

# Take the Bull by the Horns: Learning to Segment Hard Samples

## Supplementary Material

### A. Theoretical analysis

#### A.1. Generalization error bound analysis

Suppose there is a hypothesis function  $f : \mathcal{X} \rightarrow \mathcal{Y}$  outputting a target  $y \in \mathcal{Y}$ , given an input data  $x \in \mathcal{X}$ . Sample set  $\mathcal{S}$  consists of  $n$  instances  $(x_1, y_1), \dots, (x_n, y_n)$ . With this sample set, we want to use a DNN model with a weight matrix  $W$  to approximate  $f(x)$ .

Given a non-negative real-valued loss function  $\ell$ , we aim to establish a theoretical generalization error bound between the expected loss  $\mathbb{E}_P[\ell(Wx, y)]$  and the empirical loss  $\mathbb{E}_{P_n}[\ell(Wx, y)]$ , which is crucial for understanding the generalization ability of the model and providing theoretical support for model optimization. In particular, the empirical loss of given training samples distribution  $P_n$  is defined as

$$\mathbb{E}_{P_n}[\ell(Wx, y)] = \frac{1}{n} \sum_{(x_i, y_i) \in \mathcal{S}} \ell(Wx_i, y_i), \quad (1)$$

which refers to the average loss observed for each sample in the training data.

On the other hand, the expected loss is the expectation of loss under the population data distribution  $P$ , defined as

$$\mathbb{E}_P[\ell(Wx, y)] = \mathbb{E}_{(x, y) \sim P}[\ell(Wx, y)]. \quad (2)$$

Then, the difference between the empirical loss and the expected loss is called a generalization error:

$$GE(\ell) = \|\mathbb{E}_P[\ell(Wx, y)] - \mathbb{E}_{P_n}[\ell(Wx, y)]\|. \quad (3)$$

The following assumptions and conclusions are made in [12] for generalization error.

##### A.1.1. Key assumptions

We truncate the loss function during our analysis to control potential large errors. By choosing a truncation value to be  $B > 0$ , the truncated loss function is defined as

$$\tilde{\ell}(Wx, y) = \min(B, \ell(Wx, y)), \quad (4)$$

which aims to minimize the impact of extreme cases (like outliers) on the loss function, thereby enhancing the stability of the generalization error bound.

Assume that the loss function  $\ell$  is convex and satisfies the following assumption regarding its second-order derivatives.

$$\|\nabla \text{Tr}(H_\ell(\cdot, y))\|_2 \leq \tau \text{Tr}(H_\ell(\cdot, y)), \quad (5)$$

where  $H_\ell(\cdot, y)$  is the second derivative of the loss function  $\ell$  (i.e., the Hessian matrix), and  $\tau > 0$  is a control constant.

The assumption suggests that the Hessian matrix is well-controlled within the function space, preventing excessive fluctuations. Such control is crucial as it limits the curvature of the loss function, enabling better management of the generalization error.

##### A.1.2. Taylor expansion

To analyze the behavior of the loss function, we expand it around the point  $Wx_0$ , which can be chosen as the origin. The loss function  $\ell(Wx, y)$  can be approximated by the Taylor expansion

$$\ell(Wx, y) \approx \nabla \ell(0, y) \cdot Wx + \frac{1}{2} Wx^\top H_\ell(0, y) \cdot Wx, \quad (6)$$

where  $\nabla \ell(0, y)$  represents the first derivative of the loss function  $\ell$  with respect to  $Wx_0$ , that is, the Jacobian matrix;  $H_\ell(0, y)$  represents the second derivative of the loss function  $\ell$ , i.e., the Hessian matrix.

