HiEDL: Hierarchical Evidential Deep Learning for Uncertainty-Aware Tumor Segmentation from 3D CT via Boundary Regularization

Younghyun Park[1⋆], Jin Gyo Jeong[1⋆], and Sejung Yang[1,2†]

[1] Department of Precision Medicine, Yonsei University Wonju College of Medicine, Wonju, Korea
[2] Department of Medical Informatics and Biostatistics, Yonsei University Wonju College of Medicine, Wonju, Korea

Abstract. Accurate tumor segmentation in CT imaging is important for clinical diagnosis and treatment planning. However, achieving precise segmentation is still heavily challenging due to highly ambiguous boundaries. Also, conventional flat classification frameworks are not anatomically consistent, as they treat organs and tumors at the same semantic level, ignoring the fundamental anatomical prior that tumors physically reside within their corresponding organs. In this paper, we propose a novel Hierarchical Evidential Deep Learning framework (HiEDL) to accurately and efficiently segment tumors by explicitly enforcing this anatomical containment prior. Specifically, to restructure the segmentation task, we introduce a hierarchical organ-to-tumor decoupling strategy that divides the process into superordinate (organ) and subordinate (tumor) levels. Furthermore, we adopt evidential deep learning to isolate predictive tumor uncertainty and optimize the network using a novel hierarchical boundary loss combined with distance maps, directly guiding the model to resolve ambiguous boundary predictions. Experimental results demonstrate the superiority of HiEDL compared with other state-of-the-art 3D segmentation networks, by achieving the highest region-level (Dice) and boundary-level (NSD) scores on the MSD Pancreas and MSD Liver datasets. We further validate the effectiveness of the hierarchical formulation and the significant benefits of the proposed boundary loss through ablation studies. Our code is available at `https://github.com/laibi-yonsei/HiEDL`.

Keywords: Evidential Deep Learning · Uncertainty-Aware · Tumor Segmentation · Boundary Regularization.

1 Introduction

Automated segmentation of tumors from Computed Tomography (CT) scans plays a critical role in clinical workflows, providing essential anatomical information for computer-aided diagnosis, surgical planning, and radiotherapy. While

⋆ Younghyun Park and Jin Gyo Jeong contribute equally to this work.
† Corresponding email: `syang@yonsei.ac.kr`

highly accurate tumor delineation is crucial for successful clinical outcomes, current segmentation pipelines face significant challenges due to the low tissue contrast and high morphological variability of lesions [2]. Recent advances in representation learning, particularly with architectures like nnU-Net [9] and Swin-UNETR [7], have significantly improved segmentation precision. However, as highlighted in recent literature [5, 13], a major bottleneck impedes their practical adoption in safety-critical applications: model over-confidence. Conventional deterministic architectures fundamentally rely on Softmax probabilities, which often fail to capture predictive uncertainty. Consequently, they assign high confidence even to erroneous predictions in ill-defined boundary regions. Therefore, a main challenge is to learn a reliable pipeline that not only provides accurate anatomical segmentation but also quantifies its predictive uncertainty to warn clinicians of potential failures [10, 22].

To quantify uncertainty, Evidential Deep Learning (EDL) [1, 6, 17, 23] has emerged as a promising alternative to computationally expensive Bayesian approximations (e.g., Monte Carlo Dropout [4]). Grounded in Dempster-Shafer Theory, EDL formulates the network output as a Dirichlet distribution parameterized by accumulated evidence in a single forward pass. Despite its theoretical advantages, existing methods struggle to accurately delineate the critical transitions between the organ and the tumor. Standard frameworks often formulate segmentation as a *flat* classification problem, treating organs and tumors as mutually exclusive classes at the same semantic level [15, 21, 11]. This formulation fails to capture the hierarchical anatomical prior that a tumor is physically contained within its organ [20, 3, 18]. As a result, models typically resolve this by naively partitioning these regions based on deterministic confidence scores, inevitably leading to severe misclassifications at the borders. While introducing EDL mitigates this by quantifying the uncertainty at these margins [14], region-based evidential objectives alone remain insensitive to local boundary alignments [12]. Therefore, an additional regularization strategy is essential to prevent misclassifications in these highly uncertain areas.

