


Are robust loss functions still relevant for medical image segmentation with noisy labels?

Vikram Venkatraghavan^{1,2} 


V.VENKATRAGHAVAN@AMSTERDAMUMC.NL

Richard A.P. Takx³ 

R.TAKX@AMSTERDAMUMC.NL

Niek E. van der Aa⁴ 


N.VANDERAA@UMCUTRECHT.NL

Timo R. de Haan⁵ 

T.R.DEHAAN@AMSTERDAMUMC.NL

Nils R. Planken^{3,6} 

PLANKEN.NILS@MAYO.EDU

Manon J.N.L. Benders⁴ 

M.BENDERS@UMCUTRECHT.NL

Ivana Išgum^{1,3,6} 

ISGUM.IVANA@MAYO.EDU

¹Department of Biomedical Engineering & Physics, Amsterdam University Medical Center, The Netherlands. ²Informatics Institute, University of Amsterdam, The Netherlands. ³Department of Radiology & Nuclear Medicine, Amsterdam University Medical Center, The Netherlands. ⁴Department of Neonatology, Wilhelmina Children’s Hospital, University Medical Center, Utrecht, The Netherlands. ⁵Department of Neonatology, Emma Children’s Hospital, Amsterdam University Medical Center, The Netherlands. ⁶Department of Radiology, Mayo Clinic, Rochester, United States of America

Editors: Under Review for MIDL 2026

Abstract

Creating reference annotations for semantic segmentation in medical images is a labor-intensive process, where experts often disagree on the precise location of the boundary between semantic structures. Hence, noisy labels in medical image segmentation tasks are pervasive. A popular approach to tackling noisy labels is to use robust loss functions that are resilient to noise. Meanwhile, the introduction of nnU-Net has highlighted the critical role that a well-configured combination of data augmentation, model architecture, and inference pipelines play in medical image segmentation. However, the potential of such strong baselines to mitigate the impact of label noise and the additional advantage of using a robust loss function has not been thoroughly explored. Most studies proposing robust loss functions, often with tunable hyper-parameters, have shown their efficacy either with a custom U-Net architecture, or without hyper-parameter optimization. By thorough benchmarking using a standardized nnU-Net framework, along with independent hyper-parameter optimization, we found that stochastic co-teaching based small-loss sample selection and active-passive loss comprising normalized generalized cross-entropy, reverse cross entropy, and Dice loss are useful for robust segmentation in applications with high label noise. For segmentation tasks characterized by minimal label noise, none of the robust loss functions demonstrate performance improvements over a well-established baseline model. Our results highlight the need for benchmarking with strong baseline models, even when proposing a new robust loss function that is architecture and framework independent.

Keywords: Medical image segmentation, Noisy labels, Robust loss

1. Introduction

Accurate semantic segmentation (Antonelli et al., 2022) in medical images can be challenging because of the inherent ambiguity in expert annotations. Delineating the boundaries of

anatomical structures and anomalies is a subjective process, often resulting in considerable interobserver variability (Karimi et al., 2020). This variability, coupled with the demanding nature of manual labeling (Wang et al., 2021), leads to the frequent presence of imprecise or inconsistent reference standard in segmentation datasets. Addressing the consequences of such label noise has been a long-standing focus in the field, as it can significantly degrade the performance of deep learning models.

Widely adopted strategies to counteract the detrimental effects of supervision with noisy labels include using loss functions that implicitly avoid overfitting to noise or using explicit sample-selection strategies that avoid noisy samples for loss computation. These strategies have been jointly referred to as *robust loss functions* in the rest of the manuscript. They are designed to reduce the influence of mislabeled or uncertain voxels during training, thereby promoting more reliable loss computation. Various robust loss functions (Zhang and Sabuncu, 2018; Ma et al., 2020; Gonzalez-Jimenez et al., 2025; Han et al., 2018; Yu et al., 2019; de Vos et al., 2023), often equipped with tunable parameters, have been introduced and evaluated, typically within the context of a custom U-Net architecture. While these studies have provided valuable insights, the landscape of medical image segmentation has undergone a significant shift with the emergence of nnU-Net (Isensee et al., 2020, 2024) as a state-of-the-art segmentation framework for a wide variety of segmentation tasks. It has shifted the focus from architectural novelty to meticulous self-configuration of networks.

Given that many robust loss functions were proposed prior to the widespread adoption of nnU-Net, it is important to re-examine their relevance and effectiveness in the current landscape. In this work, by conducting comprehensive experiments on multiple medical image segmentation datasets, we aim to clarify whether robust loss functions offer advantages beyond what is already achieved by a strong baseline.

2. Methods

For training our models, we use a standard nnU-Net (Isensee et al., 2020) architecture with large residual encoder configuration (Isensee et al., 2024). We also use the standardized sliding window inference with the framework’s default test-time augmentation.