##### A.1.3. Inequalities and concentration bounds

We use concentration inequalities to bound the deviation between empirical loss and expected loss. Here, we apply **Bernstein's Inequality** for its tighter bounds compared to other concentration inequalities. Specifically, we define a sequence of random variables  $Z_i = \ell(Wx_i, y_i)$  and compute their variance

$$\text{Var}(Z_i) = \mathbb{E}[(Z_i - \mathbb{E}[Z_i])^2]. \quad (7)$$

Then Bernstein's inequality gives us the following bound

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n Z_i - \mathbb{E}[Z]\right| \geq \epsilon\right) \leq 2 \exp\left(-\frac{n\epsilon^2}{2\sigma^2 + \frac{2}{3}M\epsilon}\right), \quad (8)$$

where  $\sigma^2 = \text{Var}(Z_i)$  is the variance and  $M$  is a constant controlling the range of  $Z_i$ . This inequality demonstrates that as the sample size  $n$  increases, the probability of a large deviation between empirical and expected losses decreases exponentially.

##### A.1.4. Definitions

To quantify the bounds on the generalization error, we introduce two important quantities.

###### • Norm of the Jacobian matrix:

$$\mu(W) = \mathbb{E}_{P_n}[\|\nabla \ell(Wx, y)\|_2], \quad (9)$$

which measures the overall magnitude of the first-order derivative of the loss function, reflecting the sensitivity of the loss function to changes in the input  $Wx$ .

• **Trace of the Hessian matrix:**

$$\nu(W) = \mathbb{E}_{P_n} [\text{Tr}(H_\ell(Wx, y))], \quad (10)$$

which measures the second-order derivative of the loss function, describing the local curvature of the loss function in the input space.

**A.1.5. Generalization error bounds**

Combining the previous process, we decompose the error bound into several key error terms. Through calculation, we obtain:

$$\mathbb{E}_P[\ell(Wx, y)] \leq \mathbb{E}_{P_n}[\ell(Wx, y)] + \mathcal{O}\left(\frac{\sigma}{\sqrt{n}} + \frac{M}{n}\right). \quad (11)$$

By expanding  $\sigma$  and  $M$ , we obtain

$$\sigma \sim A\sqrt{B\nu(W)}\theta,$$

and

$$M \sim B \log^3(nc).$$

Reviewing the generalization error results from [12], we have:

**Theorem 1** (Theorem 4.1 in [12]). *With probability  $1 - \delta$  over the training examples, for all weight matrices  $W$  satisfying the norm bound  $\|W^T\|_{2,1} \leq A$ , the following holds*

$$\begin{aligned} \mathbb{E}_P[\ell(Wx, y)] - 1.01\mathbb{E}_{P_n}[\ell(Wx, y)] &\leq \frac{(A\mu(W))^{\frac{2}{3}}(\theta B)^{\frac{1}{3}}}{n^{\frac{1}{3}}} \\ &+ \frac{A\sqrt{B\nu(W)}\theta}{\sqrt{n}} + \frac{BA^2\theta}{n\left(\log^2\left(\frac{BA^2\theta}{\nu(W)n}\right) + 1\right)} + \zeta, \end{aligned} \quad (12)$$

where  $\mu(W), \nu(W)$  measure the Jacobian and Hessian of the loss, respectively. Additionally, we define  $\theta = \log^3(nc) \max_i \|x_i\|_2^2$  and  $\zeta = \frac{B(\log(1/\delta) + \log \log n)}{n}$  is a low-order term.

**A.2. Regularizing of Hessian trace is necessary**

**A.2.1. Impact of Hessian trace**

According to **Theorem 1**, the model achieves strong generalization when both the Jacobian norm and the Hessian trace are small. While gradient descent minimizes the loss function to reach a zero-gradient point, it primarily reduces the Jacobian norm by addressing first-order derivatives. However, it does not inherently constrain the Hessian trace, which involves second-order derivatives. Thus, imposing additional constraints on the Hessian trace is essential when updating parameters using gradient descent. Inspired by this, Liu et al. [7] incorporated the Hessian trace as an additional penalty term in the loss function

$$\ell_{all}(f(x), y) \approx \ell(f(x), y) + \lambda \cdot \text{Tr}(H_{\ell, W}), \quad (13)$$

where  $\lambda$  controls the contribution of Hessian regularization. Furthermore, they discussed the impact of penalizing the Hessian trace on the flatness of the minima and the linear stability.