To address these challenges, we propose a novel Hierarchical Evidential Deep Learning (HiEDL) framework for uncertainty-aware tumor segmentation from 3D CT scans. HiEDL explicitly integrates the anatomical hierarchy into the evidential learning process while adaptively regularizing ambiguous tumor boundaries. HiEDL has two main components as follows: (1) **Hierarchical evidential formulation.** To prioritize accurate tumor delineation, we decouple the segmentation into a hierarchical task. The superordinate level first localizes the encompassing organ as a structural constraint, allowing the subordinate level to exclusively isolate the target tumor. This explicitly enforces the anatomical containment prior. (2) **Uncertainty-aware boundary regularization.** To explicitly penalize over-confident misclassifications at ambiguous borders, we integrate a boundary loss based on Signed Distance Functions (SDF) into our framework. Coupled with evidential uncertainty, this surface regularization enables the network to accurately refine the challenging tumor boundaries.

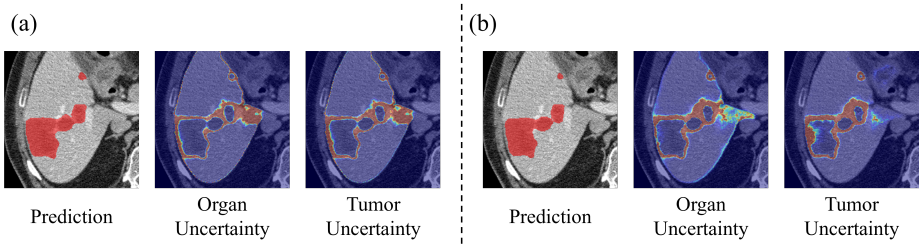

Fig. 1. (a) Flat classification completely ignores the prior that tumors physically reside within organs, causing severe errors. (b) Predictive tumor uncertainty is high at the inner organ-tumor boundary, rather than at the outer organ boundary.

2 Method

2.1 Evidential Deep Learning

Dirichlet Evidence Modeling. To address the aforementioned limitations of deterministic softmax probabilities, EDL [17] explicitly quantifies epistemic uncertainty by modeling the class probability vector as a random variable drawn from a Dirichlet distribution $\mathrm{Dir}(\mathbf{p} \mid \boldsymbol{\alpha})$. Specifically, instead of applying a standard softmax function, the segmentation network outputs voxel-wise logits that are transformed into non-negative evidence $\mathbf{e} = [e_1, \ldots, e_K]$ to quantify support for each of the K classes. The Dirichlet concentration parameters $\boldsymbol{\alpha}$ are then obtained via

$$\boldsymbol{\alpha} = \mathbf{e} + 1, \qquad S = \sum_{k=1}^{K} \alpha_k, \tag{1}$$

where S is the total evidence strength. The predictive class probability p_k and the evidential uncertainty u can then be computed as:

$$p_k = \frac{\alpha_k}{S}, \qquad u = \frac{K}{S}. \tag{2}$$

In EDL, the segmentation network outputs voxel-wise logits, which are transformed into non-negative evidence before forming $\boldsymbol{\alpha} = \mathbf{e} + 1$; the Dirichlet mean then yields the class probability map p_k, while u provides an uncertainty map.
Region-based Evidential Dice Objective. To connect EDL to dense prediction, predictions are indexed by voxel location $\mathbf{v} \in \Omega$. Li *et al.* [14] proposed optimizing region-level segmentation quality by minimizing the Bayes risk of a soft Dice objective under the voxel-wise Dirichlet posterior. This uncertainty-aware objective replaces point-estimate losses by explicitly penalizing both expected region mismatch and predictive variance.

Given the one-hot target $y_c(\mathbf{v})$ for class c, the objective admits a closed-form expression based on the first and second moments of the predicted probabilities under the Dirichlet distribution:

$$\mathbb{E}[p_c] = \frac{\alpha_c}{S}, \qquad \mathbb{E}[p_c^2] = \frac{\alpha_c(\alpha_c + 1)}{S(S+1)}. \tag{3}$$

