# Uncertainty-Supervised Interpretable and Robust Evidential Segmentation

Yuzhu Li[1,2†], An Sui[1†], Fuping Wu[3✉], and Xiahai Zhuang[1✉]

[1] School of Data Science, Fudan University, Shanghai, China
[2] Institute of Science and Technology for Brain-inspired intelligence, Fudan University, Shanghai, China
[3] Nuffield Department of Population Health, University of Oxford, Oxford, UK
zxh@fudan.edu.cn
https://zmiclab.github.io/

**Abstract.** Uncertainty estimation has been widely studied in medical image segmentation as a tool to provide reliability, particularly in deep learning approaches. However, previous methods generally lack effective supervision in uncertainty estimation, leading to low interpretability and robustness of the predictions. In this work, we propose a self-supervised approach to guide the learning of uncertainty. Specifically, we introduce three principles about the relationships between the uncertainty and the image gradients around boundaries and noise. Based on these principles, two uncertainty supervision losses are designed. These losses enhance the alignment between model predictions and human interpretation. Accordingly, we introduce novel quantitative metrics for evaluating the interpretability and robustness of uncertainty. Experimental results demonstrate that compared to state-of-the-art approaches, the proposed method can achieve competitive segmentation performance and superior results in out-of-distribution (OOD) scenarios while significantly improving the interpretability and robustness of uncertainty estimation. Code is available via https://github.com/suiannaius/SURE.

**Keywords:** Uncertainty supervision · Interpretability · Robustness.

## 1 Introduction

Accurate medical image segmentation is essential for clinical applications such as diagnosis [1] and treatment planning [4]. Beyond accuracy, reliability, interpretability, and robustness have raised increasing concerns for researchers and clinicians. Recent advances in deep learning, particularly U-Net [6] and its variants [16,9,3], have significantly improved segmentation accuracy. However, these models often neglect uncertainty in ambiguous regions like low-contrast or noisy areas, leading to over-confident predictions and errors. The absence of uncertainty further limits access to reliable predictions, hindering practical utility.

† These two authors contributed equally.
✉ Xiahai Zhuang and Fuping Wu are the corresponding authors.

To address these issues, uncertainty estimation methods, such as Bayesian approaches [7], ensemble strategies [17], test time augmentation (TTA) [23], and evidential deep learning (EDL) [10], have emerged. While Bayesian methods like Monte Carlo Dropout [25] are computationally costly, ensemble-based techniques [18] require training multiple models, and TTA depends heavily on augmentations, EDL [21] offers a computationally efficient and theoretically sound solution. Based on Dempster-Shafer theory [22] and subjective logic [12], EDL integrates uncertainty directly into the model, enabling reliable estimations with a single forward pass. While these models can provide pixel-level confidence [20], they often fail to explain the underlying mechanism for uncertainty or maintain robustness under noise perturbations.

In this study, we propose a self-supervised approach to enhance the uncertainty interpretability and robustness against noise based on EDL. Different from uncertainty calibration, differentiating the inaccurate predictions from the accurate [26], we propose three principles requiring uncertainty estimation conforming to human beings' thinking or reasoning patterns. Based on these principles, we design supervision losses accordingly, leading to our novel uncertainty supervision learning framework for medical image segmentation.

The contributions of this work are summarized as follows: (1) We introduce an uncertainty supervision approach to enhance the interpretability and robustness of evidential learning, by regularizing the relationships of uncertainty with gradients of boundaries and noise; (2) We introduced new quantitative metrics for interpretability and noise robustness of uncertainty; (3) Experimental results show that the proposed method aligns with human logic and demonstrate enhanced robustness against noise in Out-Of-Distribution (OOD) cases.

## 2 Methods

As illustrated in Fig. 1, our framework comprises three parts:(1) For segmentation prediction and uncertainty estimation, we employ EDL to generate class-specific evidence for input images, as detailed in Section 2.1. (2) For uncertainty supervision, we introduce human-inspired gradient-based supervision loss to enhance its interpretability in Section 2.2, and (3) we design novel noise-based supervision loss to improve both interpretability and robustness in Section 2.3.