We use nnU-Net model with a compound loss function comprising cross-entropy (CE) and Dice loss ($\mathcal{L}_{CE} + \mathcal{L}_{Dice}$) to provide us a **Baseline**. This configuration has previously been shown to be a strong benchmark in multiple applications (Khened et al., 2019; Isensee et al., 2020; Ma et al., 2021).

2.1. Benchmarking robust loss functions

We include three broad categories of robust loss functions: variants of Active-Passive losses (APL) (Ma et al., 2020), variants of sample selection strategies for robust loss computation (Han et al., 2018; de Vos et al., 2023; Yu et al., 2019), and model calibration based loss that offer robustness to noise as an auxiliary effect (Islam and Glocker, 2021). Furthermore, we explore a hybrid robust loss function that combine the above strategies. Finally, we performed independent optimization of the hyper-parameters involved in these robust losses, which we observed was often not investigated in validating robust loss functions.

Variants of APL: In this category, we include loss function combinations that were designed to boost each other so that the resulting composite losses are both robust to noise

and avoid underfitting (Ma et al., 2020). We consider two composite loss functions in this category:

i) APL1: Our first composite APL is a combination of reverse CE loss (RCE) and normalized CE (NCE) loss as suggested in (Ma et al., 2020). Let x denote the voxel-wise input and y the corresponding manually annotated label corresponding to one of K categories. We denote the one-hot encoding distribution of manually annotated labels by $q(k|x)$, where $k \in 1, \dots, K$ classes. By definition, $\sum_{k=1}^K q(k|x) = 1$. For computational purposes (*i.e.* avoiding to compute $\log(0)$), we set $q(k \neq y|x) = \delta$, where $0 < \delta \ll 1$. We denote the voxel-wise model predictions by $p(k|x)$. RCE and NCE can be mathematically denoted as:

$$\mathcal{L}_{RCE} = - \sum_{k=1}^K p(k|x) \log q(k|x) \quad (1)$$

$$\mathcal{L}_{NCE} = \frac{\mathcal{L}_{CE}(x, y)}{\sum_{j=1}^K \mathcal{L}_{CE}(x, j)} \quad (2)$$

where $\mathcal{L}_{CE}(x, j)$ is the CE loss, computed as:

$$\mathcal{L}_{CE}(x, j) = - \sum_{k=1}^K q(y = j|x) \log p(k|x) \quad (3)$$

Originally, the active-passive loss was formulated as $\mathcal{L}_{RCE} + \mathcal{L}_{NCE}$. However, we also add a third Dice loss term, reflecting the fact that the Baseline includes Dice loss. Thus:

$$\mathcal{L}_{APL1} = \mathcal{L}_{RCE} + \mathcal{L}_{NCE} + \mathcal{L}_{Dice} \quad (4)$$

ii) APL2: Our second composite APL again consists of RCE and Dice loss terms. However, it also includes normalized generalized cross entropy ($NGCE$), which uses GCE formulation introduced in Zhang and Sabuncu (2018) together with the loss normalization process introduced in Ma et al. (2020). $\mathcal{L}_{GCE}(x, j)$ for $j \in [1, K]$ is defined as following:

$$\mathcal{L}_{GCE}(x, j) = \frac{1 - \sum_{k=1}^K q(y = j | x) (p(k | x))^w}{w} \quad (5)$$

Here, $w \in (0, 1]$ is a tunable hyper-parameter. $NGCE$ is obtained by normalizing GCE , following normalization in Equation 2. Thus:

$$\mathcal{L}_{APL2} = \mathcal{L}_{RCE} + \mathcal{L}_{NGCE} + \mathcal{L}_{Dice} \quad (6)$$

Sample-selection strategies: In this category, we include peer network-based cross-training strategies that compute loss, robust to noisy labels realized through the selection of small-loss instances. The underlying loss function for training each peer-network is similar to our Baseline model ($\mathcal{L}_{CE} + \mathcal{L}_{Dice}$). We consider three robust loss functions in this category:

i) Co-teaching loss (CoT) (Han et al., 2018): Performs cross-training of two deep-learning networks. For training, each of the two networks selects small-loss instances of its peer

network. The threshold for selecting small-loss instances ($R(\tau)$) is chosen to reflect the expected noise level (τ). For unknown noise levels, as is often the case in medical applications, $R(\tau) \in [0, 0.1]$ is a tunable hyper-parameter;

ii) Co-teaching+ loss (CoT+) (Yu et al., 2019): During the training process, peer networks reaching consensus is sub-optimal, as it diminishes their ability to select noise-free samples for each other. To prevent the two peer networks to converge and reach consensus, CoT+ performs cross-training of two deep-learning networks exclusively exploiting disagreements between the two networks. Here as well, the threshold for selecting small-loss instances $R(\tau) \in [0, 0.1]$ is the hyper-parameter in this approach;

iii) Stochastic co-teaching loss (StoCoT) (de Vos et al., 2023), is a variant of CoT strategy, where the cross-training of two deep-learning networks is based on small-loss instances selected using stochastically chosen thresholds drawn from a *Beta*-distribution ($Beta(\alpha, \beta)$). The stochastically chosen thresholds for noisy sample selection have demonstrated effectiveness in preventing the two peer networks from reaching consensus. In this approach, the hyper-parameters are $\alpha \in [0.5, 8]$ and $\beta \in [0.5, 8]$, with a logarithmic search-space for both hyper-parameters.