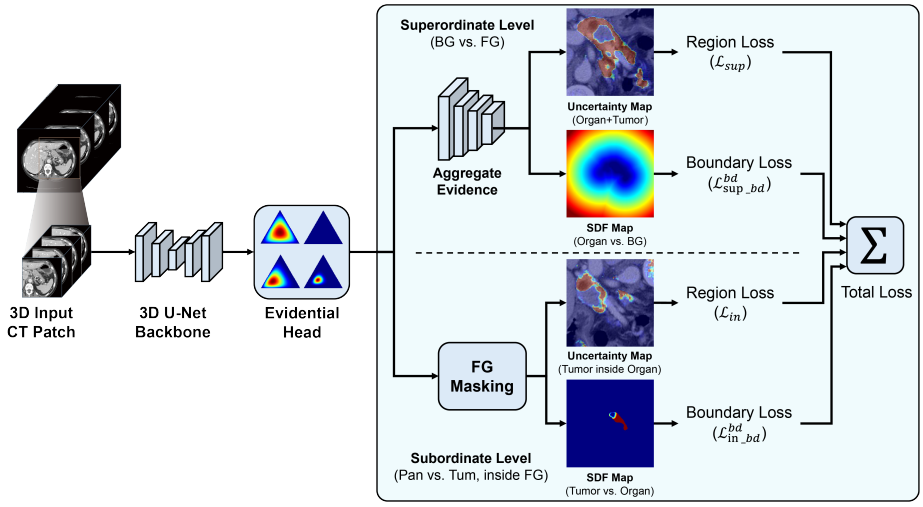

Fig. 2. Illustration of our HiEDL framework. HiEDL takes a 3D CT patch input and performs uncertainty-aware tumor segmentation by optimizing hierarchical region and boundary losses.

The evidential Dice loss is then defined as the expectation of the soft Dice score:

$$\mathcal{L}_{\text{Dice}} = \frac{1}{C}\sum_{c=1}^{C}\left(1 - \frac{2\sum_{\mathbf{v}\in\Omega} y_c(\mathbf{v})\,\mathbb{E}[p_c(\mathbf{v})]}{\sum_{\mathbf{v}\in\Omega} y_c(\mathbf{v})^2 + \sum_{\mathbf{v}\in\Omega}\mathbb{E}[p_c(\mathbf{v})^2]}\right). \tag{4}$$

2.2 HiEDL for Uncertainty-Aware Tumor Segmentation

Problem Formulation. Given an input CT volume $\mathbf{x} \in \mathbb{R}^{1\times H\times W\times D}$, we employ an evidential segmentation model that outputs a voxel-wise three-class Dirichlet prediction, where $p_k(\mathbf{v})$, $\boldsymbol{\alpha}(\mathbf{v})$, and $u(\mathbf{v})$ are derived as in Sec. 2.

Let $y(\mathbf{v}) \in \{0, 1, 2\}$ denote the ground-truth labels corresponding to the background, organ, and tumor. Anatomical structure provides a strong containment prior: tumors are physically confined within the organ of interest. We exploit this relationship by decomposing the three-class task into two hierarchical subproblems at the loss level: (i) background versus organ–tumor foreground, and (ii) discrimination between the organ and tumor within the foreground.

To this end, we define the following quantities, which are shared across all hierarchical objectives. The *superordinate* aggregation treats the organ and tumor as a unified foreground:

$$\boldsymbol{\alpha}_{\text{sup}}(\mathbf{v}) = \big[\alpha_0(\mathbf{v}),\, \alpha_1(\mathbf{v}) + \alpha_2(\mathbf{v})\big], \qquad y_{\text{sup}}(\mathbf{v}) = \mathbb{1}[y(\mathbf{v}) \neq 0]. \tag{5}$$

The *subordinate* aggregation isolates the organ–tumor channels within the foreground:

$$\boldsymbol{\alpha}_{\text{in}}(\mathbf{v}) = \big[\alpha_1(\mathbf{v}),\, \alpha_2(\mathbf{v})\big], \qquad y_c(\mathbf{v}) = \mathbb{1}[y(\mathbf{v}) = c], \quad c \in \{1, 2\}, \tag{6}$$

with the foreground mask

$$m_{\mathrm{fg}}(\mathbf{v}) = \mathbb{1}[y(\mathbf{v}) \neq 0], \tag{7}$$

and the foreground-normalized tumor score used for boundary regularization:

$$r_{\mathrm{tum}}(\mathbf{v}) = \frac{p_2(\mathbf{v})}{p_1(\mathbf{v}) + p_2(\mathbf{v}) + \epsilon}. \tag{8}$$

All hierarchical objectives in the following subsections are instantiated from these shared definitions.