### 2.1 Evidence-Based Prediction Generation

Let $\boldsymbol{X} = (\boldsymbol{x}_i) \in \mathbb{R}^V$ and $\boldsymbol{Y} = (\boldsymbol{y}_i) \in \mathbb{R}^{V \times K}$ respectively denote a 2D slice and its label, where $V$ is the number of pixels, $K$ is the number of classes. The evidential network $f_\theta$ estimates the evidence map as $\boldsymbol{E} = f_\theta(\boldsymbol{X}) = (\boldsymbol{e}_i) \in \mathbb{R}^{V \times K}$. According to subjective logic [12], for the $i$-th pixel in $\boldsymbol{X}$, its categorical probability variable $\boldsymbol{p}_i = (p_{ik}) \in [0,1]^K$ can be modeled as the Dirichlet distribution. The uncertainty of the $i$-th pixel can be derived as $u_i = \frac{K}{\sum_{j=1}^{K}(\boldsymbol{e}_{ij}+1)}$.

For supervised learning, we adopt Dice and cross-entropy loss function, denoted by $\mathcal{L}_{Dice}(\boldsymbol{P}, \boldsymbol{Y})$ and $\mathcal{L}_{CE}(\boldsymbol{P}, \boldsymbol{Y})$, proposed in [27], and the Kullback-Leibler divergence, *i.e.,* $\mathcal{L}_{KL}(\boldsymbol{P})$, proposed in [21] to avoid collecting evidence

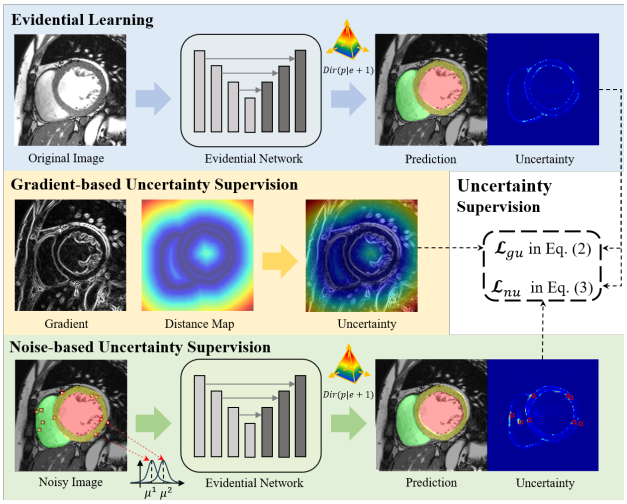



**Fig. 1.** The overview of our work. Based on EDL, our model estimates uncertainty from evidence, with gradient-based and noise-based supervision to enhance both interpretability and robustness of uncertainty.

about the incorrect classes, where $\boldsymbol{P} = (\boldsymbol{p}_i) \in \mathbb{R}^{V \times K}$. The overall EDL segmentation loss is denoted as:

$$\mathcal{L}_{Seg} = \lambda_{CE} \cdot \mathcal{L}_{CE} + \lambda_{Dice} \cdot \mathcal{L}_{Dice} + \lambda_{KL} \cdot \mathcal{L}_{KL}. \tag{1}$$

Although evidential learning offers a computationally efficient framework for uncertainty estimation, it lacks supervision regarding the underlying originality, resulting in limited interpretability and robustness. To address this, we propose two uncertainty supervision techniques: gradient-based supervision and noise-based supervision.

### 2.2 Gradient-Based Supervision

In regions near the boundaries, where uncertainty tends to be high, previous models fail to explicitly reveal the factors contributing to the uncertainty values. As a result, understanding the nature and origin of uncertainty in decision-making processes becomes challenging. To this end, we propose the following principle to align uncertainty estimation with human intuition:

**Principle 1.** *For clear boundaries, higher gradients yield lower uncertainty, whereas ambiguous boundaries with lower gradients have higher uncertainty.*

To achieve this principle, we introduce a gradient-based uncertainty supervision loss on boundary pixels, defined as:

$$\mathcal{L}_{gu} = \frac{1}{|B|} \sum_{i,j \in B, i \neq j} \max(0, (u_i - u_j)(g_i - g_j)), \tag{2}$$

where $B$ represents the set of boundary pixels, and $g_i$ denotes the gradient of pixel $i$, which is computed on a Gaussian smoothed image for stability. This formulation ensures that uncertainty estimations are aligned with human perceptual logic, thus enhancing their interpretability.