Model calibration-based losses: In this category, we include Spatially varying label smoothing loss (SVLS) (Islam and Glocker, 2021), a loss function that has been designed for improved model calibration in image segmentation tasks, which has shown performance improvements in segmentation tasks. SVLS is a soft labeling technique that captures the spatial uncertainty to counteract the segmentation network’s tendency to be overly confident. In this method, the one-hot encoded labels $g(k|x)$ are smoothed to $g(k|x)_{SVLS}$ as follows:

$$g(k|x)_{SVLS} = \frac{\sum_{j \in \mathcal{N}_x} g(k|x) \times \mathcal{W}_j}{\sum_{j \in \mathcal{N}_x} \mathcal{W}_j} \quad (7)$$

Here, \mathcal{N}_x denotes the neighbourhood of voxel x . \mathcal{W}_j denotes the weights of a discrete spatial Gaussian kernel ($G(\theta)$) centered at x . In our implementation, the size of \mathcal{N}_x was chosen as $3 \times 3 \times 3$. The spread of the kernel $G(\theta) \in [0.2, 2]$ is a hyper-parameter of this approach.

Hybrid robust loss function: We explore a hybrid robust loss (HRL) function, which is constructed by combining losses from the aforementioned categories. Since many such combinations are possible, we restrict our choice to only one comprising the most recently developed method within each category. For the Hybrid robust loss function we use StoCoT based cross-training of two deep-learning networks, with the composite loss of $\mathcal{L}_{RCE} + \mathcal{L}_{NGCE} + \mathcal{L}_{Dice}$ (\mathcal{L}_{APL2} in Equation 6) used to train each peer network. Furthermore, we use SVLS-based smoothed one-hot encoded labels (Equation 7) for training.

nnRobustLoss framework: The code for each of these robust loss functions, along with their integration into the nnU-Net framework and a self-configuring hyper-parameter tuning framework (as described in Section 3.2), is available at <https://github.com/88vikram/nnRobustLoss>.¹

¹The repository will be made openly accessible upon acceptance of the manuscript.

3. Experimental setup

3.1. Segmentation tasks and Datasets

For benchmarking performance of different robust loss functions, we included three segmentation tasks that have noisy training labels, with images differing in imaging modality and anatomical coverage.

- i) Segmentation of Hypoxic-Ischemic Encephalopathy lesions in Neonatal Brain Diffusion-Weighted Imaging (HIE lesion segmentation) (Murphy et al., 2017);
- ii) Segmentation of heart chambers and myocardium in cardiac CT angiography (Heart Segmentation) (Bruns et al., 2022; Zhuang et al., 2019);
- iii) Segmentation of Kidney, renal tumors and cysts in contrast-enhanced abdominal CT (Kidney tumor segmentation) (Myronenko et al., 2023).

For details regarding the patient demographics, scanning protocols, and scanner specifications, we refer to the original publications describing each dataset. Each dataset was approved for use by the local institutional Medical Research Ethics Committee.

1. HIE lesion segmentation:

Dataset: We used Neonatal Brain Diffusion-Weighted Imaging (DWI) dataset for this task. HIE causes diffuse lesions in neonatal brains, making demarcation highly challenging. These lesions have profound implications for patients’ developmental outcomes, highlighting the importance of accurate automatic segmentation. Our dataset consists of 1.5T Neonatal Brain MRI dataset from Utrecht UMC, used previously in Murphy et al. (2017) ($n = 20$).

Labels: Since diffuse lesions are particularly challenging to segment and often subject to interpretation differences, it was important to assess the degree of inter-observer variability (Murphy et al., 2017). Therefore, the diffuse lesions were annotated independently by two medical experts (one highly experienced and one less experienced). Noisy labels consisted of labels from multiple annotators: $n = 14$ by the highly experienced annotator and $n = 6$ by the less experienced annotator. This was done to emulate a realistic situation with multiple annotators involved in annotating mutually exclusive subsets of the dataset. However, only the annotation masks of the highly experienced annotator were used as reference labels for model evaluation.

Evaluation: To evaluate performance of different networks trained on this dataset, we computed the Dice overlap and Average surface distance (ASD) between the automatically segmented target class and its reference label.