Hierarchical Region Loss. Following the evidential Dice formulation in Eq. (4), we instantiate $\mathcal{L}_{\mathrm{Dice}}$ at two levels.

The *superordinate* term optimizes organ-centered localization over the full volume, written as

$$\mathcal{L}_{\mathrm{sup}} = \mathcal{L}_{\mathrm{Dice}}(\boldsymbol{\alpha}_{\mathrm{sup}},\, y_{\mathrm{sup}};\, m \equiv 1)\,. \tag{9}$$

The *subordinate* term refines organ–tumor discrimination strictly within the foreground:

$$\mathcal{L}_{\mathrm{in}} = \mathcal{L}_{\mathrm{Dice}}(\boldsymbol{\alpha}_{\mathrm{in}},\, \{y_1, y_2\};\, m_{\mathrm{fg}})\,. \tag{10}$$

If the current training sample contains no foreground voxels, we set $\mathcal{L}_{\mathrm{in}} = 0$.

Hierarchical Boundary Loss. While region losses capture volumetric overlap, they can be insensitive to small boundary misalignments. To explicitly regularize surfaces, we adopt a distance-map boundary loss [12] based on the SDF $\phi(\mathbf{v})$, which takes negative values inside the target region and positive values outside.

Given a predicted probability $p_{\mathrm{fg}}(\mathbf{v})$, the boundary loss is defined as:

$$\mathcal{L}_{\mathrm{bd}} = \int_{\Omega} p_{\mathrm{fg}}(\mathbf{v}) \cdot \phi(\mathbf{v})\, d\mathbf{v} \approx \frac{\sum_{\mathbf{v}} p_{\mathrm{fg}}(\mathbf{v})\, \phi(\mathbf{v})\, m(\mathbf{v})}{\sum_{\mathbf{v}} m(\mathbf{v}) + \epsilon}, \tag{11}$$

where Ω is the image domain, $\phi(\cdot)$ is precomputed from the ground-truth mask, and $m(\mathbf{v})$ is a spatial mask restricting the loss.

Using the shared definitions from Sec. 3.1, the superordinate boundary loss $\mathcal{L}_{\mathrm{sup}}^{\mathrm{bd}}$ is obtained by setting $p_{\mathrm{fg}} = p_1 + p_2$, $m \equiv 1$, and computing ϕ from y_{sup}. The subordinate boundary loss $\mathcal{L}_{\mathrm{in}}^{\mathrm{bd}}$ is computed by setting $p_{\mathrm{fg}} = r_{\mathrm{tum}}$, $m = m_{\mathrm{fg}}$, and deriving ϕ from y_2. Consequently, the overall training objective is formulated as:

$$\mathcal{L} = w_{\mathrm{sup}}\mathcal{L}_{\mathrm{sup}} + w_{\mathrm{in}}\mathcal{L}_{\mathrm{in}} + \lambda_{\mathrm{sup}}(t)\mathcal{L}_{\mathrm{sup}}^{\mathrm{bd}} + \lambda_{\mathrm{in}}(t)\mathcal{L}_{\mathrm{in}}^{\mathrm{bd}}. \tag{12}$$

where w_{sup} and w_{in} are constant weight hyperparameters for the region losses, and $\lambda(t)$ is a time-dependent ramp warmup schedule for the boundary losses.

3 Experiments

3.1 Datasets and Implementation Details

Dataset. We evaluated our proposed HiEDL framework on two publicly available CT tumor segmentation datasets: the Medical Segmentation Decathlon