### 2.3 Noise-Based Supervision

Uncertainty is intrinsically related to both the noise and the distance of the pixel from the boundary. The following principles describe the relationships:

**Principle 2.** *When a pixel is close to the boundary, a larger noise amplitude leads to higher uncertainty, and vice versa.*

**Principle 3.** *When a pixel is sufficiently far from the boundary, the uncertainty becomes negligible, regardless of the noise amplitudes.*

To capture the relationships described in **Principle 2** and **Principle 3**, we formulate the noise supervision loss function $\mathcal{L}_{nu}$ as follows for more interpretable and robust uncertainty estimation,

$$\mathcal{L}_{nu} = \sum_{i \in S} \underbrace{\mathbb{1}_{d_i \leq d_0} \cdot \max(0, -(\mu^2 - \mu^1)(u_i^2 - u_i^1))}_{\textit{nearby noise} \rightarrow \textit{interpretability}} + \underbrace{\mathbb{1}_{d_i > d_0} \cdot (u_i^0 + u_i^1 + u_i^2)}_{\textit{remote noise} \rightarrow \textit{robustness}}, \tag{3}$$

where $S$ denotes the sampled pixel set, $d_i$ represents the distance of the $i$-th pixel to the boundary, $d_0$ is a threshold, $\mu^1$ and $\mu^2$ denote mean values of two different normal distributions for noise sampling, $u_i^1$ and $u_i^2$ denote the corresponding uncertainty of the $i$-th pixel after respectively applying the two noises, and $u_i^0$ represents its uncertainty without noise. The first term in the right side of Eq.(3) is for interpretability enhancement, and the second term improves the robustness by constraining the uncertainty of pixels with distance larger than $d_0$.

Due to the large number of boundary points, we utilize an active learning strategy that selectively focuses on the most informative data points, *i.e.,* hard samples, to improve training efficiency, instead of using all pixels in Eq. (3).

**Hard Samples Detection** Inspired by multi-class active learning techniques [13], we first identify hard samples to guide the model's learning process, ensuring it focuses on more challenging instances. Specifically, we impose noise to the entire training images and feed them into the model to obtain noised uncertainty $u^1$. According to **Principle 2**, $u^1$ should be larger than the original

uncertainty $u^0$ for all pixels. We define hard samples as those pixels not meeting this condition, namely $S^{hard} = \{i | u_i^1 \leq u_i^0\}$.

To mitigate class imbalance, we adopt a class-wise sampling strategy, which equally samples pixels from the previously identified hard samples for each class. Thus, the model is encouraged to allocate balanced attention to each class region during training, thereby improving its ability to handle underrepresented classes.

### 2.4 Total Loss

In general, the total loss consists of three terms, *i.e.*, the segmentation loss from EDL given by Eq. (1), and the uncertainty supervision losses outlined in Eqs. (2) and (3),

$$\mathcal{L}_{total} = \mathcal{L}_{Seg} + \beta \cdot \mathcal{L}_{gu} + \gamma \cdot \mathcal{L}_{nu}. \tag{4}$$

## 3 Experiments

### 3.1 Dataset and Experiment Setting

We validated the proposed method with two datasets: (1) The **Automated Cardiac Diagnosis Challenge (ACDC)** dataset contains 200 annotated short-axis cardiac MR-cine images from 100 patients [2]. All slices were cropped to a size of 96 × 96. (2) The **REFUGE** dataset includes 400 color fundus photography (CFP) images for training and an additional 400 images for testing [19]. Each image was annotated with optic cup (OC) and optic disc (OD) labels. All images were cropped into 512 × 512.

**Implementation Details:** We employed U-Net as the backbone, and the Adam optimizer with a learning rate of 0.001. The batch size was set to 24 for ACDC and 8 for REFUGE (reduced to 1 for the PU [5] method due to GPU memory constraint). For hyper-parameters, we set $\lambda_{CE} = 1, \lambda_{KL} = min(1, t/20)$. We set $\lambda_{Dice} = 1 - \alpha$ and $\beta = 0.1\alpha, \gamma = 10\alpha$ for ACDC, and $\beta = \gamma = \alpha$ for REFUGE, where the annealing factor $\alpha = \alpha_0 e^{\{-(\mathrm{In}\alpha_0/T)t\}}$. $T$ and $t$ were the total epochs and the current epoch, respectively, with $\alpha_0 = 0.01$. The boundary set $B$ included pixels with $d \leq 1$. For noise supervision, we set $d_0 = 4$. All experiments were implemented on an NVIDIA Geforce RTX 2080Ti GPU.