**A.2.2. Flat minima vs. sharp minima**

During deep learning model training, parameters may converge to different minima, characterized by the distribution of eigenvalues of the Hessian matrix.

- (1) **Flat Minima:** A Hessian matrix with smaller eigenvalues ( $\lambda_i$ ) implies a smaller trace  $\text{Tr}(H) = \sum_{i=1}^d \lambda_i$ , indicating that the loss surface is relatively flat at this location, corresponding to a stronger generalization ability.
- (2) **Sharp Minima:** A large Hessian eigenvalue indicates a high trace  $\text{Tr}(H)$ , reflecting steep curvature of the loss surface. High curvature at a local minimum often indicates that the model is highly sensitive to small input perturbations, increasing the risk of overfitting.

To intuitively understand the relationship between the Hessian and the minimum properties, we examine the second-order approximation of the loss function  $\ell(Wx)$  near a local minimum point  $Wx_0$ :

$$\ell(Wx) \approx \ell(Wx_0) + \frac{1}{2}(Wx - Wx_0)^\top H(Wx_0)(Wx - Wx_0), \quad (14)$$

where  $H(Wx_0)$  represents the Hessian matrix of the loss function at the point  $Wx_0$ . If the trace of  $H(Wx_0)$  is small, it indicates that the loss function changes slowly around  $Wx_0$ , and the minimum is relatively flat. In contrast, it suggests that the loss function changes rapidly near this point, and the minimum is relatively sharp.

**A.2.3. Linear stability analysis**

Further discussion explores neural network optimization through linear stability analysis and stochastic gradient descent (SGD). By treating parameter updates as a dynamic system, the stability of equilibrium points is key to understanding convergence. The Hessian matrix is crucial for determining stability; penalizing its trace reduces the eigenvalues, which helps the optimizer to escape local minima and avoid easy-to-converge equilibrium points. It aligns with **Lyapunov stability theory** in [8], highlighting the need to destabilize certain equilibrium points for better optimization outcomes.

**A.2.4. Implicit regularization on Hessian trace**

Although [7] introduced the estimated Hessian trace as an additional penalty in the loss function, its computation via Hutchinson and Dropout methods has notable limitations: it significantly raises computational complexity, introduces potential estimation errors, and may overly smooth the model, potentially hindering its capacity to capture complex data patterns.

In contrast, our approach employs a diffusion process to introduce controlled randomness directly for feature enhancement, thereby improving the model’s segmentation accuracy on hard samples. Rather than relying on computationally heavy Hessian trace approximations, we use a diffusion process that implicitly regularizes the model. By taking expectations and applying a Taylor expansion, we derive our expected loss function as

$$\mathbb{E}[\ell(f(x + g(\epsilon)), y)] \approx \ell(f(x), y) + \frac{1}{2} \text{Tr}(H\Sigma_g), \quad (15)$$

where  $\Sigma_g$  is the covariance matrix of  $g(\epsilon)$ . Building on the previous theoretical analysis, it demonstrates how randomness augmentation implicitly regularizes the model. Here,  $\ell(f(x), y)$  represents the original loss function and  $\frac{1}{2}\text{Tr}(H\Sigma_g)$  is the regularization term. The trace of the Hessian matrix,  $\text{Tr}(H)$ , indicates the curvature of the loss surface. Larger curvature often results in sharp minima, which can harm generalization. Penalizing  $\text{Tr}(H)$  promotes convergence to flatter minima, enhancing generalization. Additionally, penalizing the Hessian trace helps the model avoid local stable points, reducing overfitting to specific hard samples. Through the integration of randomness via a diffusion process, our method achieves a more effective balance in the model’s ability to generalize and handle complex data patterns while maintaining computational efficiency.

## B. Experimental results

We expand on the experimental results by detailing the datasets and evaluation metrics and presenting additional experimental findings.

### B.1. Description of datasets

To evaluate the performance of our L2S, we carried out experiments across seven different medical image segmentation datasets, as described below.