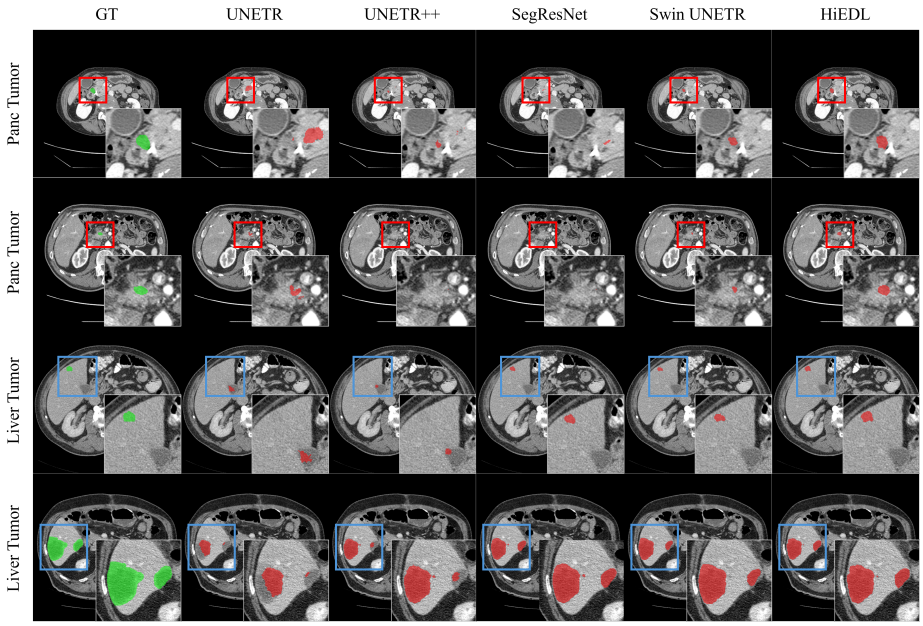

Fig. 3. Visual comparison of different methods for abdominal tumor segmentation. Green and red regions denote the GT and tumor, respectively.

(MSD) Pancreas and the MSD Liver [2].The MSD Pancreas dataset focuses on the segmentation of pancreatic tumors, while the MSD Liver dataset involves the segmentation of associated hepatic tumors.

Pre-processing. To standardize the diverse spatial resolutions and intensity distributions across different patients and scanners, we applied a systematic preprocessing pipeline. First, voxel intensities were truncated to a soft-tissue Hounsfield Unit (HU) window of $[-125, 225]$ across all datasets to enhance contrast, followed by normalization to the $[0, 1]$ range. Subsequently, all 3D volumes were resampled to a uniform physical spacing of $1.5 \times 1.5 \times 2.0$ mm^3. During training, 3D sub-volumes of size $256 \times 256 \times 32$ were randomly cropped to serve as network inputs.

Training setup. All networks were trained using the AdamW optimizer for 200 epochs, with a batch size of 1 and an initial learning rate of 2×10^{-4}. Experiments were conducted on a single NVIDIA GeForce RTX A6000 GPU with 48GB of RAM. All models were implemented in Python 3 using the PyTorch framework and executed within identical computing environments to ensure a fair comparison. We set w_{sup} and w_{in} to 1, λ_{sup} and λ_{in} to 0.2, and $t = 20$.

Evaluation metrics. To quantitatively evaluate the segmentation performance, we employed two standard metrics: the Dice Similarity Coefficient (Dice) to measure volumetric overlap, and the Normalized Surface Distance (NSD) to assess the accuracy of boundary alignment.

Table 1. Performance comparison of HiEDL with existing 3D medical image segmentation networks.

Model	MSD Pancreas		MSD Liver	
	Dice ↑	NSD ↑	Dice ↑	NSD ↑
UNETR	0.2035	0.3261	0.3090	0.2745
UNETR++	0.3213	0.4669	0.5002	0.4474
SegResNet	0.3990	0.5460	0.6287	0.6682
Swin UNETR	0.4073	0.5811	0.6218	0.6471
HiEDL	**0.4346**	**0.5906**	**0.6542**	**0.6930**

Table 2. Ablation study results of different components in HiEDL.

(a) Ablation study on each component in MSD Pancreas		
Components	Dice ↑	NSD ↑
Baseline	0.4073	0.5811
Baseline w/ Region loss	0.3965	0.5574
Baseline w/ Boundary loss (Ours)	**0.4346**	**0.5906**
(b) Ablation study on each component in MSD Liver		
Components	Dice ↑	NSD ↑
Baseline	0.6218	0.6471
Baseline w/ Region loss	0.5930	0.5758
Baseline w/ Boundary loss (Ours)	**0.6542**	**0.6930**

3.2 Experimental Results

Comparison with other methods. We compare HiEDL with four existing 3D medical image segmentation networks [7, 8, 19, 16] for tumor segmentation. Compared with other methods, our HiEDL achieves the best performance in all metrics across both datasets (Table 1). Specifically, our HiEDL surpasses the second-highest performing models (Swin UNETR for the pancreas dataset and SegResNet for the liver dataset) by obtaining 0.4346 and 0.5906 for Dice and NSD on the MSD Pancreas dataset, and 0.6542 and 0.6930 for Dice and NSD on the MSD Liver dataset, respectively. As shown in Fig. 3, the segmentation results of HiEDL are relatively close to the ground truth annotations, especially for the highly ambiguous morphological boundaries between the tumors and their corresponding organs. Compared with the previous state-of-the-art methods, HiEDL shows superior performance in terms of both region-level (Dice) and boundary-level (NSD) accuracy.

