74.1	22.4
	ϵ = 1e-6	l = 10	55.2	50.0	76.0	23.8
Asymmetric mixup	m = 0.1		52.0	48.8	68.7	32.2
	m = 0.15		58.0	55.7	70.6	31.6
	m = 0.2		59.3	57.9	70.6	27.8
	m = 0.25		60.1	55.9	78.0	23.5
	m = 0.3		59.2	55.4	77.9	17.6
	m = 0.4		52.8	45.3	80.2	21.5
	m = 0.8		51.0	43.6	79.2	18.7

the average DSC of rare classes by 4.9%. We also notice that asymmetric focal loss would decrease the segmentation performance of esophagus which is taken as a rare class. It is because esophagus is too small, and post-processing would remove the correct segmentation regions by mistake but leave the false positive predictions. We think more advanced post-processing would help improve the segmentation in this case.

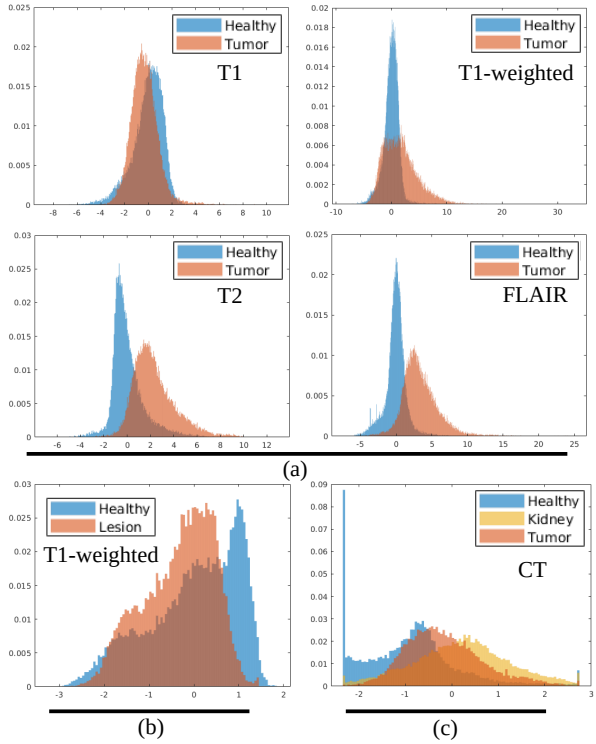

Fig. 2. (a) The intensity histogram of BRATS, (b) ATLAS and (c) KiTS. The intensity of the foreground and background classes overlap a lot for ATLAS and KiTS. This can be a potential factor due to which the asymmetric mixup does not create useful synthetic samples and cannot improve the segmentation performance that much.

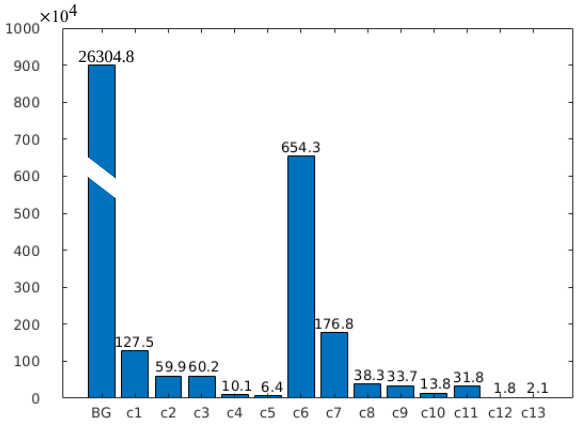

Fig. 3. The class distribution of the abdomen dataset we use in this study. We summarize the total pixel number of different classes. We take class 4, 5, 8, 9, 10, 11, 12 and 13 as rare classes.