### 3.2 Evaluation Metrics for Principle 1 and Principle 2

For quantitative evaluation of uncertainty, the conventional metrics such as Expected Calibration Error (ECE) and Uncertainty-Error Overlap (UEO) [8,14] can not measure the interpretable factors proposed in **Principle 1** and **Principle 2**. Therefore, we introduce two sets of new metrics, including Uncertainty Correlation Coefficient (UCC) and Uncertainty Ratio (UR), to quantify the interpretability of uncertainty estimations.

For **Principle 1**, UCC is defined as the Spearman correlation coefficients [24] between image gradients and the uncertainty estimations, denoted by $UCC_{[g]} =$

**Table 1.** Comparisons with various uncertainty estimation methods on the ACDC and REFUGE dataset. The **bold** indicates the best result in a column, and the underlined indicates the second best. (✓) and (×) in UCC columns denote the same and opposite signs as expected, respectively. The unit of HD95 is mm for ACDC, pixels for REFUGE.

| Methods | DSC↑ (%) | HD95↓ (mm/pixels) | UEO↑ | ECE↓ | UCC | | UR↑ | |
|---|---|---|---|---|---|---|---|---|
| | | | | | $g(-)$ | $\mu(+)$ | $g$ | $\mu$ |
| **ACDC** | | | | | | | | |
| **Ours** | **91.06** | 8.45 | 0.222 | 0.009 | -0.427(✓) | 0.170(✓) | **0.632** | 0.585 |
| **DEviS** | 90.36 | 7.42 | 0.269 | **0.007** | 0.109(×) | -0.002(×) | 0.446 | 0.500 |
| **PU** | 87.35 | 9.45 | 0.177 | 0.011 | 0.162(×) | 0.602(✓) | 0.430 | 0.801 |
| **EU** | 88.14 | **7.11** | 0.246 | 0.008 | 0.180(×) | 0.021(✓) | 0.424 | 0.511 |
| **UDrop** | 88.16 | 7.77 | 0.149 | 0.276 | -0.022(✓) | 0.636(✓) | 0.630 | **0.818** |
| **TTA** | 73.44 | 37.7 | **0.277** | 0.025 | -0.036(✓) | -0.093(×) | 0.497 | 0.453 |
| **REFUGE** | | | | | | | | |
| **Ours** | **84.46** | 56.35 | 0.275 | **0.024** | -0.056(✓) | 0.064(✓) | **0.519** | 0.532 |
| **DEviS** | 83.05 | 65.39 | 0.359 | 0.065 | 0.043(×) | 0.150(✓) | 0.486 | 0.575 |
| **PU** | 79.01 | 117.7 | **0.384** | 0.035 | -0.044(✓) | 0.106(✓) | 0.515 | 0.553 |
| **EU** | 83.60 | **56.14** | 0.160 | 0.037 | -0.034(✓) | 0.078(✓) | 0.512 | 0.539 |
| **UDrop** | 73.60 | 65.23 | 0.117 | 0.277 | 0.016(×) | 0.110(✓) | 0.497 | 0.555 |
| **TTA** | 75.34 | 98.99 | 0.305 | 0.051 | -0.052(✓) | 0.155(✓) | 0.517 | **0.578** |

$SCorr(g, u)$ for boundary pixels $B$. Similarly, for **Principle 2**, we define $UCC_{[\mu]} = SCorr(\mu, u)$ using pixels with $d \leq d_0$. Take $UCC_{[g]}$ as an example,

$$UCC_{[g]} = \frac{\sum_{i\in B}(R(g_i) - \overline{R(g)})(R(u_i) - \overline{R(u)})}{\sqrt{\sum_{i\in B}(R(g_i) - \overline{R(g)})^2 \sum_{i\in B}(R(u_i) - \overline{R(u)})^2}}. \tag{5}$$

$R(\cdot)$ denotes the ranking function, $\overline{(\cdot)}$ denotes the mean values.

The UCC value ranges from [-1, 1], where the sign indicates the direction of correlation (positive or negative), and the magnitude reflects its strength. An interpretable uncertainty should satisfy $UCC_{[g]} < 0$ and $UCC_{[\mu]} > 0$.