Component analysis. We perform ablation studies to verify the effectiveness of each loss component in our proposed HiEDL framework. First, we compare the effectiveness of each component configuration, including (1) the Baseline, (2) the Baseline with region loss, and (3) the Baseline with boundary loss (Ours). Tables 2(a) and 2(b) present the evaluation results on the MSD Pancreas and MSD Liver datasets, respectively. Interestingly, applying only the region-based loss degrades the segmentation performance compared to the Baseline, dropping the Dice score from 0.4073 to 0.3965 on the pancreas dataset, and from 0.6218 to 0.5930 on the

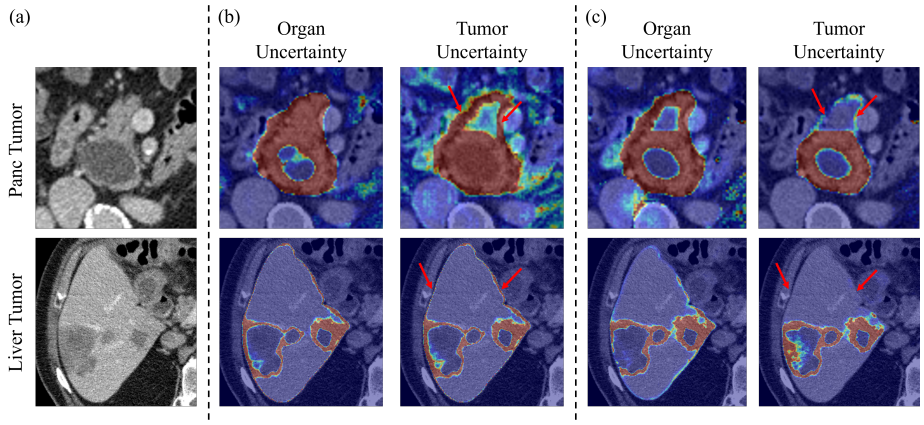

Fig. 4. Predictive uncertainty maps from the ablation study. Red arrows indicate that our proposed boundary loss effectively reduces the uncertainty outside the organs in the tumor's uncertainty maps.

liver dataset. This indicates that relying solely on region-level supervision struggles to delineate the highly ambiguous boundaries between tumors and their corresponding organs. Conversely, our proposed method effectively overcomes this limitation. By explicitly optimizing the boundary constraints, our approach significantly outperforms the other configurations, achieving the highest Dice scores of 0.4346 and 0.6542, and the highest NSD scores of 0.5906 and 0.6930 on the pancreas and liver datasets, respectively. This boundary-level improvement is further supported by the uncertainty maps in Fig.4. While standard EDL distributes high predictive uncertainty across irrelevant background regions, our hierarchical formulation strictly confines the tumor uncertainty within the organ. Consequently, HiEDL effectively redirects the network's focus to accurately resolve the ambiguous organ-tumor interfaces.

4 Conclusion

In this study, we present HiEDL, a hierarchical evidential deep learning framework for uncertainty-aware tumor segmentation. By modeling the anatomical prior that tumors reside within organs, we decouple the task into a hierarchical organ-to-tumor strategy. Experiments demonstrate this formulation, along with our uncertainty-aware boundary loss, successfully suppresses spurious uncertainties and resolves ambiguous boundaries. Consequently, HiEDL achieves superior region-level and boundary-level accuracy compared to existing state-of-the-art 3D medical image segmentation networks.