2. Heart segmentation:

Dataset: We used Cardiac CT angiography dataset for this task. Automatic segmentation of heart chambers and myocardium (left ventricle, right ventricle, left atrium, right atrium, myocardium) enables a quantitative assessment of cardiac anatomy (Chen et al., 2020). This could be of particular importance for patients undergoing cardiac intervention or surgery. We hence included patients with abnormal heart function by combining the public dataset from the multi-modal whole heart segmentation (MM-WHS) challenge (Zhuang et al., 2019) ($n = 20$) and the Amsterdam UMC dataset of patients who had been selected for transcatheter aortic valve implantation (TAVI) as described by Bruns et al. (2022) ($n = 24$).

Labels: Combining the two datasets for training ensured a natural variation in reference

labels as they were labeled by multiple annotators from different institutions. Furthermore, given that it takes approximately 5 hours per CT volume to perform these annotations manually, the annotations in the Amsterdam UMC dataset were initially performed by an annotator with limited expertise, which were later reviewed and corrected by an expert radiologist. We retained the labels of the inexperienced annotator as noisy labels in our experiments and used the corrected annotations as reference labels for evaluation.

Evaluation: The Dice overlap and ASD metrics were again computed between the automatically segmented target class and its reference.

3. Kidney tumor segmentation:

Dataset: We used contrast-enhanced abdominal CT dataset for this task. Kidney cysts and tumors are most often discovered incidentally during cross-sectional imaging, rather than on the basis of symptoms. There is hence an unmet clinical need to robustly detect these abnormalities in CT images. For this task, we utilized the dataset from the MICCAI KiTS2023 challenge (Myronenko et al., 2023) consisting of $n = 489$ CT scans with their respective reference labels, including the kidney, tumor, and cyst classes.

Labels: To simulate noisy labels in this well-curated dataset, we artificially corrupted the labels along the boundaries of the target structures by randomly performing a morphological dilation or erosion of the segmentation masks. The original labels in the KiTS2023 training dataset were used as reference labels for model evaluation.

Evaluation: The evaluation in this dataset was in accordance with the recommendation of the KiTS2023 challenge, where the evaluation was done on hierarchical classes, rather than on each target class by itself. The hierarchical classes were: Kidney + Tumor + Cyst, Tumor + Cyst, and Tumor. The evaluation metrics computed in these hierarchical classes were Dice and Surface Dice (Nikolov et al., 2021).

3.2. Model training and Hyper-parameter tuning

We trained models in a 3-fold cross-validated manner with the aforementioned Baseline and robust loss functions using noisy labels. Furthermore, we also trained our Baseline model on reference labels for each of these tasks, to provide an approximate upper-bound for the models trained on noisy labels.

The Baseline model, and the APL1 robust loss had no tunable hyper-parameters. For each approach that required hyper-parameter tuning, the optimum hyper-parameters were estimated in the training set using Optuna’s hyper-parameter optimization framework (Akiba et al., 2019). With Optuna, we used 30 iterations of Tree-structured Parzen estimator (Watanabe, 2025) to explore the hyper-parameter search space. For each iteration, a model was trained with the selected hyper-parameter(s), for 50 epochs. We identified the optimal hyper-parameter settings that maximized the model’s exponential moving average (EMA) of pseudo Dice computed in the training dataset after 50 epochs of training.

$$EMA_t = \lambda \cdot EMA_{t-1} + (1 - \lambda) \cdot \text{Pseudo Dice}_t \quad (8)$$

Here, t stands for the epoch number and Pseudo Dice is the Dice metric computed on patches sampled from the training dataset in the current epoch. The default value of $\lambda = 0.9$ was used. While the nnU-Net framework computes EMA over the validation dataset, our implementation instead computes this metric for the training dataset for hyper-parameter

optimization. Our decision to perform hyper-parameter optimization within the training dataset rather than using an independent validation dataset was motivated by two key considerations. First, the noise in the labels was not uniformly distributed across the entire dataset. As a result, selecting an independent validation set for hyper-parameter tuning would introduce variability and potentially biasing the optimization process. Second, to avoid the substantial computational burden associated with nested cross-validation of large residual encoder nnU-Nets. Once the optimum model hyper-parameters were estimated, we trained the nnU-Net with large residual encoder for 200 epochs with the chosen hyper-parameters. Testing was always performed on the held-out test data within each fold of cross-validation.

4. Results

4.1. Label noise quantification

For the HIE lesion segmentation, the mean Dice overlap between the annotations of less experienced and expert annotators was 0.56, reflecting substantial variability. Quantifying this disagreement is important because it illustrates the intrinsic ambiguity of diffuse lesion boundaries and highlights the inherent complexity and subjectivity involved in delineating diffuse lesions.