Alternatively, UR calculates the ratio of pixel pairs satisfying the relationships between image gradients (noise) and the uncertainty estimations. Thus for **Principle 1**, we define

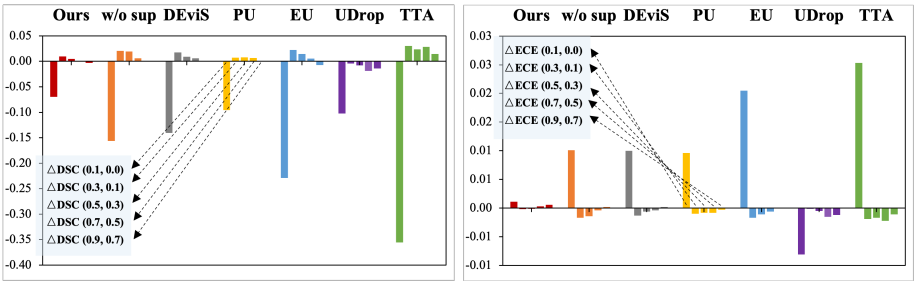
$$UR_{[g]} = \frac{\sum_{i,j\in B, i\neq j} \mathbb{1}_{((g_i - g_j)(u_i - u_j) \leq 0)}}{\sum_{i,j\in B} \mathbb{1}_{(i\neq j)}}. \tag{6}$$

Similarly, we have $UR_{[\mu]}$ for **Principle 2** defined in pixels with $d \leq d_0$.

For segmentation accuracy, we adopted the Dice Similarity Coefficient (DSC) and the 95% Hausdorff Distance (HD95)[11] as metrics.

**Table 2.** Ablation study on the REFUGE dataset.

| $\mathcal{L}_{gu}$ | $\mathcal{L}_{nu}$ | HSD | DSC↑ (%) | HD95↓ (pixels) | UEO↑ | ECE↓ | UCC $g(-)$ | UCC $\mu(+)$ | UR↑ $g$ | UR↑ $\mu$ |
|---|---|---|---|---|---|---|---|---|---|---|
| ✓ | ✓ | ✓ | 84.46 | 56.35 | 0.275 | 0.024 | -0.056(✓) | 0.064(✓) | 0.519 | **0.532** |
| × | ✓ | ✓ | <u>84.51</u> | <u>41.27</u> | **0.338** | <u>0.017</u> | -0.034(✓) | -0.013(×) | 0.511 | 0.493 |
| ✓ | × | ✓ | **84.65** | 42.53 | 0.286 | 0.019 | -0.064(✓) | -0.073(×) | <u>0.522</u> | 0.464 |
| ✓ | ✓ | × | 83.24 | **40.75** | 0.319 | **0.016** | -0.075(✓) | 0.009(✓) | **0.525** | <u>0.504</u> |
| × | × | × | 83.96 | 58.40 | <u>0.337</u> | 0.022 | -0.029(✓) | -0.033(×) | 0.510 | 0.484 |




(a) DSC change. (b) ECE change.

**Fig. 2.** Robustness evaluation on ACDC dataset, different colors represent different methods. The values are the difference of scores under two different noise levels, *e.g.*, $\triangle DSC(0.3, 0.1) = DSC(\mu = 0.3) - DSC(\mu = 0.1)$.

### 3.3 Experiment Results

**Comparison Study** We evaluated the proposed method by comparing with various uncertainty estimation approaches, including EDL-based DEviS [26], Variational Inference-based Probabilistic U-Net (PU) [5], Deep Ensemble-based EU [17], Dropout-based UDrop [15], and Test-Time Augmentation (TTA) [23].

Table 1 presents the quantitative results on ACDC and REFUGE: (1) For **segmentation accuracy**, our method achieved the best DSC scores and competitive HD95 values on both datasets. (2) For the conventional **uncertainty evaluation (UEO, ECE)**, our method obtained competitive results on both datasets. Particularly on REFUGE dataset, we achieved the best ECE value. Note that UEO measures a strong correlation between uncertainty and error, which might not match our principles to some extent. To enhance the interpretability of uncertainty with the new uncertainty supervision losses, the performance in UEO can be traded off. (3) For **uncertainty interpretability (UCC, UR)**, our method demonstrated significant superiority, as evidenced by the consistent signs of UCC and their values in the tables. Moreover, one can see that for other compared methods, none of them obtained the right signs for $UCC_{[g]}$ and $UCC_{[\mu]}$ on both datasets completely. Specifically, DEviS only had the right sign of $UCC_{[\mu]}$ on REFUGE, both PU and EU obtained the wrong