G. QUANTITATIVE RESULTS WITHOUT POST-PROCESSING

The quantitative segmentation results without post-processing are summarized in Table VI, Table VII and Table VIII. Without post-processing, the proposed asymmetric regularization methods can improve DSC but could lead to worse distance-based evaluation metrics such as Hausdorff distance (HD). It is because the regularized model, which is more sensitive for the under-represented classes, would make relatively more false positive predictions. The false positive predictions which are far from the ground truth

TABLE V

EVALUATION OF ABDOMEN SEGMENTATION WITH 25% OF TRAINING DATA WITH SYMMETRIC (SY.) AND ASYMMETRIC (ASY.) FOCAL LOSS. THE RARE CLASSES ARE MARKED WITH $\boldsymbol{r}$. AVG IS THE AVERAGE PERFORMANCE OF ALL CLASSES. $\text{AVG}_{\boldsymbol{r}}$ IS THE AVERAGE PERFORMANCE OF ALL RARE CLASSES. BEST RESULTS ARE HIGHLIGHTED IN BOLD.

	c1 (spleen)			c2 (right kidney)			c3 (left kidney)			c4 (gallbladder) $\boldsymbol{r}$			c5 (esophagus) $\boldsymbol{r}$		
	vanilla - CE	sy. focal loss	asy. focal loss	vanilla - CE	sy. focal loss	asy. focal loss	vanilla - CE	sy. focal loss	asy. focal loss	vanilla - CE	sy. focal loss	asy. focal loss	vanilla - CE	sy. focal loss	asy. focal loss
DSC	**85.3**	78.7	84.3	78.3	**84.4**	68.4	82.7	**85.0**	82.6	34.6	30.5	**53.7**	**55.7**	52.7	41.8
Sensitivity	80.0	70.5	76.9	73.7	77.2	62.9	77.5	79.5	77.8	25.8	22.9	46.3	50.7	48.2	43.8
Precision	93.6	96.0	95.9	84.6	94.8	76.2	94.7	94.4	93.0	78.4	68.5	68.9	74.4	69.5	45.0
	c6 (liver)			c7 (stomach)			c8 (aorta) $\boldsymbol{r}$			c9 (vena cava) $\boldsymbol{r}$			c10 (vein) $\boldsymbol{r}$		
	vanilla - CE	sy. focal loss	asy. focal loss	vanilla - CE	sy. focal loss	asy. focal loss	vanilla - CE	sy. focal loss	asy. focal loss	vanilla - CE	sy. focal loss	asy. focal loss	vanilla - CE	sy. focal loss	asy. focal loss
DSC	87.4	88.4	**88.6**	39.7	39.7	**42.1**	82.7	80.3	**84.0**	65.1	66.4	**73.9**	42.0	26.3	**43.3**
Sensitivity	84.1	85.2	84.0	28.8	28.5	31.0	76.6	73.1	82.9	56.8	60.3	76.9	28.4	16.5	31.0
Precision	92.3	92.8	94.2	91.3	84.1	86.3	91.6	91.5	86.3	86.1	79.3	72.7	91.5	82.1	80.0
	c11 (pancreas) $\boldsymbol{r}$			c12 (right adrenal) $\boldsymbol{r}$			c13 (left adrenal) $\boldsymbol{r}$			AVG			$\text{AVG}_{\boldsymbol{r}}$		
	vanilla - CE	sy. focal loss	asy. focal loss	vanilla - CE	sy. focal loss	asy. focal loss	vanilla - CE	sy. focal loss	asy. focal loss	vanilla - CE	sy. focal loss	asy. focal loss	vanilla - CE	sy. focal loss	asy. focal loss
DSC	17.2	24.3	**26.4**	54.3	32.8	**55.9**	34.8	28.9	**47.0**	58.5	55.3	**60.9**	48.3	42.8	**53.2**
Sensitivity	11.1	17.3	18.3	45.3	24.3	50.3	27.2	22.2	41.6	51.2	48.1	55.7	40.2	35.6	48.9
Precision	56.6	52.0	61.7	74.5	60.8	69.4	61.4	52.7	63.6	82.4	78.3	76.4	76.8	69.6	68.5