For the Heart segmentation, annotations provided by the inexperienced annotator were compared to the corrected reference annotations. The mean Dice overlap across all cardiac structures was found to be 0.96, indicating a high degree of agreement. The observed disagreements were primarily due to the thickening of papillary muscles in some patients, as well as inconsistent delineation of the boundary between the right ventricle and right atrium. These factors contributed to variability between the annotator and the reference standard. This high agreement provides a strong motivation for evaluating noise robustness in this task, as it represents a realistic low-noise setting in which robust loss functions might offer limited benefit.

For the segmentation of kidney tumor, since synthetic noise was added, the noise level was chosen to match the noise observed in the heart segmentation task as measured by the Dice overlap. Consequently, the mean Dice overlap in this task was also 0.96.

4.2. Hyper-parameter stability and Benchmarking results

Figure 1 shows the variation in achieved Pseudo Dice within the training dataset for different choices of hyper-parameters, across all three folds of cross-validation. It illustrates the sensitivity of different robust loss functions to their hyper-parameter choices, with CoT being the most sensitive. Among robust loss functions based on sample-selection strategies, StoCoT was the most resilient to hyper-parameter variation across different tasks.

Table 1 compares the mean evaluation metrics computed across all target classes and over the three cross-validation folds, using self-configured hyper-parameters where relevant. In the HIE lesion segmentation task, StoCoT, APL2, and HRL strategies outperformed the Baseline trained on noisy labels as well as on reference labels. Moreover, APL1 and CoT outperformed the Baseline model trained on noisy labels based on the Dice metric.

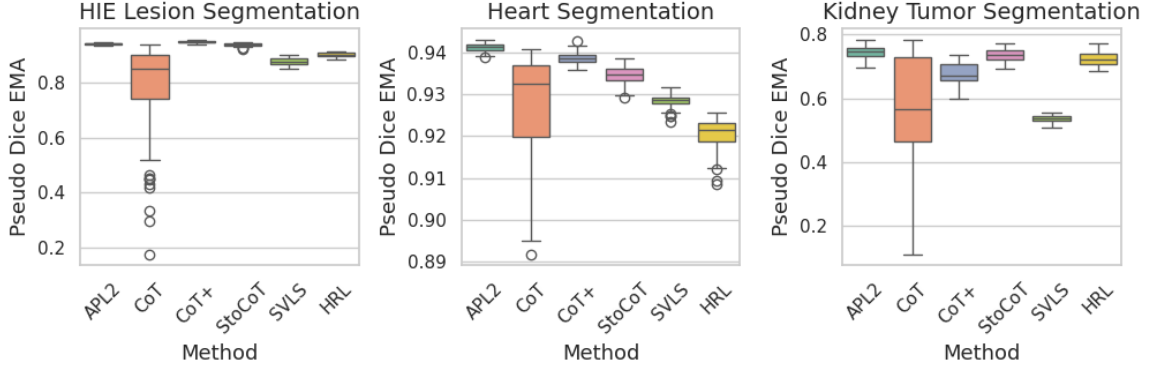


Figure 1: Hyper-parameter stability: Distribution of Pseudo Dice EMA values across different methods for three medical imaging datasets. Abbreviations: APL - Active-passive loss; CoT - co-teaching; StoCoT - Stochastic co-teaching; SVLS - spatially varying label smoothing; HRL - Hybrid robust loss.

Table 1: Quantitative comparison of segmentation performance across all methods and tasks trained on noisy labels. Baseline (ref.) refers to the Baseline model trained on reference labels without noise. Abbreviations: ASD - Average surface distance; APL - Active-passive loss; CoT - co-teaching; StoCoT - Stochastic co-teaching; SVLS - spatially-varying label smoothing; HRL - Hybrid robust loss.

Method	HIE Lesion		Heart		Kidney Tumor	
	Dice \uparrow	ASD \downarrow (mm)	Dice \uparrow	ASD \downarrow (mm)	Dice \uparrow	Surface Dice \uparrow
Baseline (ref.)	64.00	1.40	92.45	0.96	85.41	73.92
Baseline	63.33	1.31	92.22	1.00	84.94	71.07
APL1 (Eq. 4)	63.89	1.30	92.13	1.00	80.94	66.85
APL2 (Eq. 6)	64.05	1.32	92.25	0.99	82.95	69.91
CoT	63.53	1.23	92.14	1.00	84.23	70.47
CoT+	62.25	1.15	92.12	1.00	83.28	70.32
StoCoT	64.66	1.17	92.12	1.00	83.49	69.41
SVLS	62.36	1.02	91.98	1.02	74.32	56.91
HRL	64.23	1.31	92.05	1.02	80.22	64.76

All reported methods achieved better performance than the less experienced rater (with a Dice overlap of 0.56), whereas this could not be achieved in previous work on HIE lesion segmentation (Murphy et al., 2017) with a custom U-Net architecture further highlighting the importance of the nnU-Net framework.