#### B.1.1. Carotid artery segmentation

We used the carotid artery MRI dataset from the CarOtId Vessel Wall Segmentation And Atherosclerosis Diagnosis Challenge (COSMOS 2022) [13]. This dataset comprises 75 MR scans, with 45 scans for training (1,875 axial slices), 5 scans for validation (212 axial slices) and 25 scans for testing (1,241 axial slices). The annotations include both the Lumen and Outer Wall regions. Our focus is on segmenting the vessel wall, defined as the area obtained by subtracting the Lumen from the Outer Wall label.

#### B.1.2. Skin lesion segmentation

We used the ISIC2018 segmentation dataset [10], which consists of 2,594 annotated dermoscopic images aimed at accurately delineating skin lesion boundaries.

#### B.1.3. Polyp segmentation

We used the Kvasir-SEG dataset [5], which contains 1,000 polyp images, each paired with a corresponding segmentation mask. These images are derived from various procedures, including colonoscopies, encompassing a wide range of polyp sizes, shapes, and appearances.

#### B.1.4. Breast cancer segmentation

We used the BUSI dataset [1] for breast cancer segmentation. To make the model focus on segmentation on difficult samples, we excluded images labeled ‘normal’ and used 1,312 images (891 benign and 421 malignant) from this dataset.

#### B.1.5. Cardiac organ segmentation

We used the ACDC dataset [2] for cardiac organ segmentation. It contains 100 cardiac MRI scans having three sub-organs, namely the right ventricle (RV), myocardium (Myo), and left ventricle (LV). Following TransUNet [3], we used 70 cases (1,930 axial slices) for training, 10 for validation, and 20 for testing.

#### B.1.6. Abdomen organ segmentation

We used the BTCV multi-organ dataset [4] for abdomen organ segmentation. This dataset includes 30 abdominal CT scans with a total of 3,521 axial contrast-enhanced slices, averaging 127 slices per scan, each with a resolution of  $512 \times 512$  pixels.

#### B.1.7. Brain tumor segmentation

We used the BraTS2020 [9] for brain tumor segmentation. This dataset includes 369 brain MRI scans with a total of 17,391 axial contrast-enhanced slices, each with a resolution of  $240 \times 240$  pixels.

In our implementation, we used an 80:10:10 train-validation-test split for the ISIC2018, Kvasir-SEG, BUSI, BTCV and BraTS2020 datasets. For the other datasets, we adhered to the default train-validation-test splits.

## B.2. Evaluation metrics

We use the DICE score to evaluate the performance across all segmentation datasets and include IoU as an additional metric for four binary segmentation datasets. The DICE score  $DSC(Y, P)$  and IoU  $IoU(Y, P)$  are calculated as follows

$$DSC(Y, P) = \frac{2 \times |Y \cap P|}{|Y| + |P|} \times 100, \quad (16)$$

and

$$IoU(Y, P) = \frac{|Y \cap P|}{|Y \cup P|} \times 100, \quad (17)$$

where  $Y$  and  $P$  are the ground truth and predicted segmentation map, respectively.

Dataset	TS	LHS	IHS	Overlap	Jaccard Index (%)	Precision (%)
COSMOS 2022	1241	185	204	185	90.69	100.00
ISIC 2018	519	118	130	109	78.42	92.37
BUSI	262	80	101	75	70.75	93.75
Kvasir-SEG	200	33	39	28	63.64	84.85

Table 1. Evaluation of hard sample identification across different datasets, where TS denotes total samples during testing, LHS denotes the labeled hard samples, and IHS denotes the hard samples identified by our model with a Dice score below 0.7. Besides, overlap represents the number of intersections between labeled hard samples and identifiable hard samples.