TABLE VI

EVALUATION OF BRAIN TUMOR CORE SEGMENTATION USING DEEPMEDIC WITH DIFFERENT AMOUNTS OF TRAINING DATA AND DIFFERENT TECHNIQUES TO COUNTER OVERFITTING. THE RESULTS ARE CALCULATED WITHOUT ANY POST-PROCESSING. RESULTS WHICH HAVE WORSE DSC THAN THE VANILLA BASELINE ARE HIGHLIGHTED WITH GRAY SHADING. BEST AND SECOND BEST RESULTS ARE HIGHLIGHTED IN BOLD WITH THE BEST ALSO BEING 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 [5]	51.0	42.6	78.6	17.5	62.8	56.9	81.6	**13.7**	65.3	61.0	83.2	12.8	69.5	66.4	83.8	14.3
Vanilla - CE - 80% tumor	46.2	38.3	77.1	22.5	61.6	55.3	79.1	17.8	65.5	60.9	81.7	17.1	68.8	65.3	83.4	15.1
Vanilla - F1 (DSC)	47.7	38.7	82.0	**15.4**	59.4	52.2	82.7	**13.7**	64.6	59.0	82.9	13.6	67.4	63.4	84.7	13.4
Vanilla - F2 [3]	46.9	38.6	79.2	**14.9**	59.9	53.6	82.6	15.4	66.5	62.1	81.8	**12.7**	68.8	67.0	81.1	14.2
Vanilla - F4 [3]	51.6	44.0	78.6	19.8	60.1	54.1	81.7	16.5	65.9	63.1	80.9	19.1	67.8	65.7	83.1	12.2
Vanilla - F8 [3]	48.6	40.2	80.2	16.8	60.2	53.6	83.2	15.4	64.7	61.3	81.6	15.4	67.9	66.4	79.6	13.7
Large margin loss [7]	46.4	38.3	77.8	21.3	61.2	54.3	82.2	15.3	67.0	62.7	83.3	**12.5**	66.8	63.4	86.1	**11.0**
Asymmetric large margin loss	55.5	50.6	76.2	23.9	64.3	57.9	84.1	**14.3**	67.8	63.9	82.4	13.2	69.3	66.2	84.6	12.7
Focal loss [6]	53.6	46.3	78.7	20.3	62.9	56.0	82.5	17.3	65.2	61.1	82.6	19.2	67.2	63.2	84.7	15.3
Asymmetric focal loss	57.3	52.7	74.4	24.4	66.3	62.9	79.1	16.5	68.6	67.3	78.8	15.6	**71.2**	71.7	79.9	12.4
Adversarial training [2]	53.4	45.7	81.8	21.5	62.4	55.8	83.1	19.4	65.2	60.4	83.4	15.4	66.0	62.0	84.8	17.8
Asymmetric adversarial training	56.8	51.8	74.8	23.6	64.0	59.2	80.5	17.2	68.0	64.7	82.8	15.6	70.6	69.3	81.5	15.0
Mixup [9]	50.0	42.2	77.6	21.1	60.9	55.0	81.4	19.7	64.9	60.0	82.3	17.3	67.2	62.7	86.3	17.3
Asymmetric mixup	**59.2**	57.9	70.7	27.8	**68.5**	66.3	79.2	16.5	**70.6**	69.2	81.2	16.0	70.8	69.1	83.7	**11.1**
Symmetric combination	50.6	43.0	82.2	20.3	61.0	54.2	83.4	23.3	64.9	59.5	85.9	16.8	67.4	63.9	84.4	15.7
Asymmetric combination	**62.4**	64.7	71.5	27.8	**71.4**	73.8	74.3	20.8	**71.9**	74.1	79.2	20.7	**72.9**	77.0	77.8	19.0

would increase HD significantly. However, in practice most false positive predictions could be easily removed by some connected component-based post-processing, as described in Section V. In this way, eventually we can get better or similar HD with our methods, as shown in the main text.

REFERENCES

[1] N. Abraham and N. M. Khan. A novel focal tversky loss function with improved attention u-net for lesion segmentation. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 683–687. IEEE, 2019.

[2] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

[3] S. R. Hashemi, S. S. M. Salehi, D. Erdogmus, S. P. Prabhu, S. K. Warfield, and A. Gholipour. Asymmetric loss functions and deep densely-connected networks for highly-imbalanced medical image segmentation: Application to multiple sclerosis lesion detection. *IEEE Access*, 7:1721–1735, 2018.

[4] F. Isensee, P. F. Jäger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein. Automated design of deep learning methods for biomedical image segmentation. *arXiv preprint arXiv:1904.08128*, 2019.