In the Heart Segmentation task, APL2 was the only loss function that marginally outperformed the Baseline model trained on noisy labels, while none of the robust loss functions outperformed the Baseline in the Kidney Tumor segmentation task. The reported measures in the Heart Segmentation task exceeded those reported by [Bruns et al. \(2022\)](#), suggesting that these models are ready for large-scale validation in clinical studies where robust failure detection ([Zenk et al., 2025](#)) could play an important role. In the Kidney Tumor task, the results are not directly comparable with the KiTS2023 leaderboard, as our evaluation is based on 3-fold cross-validation in the training dataset, whereas the leaderboard reflects performance on previously unseen test data.

We noted that model performance with SVLS loss was consistently lower than that of other models, indicating the effect of model calibration on model performance. Table 2 compares target class-wise model performance for two models that includes model calibration loss: SVLS and Hybrid robust loss. The results suggest that combining different noise-robustness strategies with model calibration offers a balance between the two.

Table 2: Segmentation performance for all target classes, comparing SVLS loss and Hybrid robust loss-based models. Abbreviation: SVLS - spatially varying label smoothing; ASD - Average surface distance; LV - Left Ventricle; RV - Right Ventricle; LA - Left Atrium; RA - Right Atrium; Myo - Myocardium.

Dataset Target Class	SVLS		Hybrid Robust Loss	
HIE Lesion Lesion	Dice \uparrow 62.36	ASD (mm) \downarrow 1.02	Dice 64.23	ASD (mm) \downarrow 1.31
Heart	Dice \uparrow	ASD (mm) \downarrow	Dice	ASD (mm) \downarrow
LV	92.81	0.89	92.78	0.89
RV	90.60	1.24	90.68	1.23
LA	93.77	0.89	93.77	0.95
RA	92.07	1.23	92.37	1.19
Myo	90.63	0.83	90.65	0.83
Kidney Tumor	Dice \uparrow	Surface Dice \uparrow	Dice	Surface Dice \uparrow
Kidney	95.14	83.11	94.79	83.5
Cyst	65.04	44.24	71.75	53.6
Tumor	62.79	43.38	74.14	57.2

5. Discussions

In this work, we systematically investigated the robustness to reference label noise of a Baseline nnU-Net model and compared it with several robust loss functions specifically designed to handle noisy labels. The segmentation tasks spanned multiple anatomical structures (kidney, heart), pathologies (diffuse brain lesions, kidney cysts & tumors), and imaging modalities (CT, MRI). In the process, we performed a rigorous and unbiased hyper-parameter

estimation for different robust loss functions. Our self-configuring hyperparameter estimation framework ([nnRobustLoss](#)), is compatible with the nnU-Net segmentation framework, allowing future studies to adopt it readily using our publicly available code. Our results demonstrate that the Baseline nnU-Net model is remarkably robust to label noise for low levels of noise present in Heart and Kidney segmentation tasks. Despite the strong performance of our Baseline model, through our experiments we find that robust loss functions remain relevant, albeit with certain caveats discussed below.

When there is too much label noise: In scenarios where reference labels are difficult to annotate even for experts, such as delineating diffuse lesions, robust loss functions consistently outperform the Baseline nnU-Net model. Approaches like stochastic co-teaching and active-passive loss functions, which combine normalized generalized cross-entropy, reverse cross-entropy, and Dice loss, are particularly effective. These methods also remain stable across a wide range of hyper-parameter settings, making them suitable for applications with variable, annotator-dependent label noise. Future work could explore combining these robust loss functions with iterative mask update strategies ([Li et al., 2021](#)), to potentially yield further improvements.

When model calibration is crucial: Robust loss functions that are designed to improve model calibration, such as spatially-varying label smoothing, ensure that predicted probabilities are reflective of true outcome likelihoods. This is particularly important in safety-critical applications, where overconfident predictions can have serious consequences. However, our results in Table 1 show that model calibration can lead to reduction in model performance. With the right combination of robust loss functions, such as in our Hybrid Robust loss, we were able to train calibrated models without a substantial reduction in model performance (as seen in HIE lesion and Heart segmentation tasks).

When the target classes are balanced: However, all the explored robust loss functions failed to outperform the Baseline model in the presence of a large imbalance in the size of the target classes, as observed in the Kidney Tumor segmentation task. Future work on robust loss functions could focus on addressing such imbalances in target classes.