Method	COSMOS2022	ISIC2018	Kvasir-SEG	BUSI	$p$ -value
UNet	82.44 $\pm$ 0.8	87.14 $\pm$ 1.8	85.63 $\pm$ 1.4	76.41 $\pm$ 1.2	0.007
PolypPVT	82.33 $\pm$ 1.3	89.84 $\pm$ 2.4	91.44 $\pm$ 0.9	81.05 $\pm$ 1.8	0.035
nnUNet	84.12 $\pm$ 0.6	88.91 $\pm$ 1.6	90.55 $\pm$ 1.1	80.94 $\pm$ 1.2	0.037
MedT	82.81 $\pm$ 1.7	88.84 $\pm$ 2.5	89.68 $\pm$ 1.4	80.44 $\pm$ 1.8	0.001
TransUNet	82.92 $\pm$ 1.3	89.44 $\pm$ 2.3	91.04 $\pm$ 1.8	80.32 $\pm$ 1.4	0.001
TransFuses	83.17 $\pm$ 1.6	89.96 $\pm$ 3.1	91.27 $\pm$ 2.2	81.50 $\pm$ 2.3	0.010
SwinUNet	83.63 $\pm$ 1.1	89.56 $\pm$ 2.0	90.22 $\pm$ 1.8	79.76 $\pm$ 1.9	0.012
MedSegDiffV2	82.89 $\pm$ 3.2	89.78 $\pm$ 3.3	91.06 $\pm$ 2.8	80.53 $\pm$ 2.4	0.001
Mask2former	82.76 $\pm$ 1.3	89.59 $\pm$ 2.1	90.86 $\pm$ 2.0	80.19 $\pm$ 1.4	0.001
<b>L2S</b>	<b>84.16 <math>\pm</math> 1.8</b>	<b>90.97 <math>\pm</math> 2.9</b>	<b>91.87 <math>\pm</math> 2.3</b>	<b>81.82 <math>\pm</math> 2.0</b>	-

Table 2. The Wilcoxon Signed-Rank Test on the Binary Segmentation Task Dataset Based on DSC.

Method	ISIC 2018		Kvasir-SEG	
	DSC	DSC <sup>†</sup>	DSC	DSC <sup>†</sup>
SAM	89.58 $\pm$ 2.5	59.41 $\pm$ 3.4	88.26 $\pm$ 2.0	58.92 $\pm$ 3.8
SAM2	88.45 $\pm$ 2.2	57.93 $\pm$ 2.8	86.73 $\pm$ 1.6	58.18 $\pm$ 4.1
<b>L2S</b>	<b>90.97 <math>\pm</math> 2.9</b>	<b>63.90 <math>\pm</math> 4.7</b>	<b>91.87 <math>\pm</math> 2.3</b>	<b>64.34 <math>\pm</math> 3.9</b>

Table 3. Comparison with SAM-based methods. DSC<sup>†</sup> is computed on the hard samples with dice score below 0.7.

### B.3. Effectiveness of hard sample identification

We evaluated whether the predicted hard samples genuinely reflect the characteristics of dataset-specific hard samples. While the dataset lacks a precise definition of hard samples, it is widely accepted that samples with lesion regions that are challenging to segment accurately qualify as hard samples. Accordingly, we utilized the lesion labels provided in the segmentation dataset to define hard samples for this analysis. Specifically, taking the four datasets from the first experiment in the main text as examples, the hard samples are defined as follows: bifurcations and atherosclerotic regions in COSMOS 2022; Melanoma, Basal Cell Carcinoma, Dermatofibroma, and hair artifacts in ISIC 2018; malignant lesions in BUSI; and polyps and flat lesions in Kvasir-SEG. These hard samples all exhibit significant class imbalance.

Table 1 shows the comparison between the labeled hard samples (LHS) and identified hard samples (IHS) selected by our hard sample identification module. It can be ob-

served that the majority of the predicted hard samples overlap with the labeled hard samples, indicating that our model effectively identifies hard samples that exhibit dataset-specific characteristics. Even for datasets like Kvasir-SEG, which have a relatively low Jaccard Index (overlap ratio), our method maintains a consistently high precision across all datasets. It indicates that despite variations in overlap metrics, the majority of true hard samples are still included among the identified hard samples.