[5] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Med. Image Anal.*, 36:61–78, 2017.

[6] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[7] W. Liu, Y. Wen, Z. Yu, and M. Yang. Large-margin softmax loss for convolutional neural networks. In *International Conference on Machine Leanring (ICML)*, pages 507–516, 2016.

[8] K. C. Wong, M. Moradi, H. Tang, and T. Syeda-Mahmood. 3d segmentation with exponential logarithmic loss for highly unbalanced object sizes. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 612–619. Springer, 2018.

[9] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018.

TABLE VII

EVALUATION OF BRAIN STROKE LESION SEGMENTATION ON ATLAS USING DEEPMEDIC WITH DIFFERENT AMOUNTS OF TRAINING DATA AND DIFFERENT TECHNIQUES TO COUNTER OVERFITTING. THE RESULTS ARE CALCULATED WITHOUT POST-PROCESSING. RESULTS WHICH HAVE WORSE DSC THAN THE VANILLA BASELINE ARE HIGHLIGHTED WITH SHADING. BEST AND SECOND BEST RESULTS ARE IN BOLD WITH THE BEST ALSO UNDERLINED.

Method	30% training				50% training				100% training			
	DSC	SEN	PRC	HD	DSC	SEN	PRC	HD	DSC	SEN	PRC	HD
Vanilla - w/ augmentation [5]	22.9	22.4	52.3	41.7	47.7	48.5	55.3	**32.8**	55.7	56.8	62.2	30.2
Vanilla - w/o augmentation	17.1	15.2	51.4	38.8	40.5	46.8	45.8	46.4	53.5	56.0	58.2	33.3
Vanilla - asymmetric augmentation	23.3	22.6	49.7	41.8	48.7	51.3	55.0	35.0	**57.1**	58.9	62.7	**29.1**
Large margin loss [7]	20.2	17.0	59.4	**37.2**	46.8	46.2	57.4	34.9	56.0	55.3	64.3	**29.0**
Asymmetric large margin loss	24.0	23.8	50.4	41.4	49.2	52.6	54.7	35.4	56.9	59.9	60.8	27.7
Focal loss [6]	21.9	20.2	55.0	**37.7**	47.8	49.3	55.6	33.7	**57.1**	59.6	63.4	32.6
Asymmetric focal loss	24.7	27.2	42.8	49.0	49.6	56.3	51.5	36.6	56.6	64.7	56.9	31.0
Adversarial training [2]	21.1	18.3	55.1	49.2	48.3	46.1	59.1	35.9	56.4	55.2	65.2	34.0
Asymmetric adversarial training	**27.5**	28.1	52.3	42.2	**50.8**	52.6	57.6	33.3	56.7	58.5	64.3	33.4
Mixup [9]	15.9	14.9	46.0	41.0	47.6	47.5	56.8	**31.6**	55.9	57.4	63.6	30.3
Asymmetric mixup	21.5	24.9	39.4	48.0	47.8	60.3	46.3	45.8	57.0	56.3	67.1	33.8
Symmetric combination	24.6	20.9	63.5	41.3	49.7	51.7	56.0	34.7	56.8	57.0	65.1	29.9
Asymmetric combination	**29.9**	34.2	47.1	47.4	**51.1**	58.6	52.2	38.6	**57.9**	62.4	61.5	32.1

TABLE VIII

EVALUATION OF KIDNEY AND KIDNEY TUMOR SEGMENTATION BASED ON 3D U-NET WITH DIFFERENT AMOUNTS OF TRAINING DATA AND DIFFERENT TECHNIQUES TO COUNTER OVERFITTING. THE RESULTS ARE CALCULATED WITHOUT ANY POST-PROCESSING. RESULTS WHICH HAVE WORSE DSC THAN THE VANILLA BASELINE ARE HIGHLIGHTED WITH SHADING. BEST AND SECOND BEST RESULTS ARE IN BOLD WITH THE BEST ALSO UNDERLINED.