Acknowledgments. This work was supported by the Regional Innovation System & Education (RISE) program through the Gangwon RISE Center, funded by the Ministry of Education (MOE) and Gangwon State, Republic

of Korea (Grant No. 2026-RISE-10-006); the Bio & Medical Technology Development Program of the National Research Foundation of Korea (NRF), funded by the Korean government (MSIT) (Grant No. RS-2024-00440802); and an NRF grant funded by the Korean government (MSIT) (Grant No. 2022R1A2C2091160).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Amini, A., Schwarting, L., Soleimany, A., Rus, D.: Deep evidential regression. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 33, pp. 14927–14937 (2020)
2. Antonelli, M., et al.: The medical segmentation decathlon. Nature Communications **13**(1), 4128 (2022)
3. Becker, S., Baldini, G., Schmidt, C., Pollok, O., Pelka, O., Kohnke, J., Borys, K., Friedrich, C., Schaarschmidt, B., Forsting, M., Umutlu, L., Haubold, J., Nensa, F., Hosch, R.: SALT: Introducing a framework for hierarchical segmentations in medical imaging using label trees. Scientific Reports **15**(1) (2025)
4. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning (ICML). pp. 1050–1059. PMLR (2016)
5. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.: On calibration of modern neural networks. In: International Conference on Machine Learning (ICML). pp. 1321–1330. PMLR (2017)
6. Han, Z., Zhang, C., Fu, H., Zhou, J.: Trusted multi-view classification with dynamic evidential fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(2), 2551–2566 (2022)
7. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H., Xu, D.: Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In: International MICCAI Brainlesion Workshop. pp. 272–284. Springer (2022)
8. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 574–584 (2022)
9. Isensee, F., Jaeger, P., Kohl, S., Petersen, J., Maier-Hein, K.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
10. Jungo, A., Balsiger, F., Reyes, M.: Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. Frontiers in Neuroscience **14**, 282 (2020)
11. Karimi, D., Gholipour, A.: Improving calibration and out-of-distribution detection in deep models for medical image segmentation. IEEE Transactions on Artificial Intelligence **4**(2), 383–397 (2023)
12. Kervadec, H., Bouchbout, J., Piantoni, G., Benali, I., Rousseau, J., Ayed, I.: Boundary loss for highly unbalanced segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. pp. 285–296. Springer (2019)

13. Kompa, B., Snoek, J., Beam, A.: Second opinion needed: communicating uncertainty in medical machine learning. npj Digital Medicine **4**(1), 4 (2021)
14. Li, H., Nan, Y., Del Ser, J., Yang, G.: Region-based evidential deep learning to quantify uncertainty and improve robustness of brain tumor segmentation. Neural Computing and Applications **35**(30), 22071–22085 (2023)
15. Mehrtash, A., Wells, W., Tempany, C., Abolmaesumi, P., Kapur, T.: Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. IEEE Transactions on Medical Imaging **39**(12), 3868–3878 (2020)
16. Myronenko, A.: 3d MRI brain tumor segmentation using autoencoder regularization. In: International MICCAI Brainlesion Workshop. pp. 311–320. Springer (2018)
17. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 31 (2018)
18. Seo, H., Huang, C., Bassenne, M., Xiao, R., Xing, L.: Modified U-Net (mu-net) with incorporation of object-dependent high level features for improved liver and liver-tumor segmentation in CT images. IEEE Transactions on Medical Imaging **39**(5), 1316–1325 (2020)
19. Shaker, A., Maaz, M., Rashwan, H., Salman, S., Al-Garadi, M., Khan, F.: Unetr++: Delving into inter-and intra-decoding for 3d medical image segmentation. arXiv preprint arXiv:2212.04497 (2022)
20. Tang, Y., Gao, R., Lee, H., Han, S., Chen, Y., Gao, D., Nath, V., Bermudez, C., Savona, M., Abramson, R., Bao, S., Lyu, I., Huo, Y., Landman, B.: High-resolution 3D abdominal segmentation with random patch network fusion. Medical Image Analysis **69**, 101894 (2021)
21. Yeung, M., Rundo, L., Nan, Y., Sala, E., Schönlieb, C., Yang, G.: Calibrating the dice loss to handle neural network overconfidence for biomedical image segmentation. Journal of Digital Imaging **36**(2), 739–752 (2022)
22. Zenk, M., Zimmerer, D., Isensee, F., Traub, J., Norajitra, T., Jäger, P., Maier-Hein, K.: Comparative benchmarking of failure detection methods in medical image segmentation: Unveiling the role of confidence aggregation. Medical Image Analysis **101**, 103392 (2025)
23. Zhang, Y., et al.: DEviS: Toward reliable medical image segmentation by modeling evidential calibrated uncertainty. IEEE Transactions on Cybernetics **55**(12), 5975–5988 (2025)