As observed in Table 1, the number of hard samples identified by our model exceeds the number of labeled hard samples in the dataset. This is because hard samples are not limited to lesion-related instances. To further analyze this, we visualized other samples identified as hard samples, as shown in Figure 1. The results demonstrate that the identified hard samples encompass the types described in the main text, including structural changes, pathological variations, and artifacts and noise. It highlights the robust capability of our hard sample identification module in uncovering various forms of hard samples within the dataset.

### B.4. More results on binary segmentation

We have supplemented the corresponding standard deviations and statistical significance of Table 1 in the main text within Table 2. Our model demonstrates superior performance in five-fold cross-validation, with significantly higher DSC scores than other models, which is statistically significant. We further conducted a comparative analysis of our model against the fine-tuned SAM [6] and SAM2 [11]



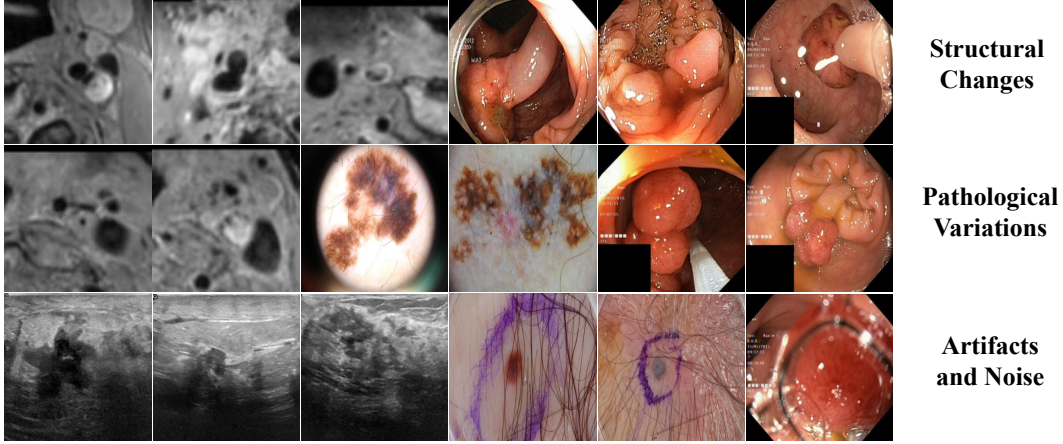


Figure 1. Instances identified as hard samples by our model in different datasets.

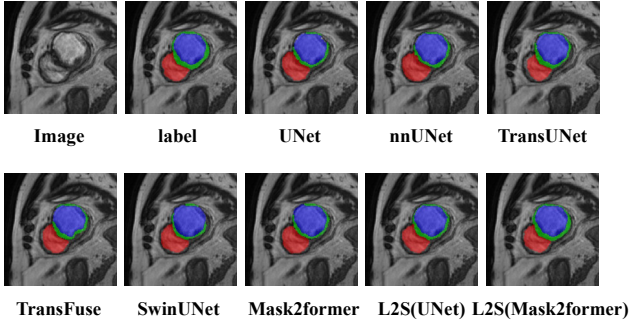


Figure 2. Qualitative results of cardiac organ segmentation on ACDC dataset.

on the ISIC2018 and Kvasir-SEG datasets. As shown in Table 3, despite fine-tuning, both SAM models underperform compared to our approach. SAM-based methods validate the scaling laws in image segmentation, but human-like data efficiency—learning from minimal samples—remains unachieved. We propose an architecture with improved data efficiency and elucidate its theoretical foundations.

### B.5. More results on multi-phase segmentation

We provide qualitative results on the ACDC, BTCV and BraTS2020 datasets for multi-phase segmentation using various methods, including our L2S. Figure 2 shows that the predicted segmentation results of our UNet-based and Mask2Former-based L2S models have a high overlap with the Ground Truth mask, especially in the Myocardium (Myo) region, while existing state-of-the-art methods show segmentation errors. Among these, our L2S (Mask2Former) achieves the best segmentation results. Figure 3 presents the qualitative results of different segmentation methods on the BTCV multi-organ dataset, where most methods face challenges in segmenting the Adrenal

gland (cyan-blue). The comparison results show that our L2S method performs excellently in segmenting this organ (see red rectangular box), while also achieving the best segmentation results for other organs. Furthermore, we validated the effectiveness of our model on the BraTS2020 brain tumor segmentation dataset. As shown in Figure 4, compared to the competing algorithms, our L2S method demonstrates outstanding performance in the enhancing tumor (red) region. Table 4 presents the accuracy metrics for the BraTS2020 brain tumor segmentation task, highlighting its superiority in challenging sample segmentation.