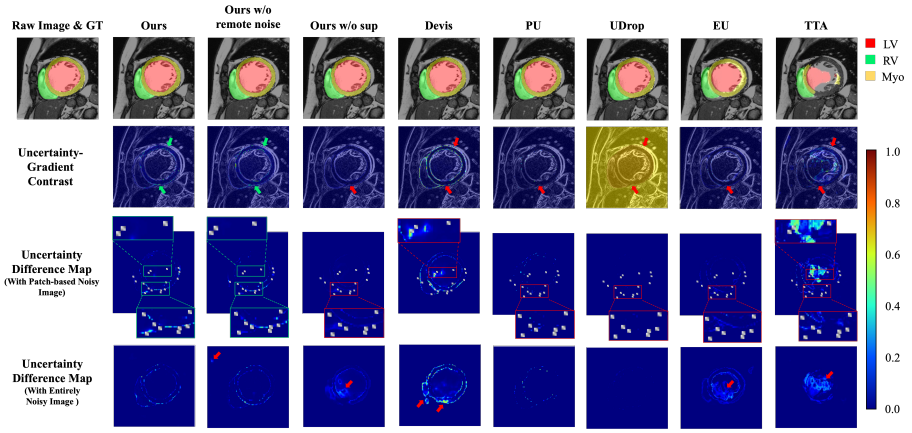



**Fig. 3.** Illustration of prediction results and uncertainty maps of different methods. Red arrows/boxes highlight erroneous uncertainty estimations, while green indicates correct ones.

sign of $UCC_{[g]}$ on ACDC, UDrop and TTA respectively got the wrong sign of $UCC_{[g]}$ on REFUGE and $UCC_{[\mu]}$ on ACDC.

**Ablation Study** We analyzed the effectiveness of three techniques adopted in the proposed method, including (1) gradient supervision loss $\mathcal{L}_{gu}$ (Eqs.(2)), (2) noise supervision loss $\mathcal{L}_{nu}$ (Eqs.(3)), and (3) hard sample detection (HSD) (Sec.2.3). As showed in Table 2 , without either $\mathcal{L}_{gu}$ or $\mathcal{L}_{nu}$, our model delivered an opposite sign of $UCC_{[\mu]}$, although other metrics were slightly improved. Note that HSD was used in the sampling process of the noise supervision loss, its removal also led to a decrease of $UCC_{[\mu]}$ and $UR_{[\mu]}$.

**Discussion** To validate **Principle 3** for robustness enhancing, we evaluated the performance change when exposed to different levels of noise. Fig. 2 (a) illustrates the change of DSC values for all compared methods. Specifically, we use $\triangle DSC(\mu^i, \mu^j)$ representing the difference of DSC scores when applying two noises with mean value being $\mu_i$ and $\mu_j$ respectively. We chose $\mu_i \in \{0.0, 0.1, 0.3, 0.5, 0.7, 0.9\}$. Similarly, Fig. 2 (b) shows the results of $\triangle ECE(\mu^i, \mu^j)$ for all methods. One can see that our method (red) shows the best overall stability, regardless of the level of noise added.

Fig. 3 illustrates the segmentation results and uncertainty maps of different methods. Specifically, for **Uncertainty Interpretability**: (1) The second row overlays uncertainty maps with image gradients. Our method shows lower uncertainty in regions with higher gradients, conforming to **Principle 1**, while other methods lack this clear relationship. (2) The third row displays uncertainty difference maps between a test image and the image with patch noises. Our method effectively highlights noisy patches near edges, which conforms to

**Principle 2**. For **Uncertainty Robustness**, the fourth row shows uncertainty difference maps between an original image and the noisy one with noise imposed on the entire image. Our method emphasizes edge regions while maintaining stable uncertainty in non-edge areas (**Principle 3**), while other methods might exhibit significant uncertainty variations in regions beyond the boundary.

## 4 Conclusion

In this paper, we introduce a human-inspired uncertainty supervision method within the evidential learning framework. By utilizing image gradients and noise to constrain the uncertainty estimation, we not only provide reliable predictions but also offer interpretable and robust uncertainty estimations, which aligns with human experience. The proposed approach aids in understanding the sources of uncertainty, thereby facilitating better decision-making.