Method	Kidney											
	10% training				50% training				100% training			
	DSC	SEN	PRC	HD	DSC	SEN	PRC	HD	DSC	SEN	PRC	HD
Vanilla - w/ augmentation [4]	93.7	91.7	96.9	5.6	96.5	96.1	97.0	3.6	**96.8**	96.4	97.2	**2.2**
Vanilla - w/o augmentation	92.8	90.2	96.6	12.4	96.3	93.1	96.6	2.5	96.5	96.4	96.8	3.8
Vanilla - asymmetric augmentation	94.7	93.0	96.9	5.2	95.6	95.9	95.8	5.9	96.5	96.6	96.6	3.7
Large margin loss [7]	**94.9**	93.1	97.0	4.7	96.4	96.3	96.7	4.0	96.3	96.7	96.1	4.6
Asymmetric large margin loss	94.1	91.9	97.1	5.9	96.3	95.9	96.8	3.2	**96.8**	96.8	96.8	4.0
Focal loss [6]	91.6	86.1	99.1	6.6	94.1	89.9	99.0	5.5	94.4	90.2	99.1	4.1
Asymmetric focal loss	92.4	87.3	98.9	5.8	94.9	91.3	98.9	3.4	94.9	91.2	99.1	3.0
Adversarial training [2]	94.3	92.3	97.3	6.3	96.5	96.1	97.0	**2.4**	**96.8**	96.5	97.2	**2.2**
Asymmetric adversarial training	94.6	92.8	97.2	4.6	**96.6**	93.4	97.0	3.6	**97.0**	96.7	97.3	**2.1**
Mixup [9]	**95.2**	93.6	97.3	**4.0**	**96.9**	96.4	97.5	**2.2**	**97.0**	96.6	97.3	2.5
Asymmetric mixup	94.8	92.9	97.3	**4.3**	96.1	95.4	97.0	3.1	96.5	95.9	97.3	2.5
Symmetric combination	94.3	91.9	97.4	5.9	94.7	91.2	98.7	4.1	**96.8**	96.4	97.2	**2.1**
Asymmetric combination	94.0	90.5	98.4	4.8	94.3	90.8	95.4	5.1	**96.8**	95.9	97.7	3.3
Method	Kidney tumor											
	10% training				50% training				100% training			
	DSC	SEN	PRC	HD	DSC	SEN	PRC	HD	DSC	SEN	PRC	HD
Vanilla - w/ augmentation [4]	54.9	47.9	77.5	93.8	75.6	73.7	83.9	49.8	79.3	78.3	84.8	48.3
Vanilla - w/o augmentation	37.8	32.9	62.8	121.2	64.1	62.1	74.6	85.6	71.3	69.7	79.3	54.9
Vanilla - asymmetric augmentation	56.1	50.4	74.5	97.1	75.3	73.6	84.3	60.3	79.6	79.5	85.2	**35.0**
Large margin loss [7]	54.8	48.2	77.2	84.0	77.1	75.2	84.5	58.8	80.9	82.1	83.5	47.4
Asymmetric large margin loss	55.7	50.1	75.4	99.6	77.9	76.0	84.9	54.6	**81.9**	82.3	84.0	56.2
Focal loss [6]	48.4	39.3	78.1	80.7	73.0	66.8	86.1	63.0	78.6	73.6	88.1	52.5
Asymmetric focal loss	**57.0**	49.9	74.9	95.9	**78.2**	76.6	84.6	**43.8**	80.8	81.1	83.5	48.9
Adversarial training [2]	51.6	45.0	79.5	**78.3**	73.6	71.6	83.2	55.3	81.4	81.6	84.0	52.2
Asymmetric adversarial training	56.9	51.1	79.4	87.5	77.4	75.6	85.5	58.3	81.8	81.5	86.1	**30.8**
Mixup [9]	54.5	48.3	79.6	**74.3**	77.1	73.8	86.2	48.3	80.6	79.5	84.9	52.0
Asymmetric mixup	55.1	48.6	79.9	92.3	77.9	74.4	87.9	**40.8**	79.8	79.0	85.9	54.9
Symmetric combination	54.2	47.1	79.0	105.4	73.6	67.5	86.0	51.5	80.5	80.0	84.5	48.4
Asymmetric combination	**59.2**	54.1	77.1	82.8	**79.4**	79.0	85.1	50.6	**82.2**	82.7	85.0	36.7