Methods	WT	TC	ET	ET*	Avg.
UNet	87.74	83.27	75.23	35.38	82.08
nnUNet	90.38	86.63	78.84	46.66	85.28
TransUNet	88.45	84.97	76.45	38.77	83.29
SwinUNet	88.71	85.18	76.52	40.21	83.47
TransFuses	90.11	86.44	78.62	45.04	85.06
TransBTS	90.62	86.86	79.61	47.12	85.70
MedSegDiffV2	89.23	85.94	77.28	42.23	84.15
Mask2former	88.84	85.26	77.15	39.47	83.75
L2S	<b>91.13</b>	<b>87.42</b>	<b>80.29</b>	<b>55.15</b>	<b>86.28</b>

Table 4. Dice score for brain tumor segmentation on the BraTS2020 dataset. ET\* denotes the mean value of enhancing tumor samples with dice score below 0.7.

### B.6. Visual comparison on ablation study

Finally, we performed a detailed visualization of the results from a comprehensive ablation study. Figure 5 illustrates representative visual outcomes, accompanied by standard deviations, to elucidate the impact of distinct model components. The first two examples (rows 1 and 2) demonstrate that the integration of CLS and DDPM significantly enhances the model’s segmentation accuracy. However, for

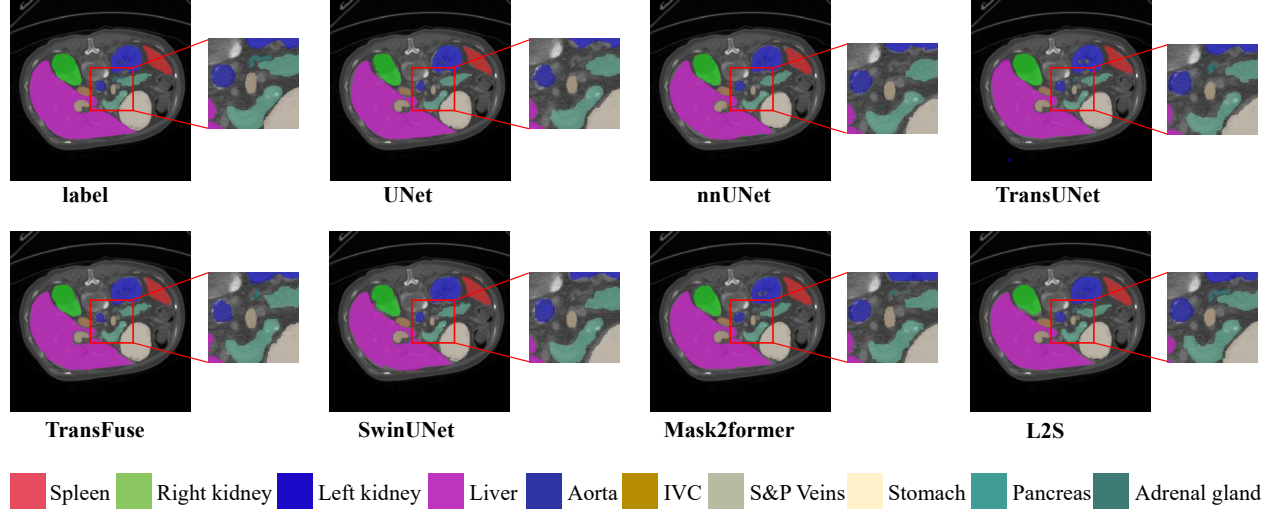


Figure 3. Qualitative results of multi-organ segmentation on BTCV dataset. The red rectangular box highlights incorrectly segmented organs by SOTA methods.