**Acknowledgments.** This work was funded by the National Natural Science Foundation of China (grant No. 62372115) and Shanghai Municipal Education Commission-Artificial Intelligence Initiative to Promote Research Paradigm Reform and Empower Disciplinary Advancement Plan (grant no. 24KXZNA13).

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

## References

1. Addimulam, S., Mohammed, M.A., Karanam, R.K., Ying, D., Pydipalli, R., Patel, B., Shajahan, M.A., Dhameliya, N., Natakam, V.M.: Deep learning-enhanced image segmentation for medical diagnostics. Malaysian Journal of Medical and Biological Research **7**(2), 145–152 (2020)
2. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., 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. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision. pp. 205–218. Springer (2022)
4. Erdur, A.C., Rusche, D., Scholz, D., Kiechle, J., Fischer, S., Llorián-Salvador, Ó., Buchner, J.A., Nguyen, M.Q., Etzel, L., Weidner, J., et al.: Deep learning for autosegmentation for radiotherapy treatment planning: State-of-the-art and novel perspectives. Strahlentherapie und Onkologie pp. 1–19 (2024)
5. Eslami, A., Paredes, B.R., Meyer, C., Rezende, D.J., De Fauw, J., Ledsam, J., Maier-Hein, K.H., Ronneberger, O., Kohl, S.: A probabilistic u-net for segmentation of ambiguous images (2018)
6. Falk, T., Mai, D., Bensch, R., Çiçek, Ö., Abdulkadir, A., Marrakchi, Y., Böhm, A., Deubner, J., Jäckel, Z., Seiwald, K., et al.: U-net: deep learning for cell counting, detection, and morphometry. Nature methods **16**(1), 67–70 (2019)

7. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059. PMLR (2016)
8. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International conference on machine learning. pp. 1321–1330. PMLR (2017)
9. Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.W., Wu, J.: Unet 3+: A full-scale connected unet for medical image segmentation. In: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 1055–1059. IEEE (2020)
10. Huang, L., Ruan, S., Decazes, P., Denœux, T.: Lymphoma segmentation from 3d pet-ct images using a deep evidential network. International Journal of Approximate Reasoning **149**, 39–60 (2022)
11. Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J.: Comparing images using the hausdorff distance. IEEE Transactions on pattern analysis and machine intelligence **15**(9), 850–863 (1993)
12. Jøsang, A.: Subjective logic, vol. 3. Springer (2016)
13. Joshi, A.J., Porikli, F., Papanikolopoulos, N.: Multi-class active learning for image classification. In: 2009 ieee conference on computer vision and pattern recognition. pp. 2372–2379. IEEE (2009)
14. Jungo, A., Reyes, M.: Assessing reliability and challenges of uncertainty estimations for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22. pp. 48–56. Springer (2019)
15. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? Advances in neural information processing systems **30** (2017)
16. Krithika Alias AnbuDevi, M., Suganthi, K.: Review of semantic segmentation of medical images using modified architectures of unet. Diagnostics **12**(12), 3064 (2022)
17. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems **30** (2017)
18. Mehrtash, A., Wells, W.M., Tempany, C.M., 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)
19. Orlando, J.I., Fu, H., Breda, J.B., van Keer, K., Bathula, D.R., Diaz-Pinto, A., Fang, R., Heng, P.A., Kim, J., Lee, J., et al.: Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Medical image analysis **59**, 101570 (2020)
20. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., Snoek, J.: Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. Advances in neural information processing systems **32** (2019)
21. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. Advances in neural information processing systems **31** (2018)
22. Shafer, G.: Dempster-shafer theory. Encyclopedia of artificial intelligence **1**, 330–331 (1992)
23. Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T.: Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing **338**, 34–45 (2019)

24. Wissler, C.: The spearman correlation formula. Science **22**(558), 309–311 (1905)
25. Yu, L., Wang, S., Li, X., Fu, C.W., Heng, P.A.: Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22. pp. 605–613. Springer (2019)
26. Zou, K., Chen, Y., Huang, L., Yuan, X., Shen, X., Wang, M., Goh, R., Liu, Y., Fu, H.: Towards reliable medical image segmentation by utilizing evidential calibrated uncertainty. arXiv preprint arXiv:2301.00349 (2023)
27. Zou, K., Yuan, X., Shen, X., Chen, Y., Wang, M., Goh, R.S.M., Liu, Y., Fu, H.: Evidencecap: towards trustworthy medical image segmentation via evidential identity cap. arXiv preprint arXiv:2301.00349 (2023)