We noted that none of the previous work that proposed loss functions for addressing label noise had reported on hyper-parameter optimization. Consequently, model comparisons in prior work also did not rely on optimal hyper-parameters for the competing methods, making the resulting comparisons less reliable. This applies both to the approaches evaluated in our study as well as to those excluded, such as Neighbour-aware Calibration Loss ([Murugesan et al., 2025](#)) and Robust T-loss ([Gonzalez-Jimenez et al., 2025](#)). We did not include these robust losses because the former relies solely on CE loss and would have required altering our $\mathcal{L}_{CE} + \mathcal{L}_{Dice}$ Baseline, whereas the latter was developed for 2D segmentation and does not readily extend to 3D tasks. As our results show, independent hyper-parameter optimization, fair and methodologically consistent comparison with prior work are essential for meaningful progress in developing robust loss functions.

In conclusion, our experiments demonstrate that robust loss functions remain valuable in medical image segmentation, particularly in settings with substantial label noise or where calibrated predictions are essential. Our results underscore the importance of rigorous benchmarking against well-configured baselines and provide an openly available evaluation framework to support future advances in robust learning under label noise.

Acknowledgments

This work was supported by the University of Amsterdam Research Priority Area Artificial Intelligence for Health Decision-making. This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-16134.

Disclosure

Dr. Richard A.P. Takx has received consultancy fees from Hemolens Diagnostics.

References

- Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.
- Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M. Summers, Bram van Ginneken, Michel Bilello, Patrick Bilic, Patrick F. Christ, Richard K. G. Do, Marc J. Gollub, Stephan H. Heckers, Henkjan Huisman, William R. Jarnagin, Maureen K. McHugo, Sandy Napel, Jennifer S. Golia Pernicka, Kawal Rhode, Catalina Tobon-Gomez, Eugene Vorontsov, James A. Meakin, Sebastien Ourselin, Manuel Wiesenfarth, Pablo Arbeláez, Byeonguk Bae, Sihong Chen, Laura Daza, Jianjiang Feng, Baochun He, Fabian Isensee, Yuanfeng Ji, Fucang Jia, Ildoo Kim, Klaus Maier-Hein, Dorit Merhof, Akshay Pai, Beomhee Park, Mathias Perslev, Ramin Rezaiifar, Oliver Rippel, Ignacio Sarasua, Wei Shen, Jaemin Son, Christian Wachinger, Liansheng Wang, Yan Wang, Yingda Xia, Daguang Xu, Zhanwei Xu, Yefeng Zheng, Amber L. Simpson, Lena Maier-Hein, and M. Jorge Cardoso. The medical segmentation decathlon. *Nature Communications*, 13(1), July 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-30695-9.
- Steffen Bruns, Jelmer M. Wolterink, Thomas P.W. van den Boogert, Jurgen H. Runge, Berto J. Bouma, José P. Henriques, Jan Baan, Max A. Viergever, R. Nils Planken, and Ivana Išgum. Deep learning-based whole-heart segmentation in 4d contrast-enhanced cardiac ct. *Computers in Biology and Medicine*, 142:105191, March 2022. ISSN 0010-4825. doi: 10.1016/j.combiomed.2021.105191.
- Chen Chen, Chen Qin, Huaqi Qiu, Giacomo Tarroni, Jinming Duan, Wenjia Bai, and Daniel Rueckert. Deep learning for cardiac image segmentation: A review. *Frontiers in Cardiovascular Medicine*, 7, March 2020. ISSN 2297-055X. doi: 10.3389/fcvm.2020.00025.
- Bob D de Vos, Gino E Jansen, and Ivana Išgum. Stochastic co-teaching for training neural networks with unknown levels of label noise. *Scientific reports*, 13(1):16875, October 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-43864-7.
- Alvaro Gonzalez-Jimenez, Simone Lionetti, Philippe Gottfrois, Fabian Gröger, Alexander Navarini, and Marc Pouly. Robust t-loss for medical image segmentation. *Medical Image Analysis*, 105:103735, October 2025. ISSN 1361-8415. doi: 10.1016/j.media.2025.103735.

- Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.
- Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, December 2020. ISSN 1548-7105. doi: 10.1038/s41592-020-01008-z.
- Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F. Jäger. *nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation*, page 488–498. Springer Nature Switzerland, 2024. ISBN 9783031721144. doi: 10.1007/978-3-031-72114-4_47.
- Mobarakol Islam and Ben Glocker. *Spatially Varying Label Smoothing: Capturing Uncertainty from Expert Annotations*, page 677–688. Springer International Publishing, 2021. ISBN 9783030781910. doi: 10.1007/978-3-030-78191-0_52.
- Davood Karimi, Haoran Dou, Simon K. Warfield, and Ali Gholipour. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical Image Analysis*, 65:101759, October 2020. ISSN 1361-8415. doi: 10.1016/j.media.2020.101759.
- Mahendra Khened, Varghese Alex Kollerathu, and Ganapathy Krishnamurthi. Fully convolutional multi-scale residual densenets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers. *Medical Image Analysis*, 51:21–45, January 2019. ISSN 1361-8415. doi: 10.1016/j.media.2018.10.004.
- Shuailin Li, Zhitong Gao, and Xuming He. Superpixel-guided iterative learning from noisy labels for medical image segmentation, 2021. URL <https://arxiv.org/abs/2107.10100>.
- Jun Ma, Jianan Chen, Matthew Ng, Rui Huang, Yu Li, Chen Li, Xiaoping Yang, and Anne L. Martel. Loss odyssey in medical image segmentation. *Medical Image Analysis*, 71:102035, July 2021. ISSN 1361-8415. doi: 10.1016/j.media.2021.102035.
- Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *Proceedings of the 37th International Conference on Machine Learning*, ICML’20. JMLR.org, 2020.
- Keelin Murphy, Niek E. van der Aa, Simona Negro, Floris Groenendaal, Linda S. de Vries, Max A. Viergever, Geraldine B. Boylan, Manon J.N.L. Benders, and Ivana İşgum. Automatic quantification of ischemic injury on diffusion-weighted mri of neonatal hypoxic ischemic encephalopathy. *NeuroImage: Clinical*, 14:222–232, 2017. ISSN 2213-1582. doi: 10.1016/j.nicl.2017.01.005.
- Balamurali Murugesan, Sukesh Adiga Vasudeva, Bingyuan Liu, Herve Lombaert, Ismail Ben Ayed, and Jose Dolz. Neighbor-aware calibration of segmentation networks with

- penalty-based constraints. *Medical Image Analysis*, 101:103501, April 2025. ISSN 1361-8415. doi: 10.1016/j.media.2025.103501.
- Andriy Myronenko, Dong Yang, Yufan He, and Daguang Xu. Automated 3d segmentation of kidneys and tumors in miccai kits 2023 challenge, 2023.
- Stanislav Nikolov, Sam Blackwell, Alexei Zverovitch, Ruheena Mendes, Michelle Livne, Jeffrey De Fauw, Yojan Patel, Clemens Meyer, Harry Askham, Bernadino Romera-Paredes, Christopher Kelly, Alan Karthikesalingam, Carlton Chu, Dawn Carnell, Cheng Boon, Derek D’Souza, Syed Ali Moinuddin, Bethany Garie, Yasmin McQuinlan, Sarah Ireland, Kiarna Hampton, Krystle Fuller, Hugh Montgomery, Geraint Rees, Mustafa Suleyman, Trevor Back, Cían Owen Hughes, Joseph R Ledsam, and Olaf Ronneberger. Clinically applicable segmentation of head and neck anatomy for radiotherapy: Deep learning algorithm development and validation study. *Journal of Medical Internet Research*, 23(7): e26151, July 2021. ISSN 1438-8871. doi: 10.2196/26151.
- Shanshan Wang, Cheng Li, Rongpin Wang, Zaiyi Liu, Meiyun Wang, Hongna Tan, Yaping Wu, Xinfeng Liu, Hui Sun, Rui Yang, Xin Liu, Jie Chen, Huihui Zhou, Ismail Ben Ayed, and Hairong Zheng. Annotation-efficient deep learning for automatic medical image segmentation. *Nature Communications*, 12(1), October 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-26216-9.
- Shuhei Watanabe. Tree-structured parzen estimator: Understanding its algorithm components and their roles for better empirical performance, 2025.
- Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *International conference on machine learning*, pages 7164–7173. PMLR, 2019.
- Maximilian Zenk, David Zimmerer, Fabian Isensee, Jeremias Traub, Tobias Norajitra, Paul F. Jäger, and Klaus Maier-Hein. Comparative benchmarking of failure detection methods in medical image segmentation: Unveiling the role of confidence aggregation. *Medical Image Analysis*, 101:103392, April 2025. ISSN 1361-8415. doi: 10.1016/j.media.2024.103392.
- Zhilu Zhang and Mert R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels, 2018.
- Xiahai Zhuang, Lei Li, Christian Payer, Darko Štern, Martin Urschler, Mattias P. Heinrich, Julien Oster, Chunliang Wang, Örjan Smedby, Cheng Bian, Xin Yang, Pheng-Ann Heng, Aliasghar Mortazi, Ulas Bagci, Guanyu Yang, Chenchen Sun, Gaetan Galisot, Jean-Yves Ramel, Thierry Brouard, Qianqian Tong, Weixin Si, Xiangyun Liao, Guodong Zeng, Zenglin Shi, Guoyan Zheng, Chengjia Wang, Tom MacGillivray, David Newby, Kawal Rhode, Sebastien Ourselin, Raad Mohiaddin, Jennifer Keegan, David Firmin, and Guang Yang. Evaluation of algorithms for multi-modality whole heart segmentation: An open-access grand challenge. *Medical Image Analysis*, 58:101537, December 2019. ISSN 1361-8415. doi: 10.1016/j.media.2019.101537.