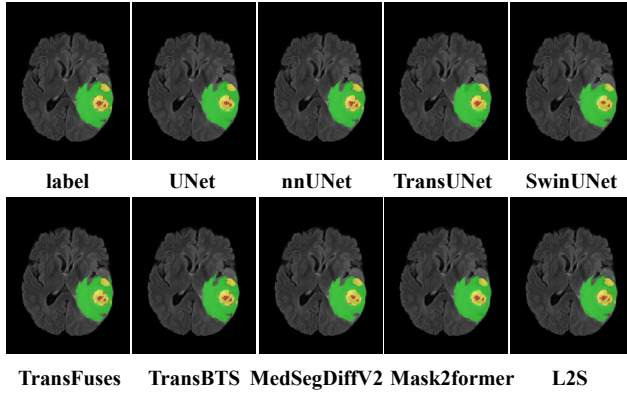


Figure 4. Qualitative results of brain tumor segmentation on BraTS2020 dataset.

images with complex edge structures (row 3), a notable disparity remains between the model’s segmentation predictions and the ground truth labels, highlighting a critical area for future research and refinement.

## References

- [1] Walid Al-Dhabyani, Mohammed Goma, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020. 3
- [2] Olivier Bernard, Alain Lalonde, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018. 3
- [3] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan

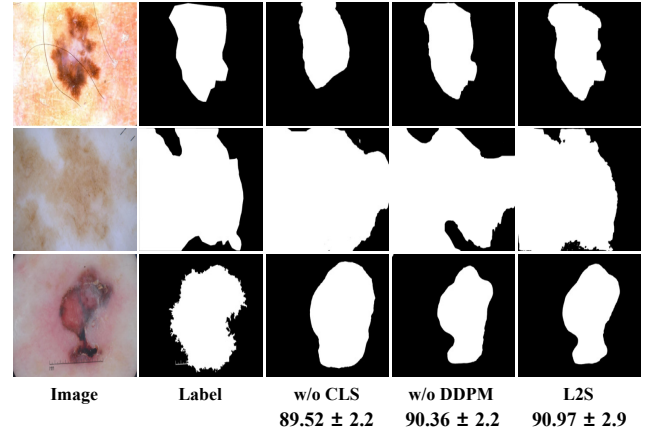


Figure 5. Ablation visualization of different modules for hard sample segmentation on ISIC2018.

Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 3

- [4] Xi Fang and Pingkun Yan. Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction. *IEEE Transactions on Medical Imaging*, 39(11):3619–3629, 2020. 3
- [5] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *MultiMedia modeling: 26th international conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, proceedings, part II 26*, pages 451–462. Springer, 2020. 3
- [6] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer White-

- head, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. [4](#)
- [7] Yucong Liu, Shixing Yu, and Tong Lin. Regularizing deep neural networks with stochastic estimators of hessian trace. *arXiv preprint arXiv:2208.05924*, 2022. [2](#)
- [8] Aleksandr Mikhailovich Lyapunov. The general problem of the stability of motion. *International journal of control*, 55(3):531–534, 1992. [2](#)
- [9] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014. [3](#)
- [10] Md Ashraful Alam Milton. Automated skin lesion classification using ensemble of deep neural networks in isic 2018: Skin lesion analysis towards melanoma detection challenge. *arXiv preprint arXiv:1901.10802*, 2019. [3](#)
- [11] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. [4](#)
- [12] Colin Wei, Sham Kakade, and Tengyu Ma. The implicit and explicit regularization effects of dropout. In *International conference on machine learning*, pages 10181–10192. PMLR, 2020. [1](#), [2](#)
- [13] Chenglu Zhu, Xiaoyan Wang, Shengyong Chen, Zhongzhao Teng, Cong Bai, Xiaojie Huang, Ming Xia, Zhanpeng Shao, Zheng Gu, and Peiliang Sun. Complex carotid artery segmentation in multi-contrast mr sequences by improved optimal surface graph cuts based on flow line learning. *Medical & Biological Engineering & Computing*, 60(9):2693–2706, 2022. [3](#)